A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think Task

Wessel, Ineke; Albers, Casper J; Zandstra, Anna Roos E; Heininga, Vera E

Published in:
Memory

DOI:
10.1080/09658211.2020.1797095

Routledge
Taylor & Francis Group

Memory

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think Task

Ineke Wessel , Casper J. Albers , Anna Roos E. Zandstra & Vera E. Heininga

To cite this article: Ineke Wessel , Casper J. Albers , Anna Roos E. Zandstra & Vera E. Heininga (2020) A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think Task, Memory, 28:7, 870-887, DOI: 10.1080/09658211.2020.1797095

To link to this article: https://doi.org/10.1080/09658211.2020.1797095

Published online: 23 Jul 2020.

MEMORY
2020, VOL. 28, NO. 7, 870–887
https://doi.org/10.1080/09658211.2020.1797095

Routledge
Taylor & Francis Group

A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think Task

Ineke Wessel [a], Casper J. Albers [b], Anna Roos E. Zandstra [a] and Vera E. Heininga [c,d]

[a]Department of Clinical Psychology and Experimental Psychopathology, University of Groningen, Groningen, Netherlands; [b]Department of Psychometrics & Statistics, University of Groningen, Groningen, Netherlands; [c]Department of Developmental Psychology, University of Groningen, Groningen, Netherlands; [d]Research group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium

ABSTRACT

In 2001, Anderson and Green [2001. Suppressing unwanted memories by executive control. *Nature*, *410*(6826), 366–369] showed memory suppression using a novel Think/No-think (TNT) task. When participants attempted to prevent studied words from entering awareness, they reported fewer of those words than baseline words in subsequent cued recall (i.e., suppression effect). The TNT literature contains predominantly positive findings and few null-results. Therefore we report unpublished replications conducted in the 2000s ($N = 49$; $N = 36$). As the features of the data obtained with the TNT task call for a variety of plausible solutions, we report parallel "universes" of data-analyses (i.e., multiverse analysis) testing the suppression effect. Two published studies (Wessel et al., 2005. Dissociation and memory suppression: A comparison of high and low dissociative individuals' performance on the Think–No think Task. *Personality and Individual Differences*, *39*(8), 1461–1470, $N = 68$; Wessel et al., 2010. Cognitive control and suppression of memories of an emotional film. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(2), 83–89. https://doi.org/10.1016/j.jbtep.2009.10.005, $N = 80$) were reanalysed in a similar fashion. For recall probed with studied cues (Same Probes, SP), some tests (sample 3) or all (samples 2 and 4) showed statistically significant suppression effects, whereas in sample 1, only one test showed significance. Recall probed with novel cues (Independent Probes, IP) predominantly rendered non-significant results. The absence of statistically significant IP suppression effects raises problems for inhibition theory and its implication that repression is a viable mechanism of forgetting. The pre-registration, materials, data, and code are publicly available (https://osf.io/qgcy5/).

ARTICLE HISTORY
Received 27 March 2020
Accepted 9 July 2020

KEYWORDS
TNT task; suppression-induced forgetting; suppression effect; same probes; independent probes; multiverse analysis

For many years, (clinical) psychologists have been interested in studying motivated forgetting (Anderson & Huddleston, 2012). Especially the recovered memory debate of the 1990s focused on the question of whether severe trauma such as child sexual abuse can be forgotten completely, only to be retrieved in psychotherapy (see Loftus, 1993). The mechanism underlying such massive forgetting that was under attack in this debate was (Freudian) repression, referring to the idea that unwanted traumatic memories become unavailable to conscious awareness but are expressed as behaviour and/or psychopathological symptoms. Although ideas resembling this definition are still prevalent, empirical research has failed to find evidence for the existence of repression (Holmes, 1990; Otgaar et al., 2019). However, efforts to study repression empirically are complicated by the assumption that it acts unconsciously. To circumvent this problem, some scholars extended the concept to encompass forgetting resulting from the deliberate avoidance of unwanted memories (see Brewin & Andrews, 2014; Conway, 2001; Erdelyi, 2006).

The Think/No-think task (TNT) was designed specifically to show that deliberate retrieval avoidance hampers the subsequent recall of the avoided material (Anderson & Huddleston, 2012). Conway (2001) welcomed the first publication (Anderson & Green, 2001) on the task as providing "an unambiguous model for exploring memory repression in the laboratory" (p. 319). Others were skeptical about the fit between the theoretical construct and the experimental task (Kihlstrom, 2002) and/or the replicability of the findings (Bulevich et al., 2006). Nevertheless, research on the TNT gained momentum and by 2012, Anderson and Huddleston could identify 32 published studies for their literature review.

The initial version of the TNT task (Anderson & Green, 2001) consisted of several phases. First, participants studied cue – target word pairs (e.g., "tattoo – uncle";

CONTACT Ineke Wessel j.p.wessel@rug.nl Grote Kruisstraat 2/1, Groningen 9712 TS, The Netherlands @inekewessel; Albers @caal; Zandstra @anrozan; Heininga @HeiningaVE

"braid – doll"). In a subsequent TNT phase, the studied cue words were presented in either respond or suppress trials. For respond trials, participants were instructed to say the studied target word out loud as quickly as possible after cue presentation (e.g., when the word "tattoo" was presented, participants had to respond with "uncle"). For suppress trials, participants were instructed to keep the target (e.g., "doll") out of awareness in any way they could when the corresponding cue (e.g., "braid") was presented. Some cues that had been part of the cue – target pairs in the study phase were not presented in the TNT phase at all and were designated to be baseline cues. During a final recall phase, all baseline, respond and suppress cues were presented. The participants were instructed to respond to all cues, regardless of previous (respond or suppress) instructions in the TNT phase. Anderson and Green (2001) showed in several experiments that the recall of respond targets had improved relative to baseline. That is hardly surprising: responding to cues invited more practice, especially because the target was briefly presented during the TNT phase if the participant had not responded to the cue at all. In contrast, baseline cues were never presented and thus, the corresponding targets would not have been retrieved to begin with. The key finding, however, was a suppression effect: participants recalled fewer suppress words than baseline words, suggesting that intentionally keeping targets out of awareness had hampered recall more than simply doing nothing. In addition, Anderson and Green (2001) observed that the magnitude of the suppression effect increased linearly with the number of repetitions.

A central idea in the TNT literature is that suppression effects are due to memory inhibition (Anderson & Green, 2001; Anderson & Huddleston, 2012). That is, rather than obstructing the access route towards a memory, avoidance would result in decreased activation of the representation itself. However, Anderson and Green (2001) noted that if recall during the test phase is prompted with the same cues as were presented during the TNT phase, it is impossible to infer the precise mechanism underlying the results. That is, a suppression effect obtained with such a Same Probe (SP) test may result from inhibitory as well as non-inhibitory mechanisms. For example, to avoid responding with the target during the TNT phase, participants may create distracting thoughts upon cue presentation. Such other thoughts might then become associated with the cue, acting as substitutes for the target and interfering with the retrieval of the correct target in the subsequent SP test. Alternatively, repeatedly attempting to avoid thinking of the target in itself may make cues less effective during recall, because the associations between cues and targets are weakened (i.e., unlearning). In order to differentiate between non-inhibitory and inhibitory accounts of the suppression effect, Anderson and Green (2001) devised an Independent Probe (IP) test. In this test, the original cues were replaced by the semantic categories (e.g., "toy") of the original targets (e.g., "doll"). Because the IP test relies on cues that are only presented during the recall phase, a suppression effect may be attributed to inhibition rather than interference or unlearning. Indeed, the authors reported that the results obtained with the IP and SP tests were comparable and thus interpreted all results in terms of inhibition, meaning a reduced activation of the memory trace itself.

Since Anderson and Green's initial paper, the TNT task was used in a substantial number of publications, using a variety of stimuli (e.g., words, pictures, autobiographical memories) and outcome measures (e.g., number of items recalled, reaction times, fMRI data). Anderson and Huddleston (2012) summarised the findings of 32 articles published up to and including 2011, together with all published and unpublished recall data collected in their lab (see also Levy & Anderson, 2008). Overall, the suppression effect in this combined sample of psychologically healthy participants was 8% for SP and 6% for IP tests. Likewise, a recent unpublished meta-analysis (Stramaccia et al., 2019, preprint) reported a reliable small to medium effect size for SP suppression in healthy control groups in TNT studies focusing on psychopathology (e.g., PTSD, depression).

The findings in the TNT literature seem to be mainly positive (Anderson & Huddleston, 2012). Only a few non-replications are available (e.g., Bulevich et al., 2006; Mecklinger et al., 2009). This positive overall picture suggests that the effects of retrieval suppression are robust. However, in recent years it has become apparent that as a whole, the scientific literature in psychology suffers from publication bias (e.g., Ioannidis et al., 2014; Munafò et al., 2017; Nelson et al., 2018). Publication bias comes in various forms (Ioannidis et al., 2014). One form is the file drawer problem (Rosenthal, 1979), referring to the situation that not all completed studies are published. Other forms are that positive findings are selectively reported in publications (disregarding the negative results from the same study) or that findings are false positives due to flexible choices in data analysis. As Simmons et al. (2011) put it: "it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields 'statistical significance', and to then report only what 'worked'" (p. 1359). This behaviour is believed to be unintentional, resulting from a combination of cognitive biases in authors (e.g., confirmation bias, Munafò et al., 2017) and ambiguity about which data-analytic choice is the most optimal. However, various (subtle) analytic decisions imply multiple potential outcomes and this can be viewed as a multiple comparison problem (Gelman & Loken, 2014; Steegen et al., 2016). Choosing one route out of many possible routes harbours an inflated probability of incorrectly rejecting the null-hypothesis (i.e., Type I error).

It is unknown to what extent the published literature on the TNT task suffers from a publication bias. There are some hints that it does. For example, Barnier (2012) mentions that her team tried to extend the TNT task to

autobiographical memory in four studies, but failed to find an effect. A literature search suggests that these results did not make it into the published literature as a separate article. Furthermore, there is a hint of a file drawer effect in the systematic review by Stramaccia et al. (2019, preprint) of studies employing the TNT in samples with psychopathology. The authors plotted the magnitude of the effects against study precision (i.e., standard errors) for clinical samples and healthy control groups separately. It appeared that there was no reason to believe that studies were missing for clinical groups. For control groups, however, a trim-and-fill procedure suggested that about six data points for healthy controls would be missing (from a plot containing 27 data points). Yet, this unpublished meta-analysis was conducted on a subset of studies, and to date, we are unaware of any published systematic review of all studies using the TNT task since 2001. In addition, the meta-analysis by Stramaccia et al. (2019, preprint) did not include IP effects because only a few studies in their sample contained a report of such a test. Nevertheless, a meta-analytic summary of IP findings is crucial as it would allow an unambiguous interpretation of suppression effects as a result of inhibition.

The aim of the present paper is twofold. First, in light of the growing recognition that both positive and negative results should be published to obtain a balanced scientific record, we report on unpublished findings with the TNT paradigm. In 2001, one of us (IW) set out to replicate Anderson and Green's (2001) findings, using their task specifications and literal instructions (albeit translated to Dutch). Two experiments did not produce statistically significant suppression effects, neither for SP, nor for IP performance. These data were never published. Some years later, we (IW & AREZ) tried to replicate the basic effect with an optimised version of the TNT task (i.e., drop-off TNT; Levy & Anderson, 2012). In contrast with the 50% correct learning criterion of the original TNT, a drop-off procedure ensured that participants learned all word pairs in the study list. Here, the initial analyses showed a below baseline suppression effect for SP data, but not IP data. Again, the data were not published. We revisit these unpublished results in the present paper.

The second aim of the present paper is to provide an overview of potential results given the various analytic choices that are inherent in this type of research. That is, data generated with the TNT task are typically skewed and there are various ways of handling violation of the normality assumption in parametric tests. In addition, although within-participant comparisons are the main interest, the stimuli are rotated over different versions and the order of testing (SP test first or IP test first) is counterbalanced. To statistically control for their influence, one may choose to include either one or both of these between-participant factors in the statistical analysis. In the present paper, we vary such plausible choices in the data analysis in a systematic way and present a multiverse of outcomes (Heininga et al., 2015; Steegen et al., 2016). A multiverse analysis explores all different but plausible parallel "universes" that exist for testing one and the same hypothesis (for example, analyses with and without outliers), thus providing insight into the robustness of a result. That is, exploring the robustness in outcomes across all reasonable "universes" may protect researchers from drawing conclusions that are supported by a specific set of modelling choices, but fail to hold more generally. In addition to the multiverse analysis of the unpublished datasets, we revisit the data of two published studies (Wessel et al., 2005, 2010) in a similar fashion. These data were analysed in a time that we were less conscious of the impact that data-driven decisions may have on the probability of false positives. A multiverse analysis should shed light on how the published results compare to the array of possible results. Specifically, we report a multiverse analysis for each sample focusing on various ways of handling violation of the normality assumption in parametric tests (i.e., outliers, skewed distributions) and the inclusion of covariates.

Pre-registered hypotheses

We explored a multiverse of analyses in four samples to test two hypotheses. First, we concentrated on the suppression effect. We expected that compared to never-suppressed targets (baseline), participants would recall a lower percentage of targets that were to be suppressed for 16 times (suppress-16). The suppression effect was expected to occur for cues that were part of the original study context (SP test), as well as for cues that were semantically related to the words in the original study context (IP test). Second, a pattern of increasing recall for respond trials and decreasing recall for suppress trials is typically reported in the literature. Therefore we examined the multiverse for the instructions (respond, suppress) by repetitions (0, 1, 8, 16 times) interaction. We expected the typical pattern to emerge for SP as well as IP recall.

Method

Statement of transparency

The current report is based on the data obtained with the TNT task in five separate experiments conducted in the 2000s. Three experiments are unpublished; two were reported in the literature (Wessel et al., 2005, 2010). The data of experiments 1 and part of experiment 2 originated from a similar TNT task and were combined into one sample. Thus, four samples were derived from five experiments. In the present paper, we only report on the recall data obtained with the TNT task in those samples. We preregistered the hypotheses and analysis plan (https://osf.io/mcs8r). A description of the deviations from the preregistered analysis plan can be found in the "analytical approach" section below. We publicly provide the materials (https://osf.io/ga8db/) as well as the anonymised recall data (https://osf.io/jeu3f/). In addition to testing the

replicability of the suppression effect, all experiments except experiment 1 examined other hypotheses. We publicly provide the additional materials (https://osf.io/ga8db/) that were still available, and the anonymised data obtained with the additional measures (https://osf.io/j7aqz/).

Participants

All participants were undergraduate students and native Dutch speakers. They were financially reimbursed for their participation (sample 1, experiment 1: Fl.25,- / appr. €11; sample 1, experiment 2: €7; sample 2: €10; sample 3: €5; sample 4: €7.50).

Sample 1

Participants ($N = 49$) were tested at Maastricht University in November / December, 2001 (experiment 1, $n = 32$; 27 women, 5 men) and in June 2002 (experiment 2; $n = 17$; age and gender unknown[2]). They were 18–25 years old ($M = 19.47$, $SD = 1.59$, $n = 17$ no data). Experiment 2 originally included two conditions. One condition contained a TNT task that was similar to the one used in experiment 1, whereas the task completed by the other participants ($n = 18$) was substantially shorter. We merged the data of the participants in experiments 1 and 2 who had completed TNT tasks of similar duration into sample 1.

Sample 2

Participants ($N = 36$; 23 women, 13 men) were tested at the University of Groningen in spring 2007. Their age ranged from 18 to 26 years ($M = 21.22$, $SD = 1.69$). Recruitment had been combined with recruitment for an unrelated (and also unpublished) study comparing students high and low in neuroticism. Therefore, the participants in sample 2 all scored in the mid-range (i.e., scores between 41 and 51) of the Neuroticism subscale of the Five Factor Personality Inventory (FFPI; Hendriks et al., 1999).

Sample 3

Sample 3 consisted of the participants ($N = 68$; 63 women, 5 men) in Wessel et al. (2005), who were tested at Maastricht University in the academic year 2002–2003. They were 17–27 years old ($M = 19.69$, $SD = 2.02$). They scored either high (≥ 20; $n = 35$) or low (< 10; $n = 33$) on the Dissociative Experiences Scale (DES; Bernstein & Putnam, 1986).

Sample 4

Sample 4 ($N = 80$; 64 women, 16 men) were participants in Wessel et al. (2010) and were tested at the University of Groningen in the academic year 2004–2005. Their age range was 17–25 years ($M = 20.60$, $SD = 1.87$). All were evening types according to their scores (< 41) on the Morningness-Eveningness Questionnaire (MEQ; Horne & Östberg, 1976).

Materials

Sample 1

In the TNT task completed by sample 1, Anderson and Green's (2001) original procedure was followed as closely as possible. Other than a post-experimental questionnaire, no additional measures were administered.

TNT Task Stimuli. The stimuli were 50 neutral cue – target word pairs (e.g., tattoo – uncle; braid – doll), 10 of which were fillers. Most word pairs were Dutch translations of Anderson and Green's (2001) original stimuli. Whenever a Dutch translation was not usable (e.g., because of ambiguity due to multiple meanings), it was replaced. We followed recommendations (M.C. Anderson, personal communication, September 6, 2001) that the cues and targets should be semantically unrelated to all other words in the task, but should weakly relate to each other to enhance learning. The translation, including an indication of the differences with Anderson and Green's list, can be found at https://osf.io/u4yxp/. The cue – target pairs were clustered into eight sets of five words, which were rotated across two within subjects factors in the experimental design (i.e., respond vs suppress instruction, number of repetitions, see below). This resulted in eight versions of the task. With 3 exceptions, experiments 1 and 2 used the same cue – target pairs. However, the word pairs were distributed differently across word sets (see https://osf.io/fqtnr/ and https://osf.io/yrj8w/). Therefore the combination of Experiments 1 and 2 contains 16 versions, rather than the eight versions in Anderson and Green (2001).

TNT Task Procedure. The TNT task was programmed using Experimental RunTime Software (ERTS; Beringer, 1996; Dutta, 1995). The task consisted of 6 phases: (1) study; (2) test-feedback; (3) hint-training; (4) warm-up Think/No-think (TNT); (5) TNT; and (6) test. The test phase had two versions: the Same Probe (SP) and the Independent Probe (IP) test. The task ran on a desktop computer with a 640×480 pixel (235×170 mm) VGA monitor. Participants' responses were picked up by a microphone that was connected to a voice key. Experimenter feedback about accuracy was delivered through a response box.

Unless stated otherwise, the following applied to each phase in the task procedure. The stimuli were presented in the middle of the screen in black font (Geneva, 24 point) against a light grey background. Fixation crosses (Geneva, 32 point) were presented for 200 ms. At the start of each phase, an instruction to "press a key when ready" was followed by a count-down trial displaying the digits 3, 2, and 1 successively for 1.5 s. Next, the actual trial sequence was presented in a fixed order. Each sequence started and ended with two filler trials. Participants were notified ("End") after all trials were presented. Instructions, count-down and end trials were presented in NRC7BIT font.

Study Phase. Each trial in the study phase consisted of the presentation of a word pair for 5 s, followed by a 600 ms interval. All 50 word pairs were presented.

Test-feedback Phase. The trials consisted of a fixation cross followed by a cue word (e.g., tattoo). Upon cue presentation, participants had to respond with the corresponding target word (e.g., uncle). After the participant's response or after 3 s, the correct target word was presented next to the cue in blue font for 2 s, followed by a 300 ms delay. Participants were instructed to use the blue target word feedback to strengthen their memory for the word pairs. All 50 cue – target pairs were presented. The experimenter marked correct responses. A criterion of 50% correct recall should be reached before terminating this phase. Otherwise the trial sequence was repeated in a different fixed order until the criterion was reached, or until a maximum of three cycles was completed. Presentation times differed with each cycle (Cycle 2: cue 4 s and blue target 1 s; Cycle 3: cue 4.5 s and blue target 500 ms).

Hint-training Phase. The 15 cue words that were to be suppressed in the upcoming TNT phase were presented in a list on the computer screen for 1 min. The place of each cue word in the list was determined randomly. Participants were instructed to learn the words without thinking of the corresponding target words. Subsequently, participants were asked to identify the 15 to-be-suppressed words in a 30 item paper and pencil recognition test. This sequence (list presentation and recognition) was repeated until 100% recognition was reached or a maximum of 6 times.

Warm-up TNT Phase. For the purpose of practicing, 32 trials containing filler words were presented in the same fashion as in the TNT phase (see below). Participants were instructed to not respond to one of the filler cues (i.e., album), which was presented 8 times.

TNT Phase. There were 377 trials separated by 800 ms intervals. All trials started with the presentation of a fixation cross, followed by the 3 s presentation of a cue word[3] Two-thirds of the trials were Respond trials. The participants were instructed to say a target word out loud as quickly as possible upon presentation of the corresponding cue word. As soon as the voice key picked up a response, the cue disappeared from the screen. Not responding resulted in the presentation of the corresponding target for 500 ms in blue font. One-third of the trials were suppress trials. For these, participants were instructed to withhold their response and to not even think about the corresponding target word while they focused on the cue word. If the voice-key registered a response despite this instruction, a 2000 Hz beep was delivered through headphones. Twenty cue – target pairs served as respond pairs and twenty pairs were suppress pairs. Filler word pairs were used as respond trials to ensure a 2:1 ratio, such that responding was the dominant response. In addition, cue words in both the respond and suppress conditions were presented either 0, 1, 8 or 16 times. The cues that were never presented during the TNT phase served as baseline for later recall testing.

Test Phase. Two types of cued recall tasks were administered. Both tests consisted of 44 trials, separated by 400 ms intervals and starting with 4 filler trials. After presentation of a fixation cross, cues were presented for 4 s. Participants were instructed to respond to all cues, irrespective of previous instructions, by saying out loud the corresponding target. In the Same Probe (SP) test, participants were presented with all original cue words and were to recall the corresponding target word. In the independent probe (IP) test, participants were presented with a category cue combined with the first letter of the target word (e.g., relative – u_). Independent probes were unrelated to the original cues and were category words that had more than one exemplar starting with the same letter as the target word to reduce correct responding by guessing. The order of test administration (SP / IP or IP / SP) was counterbalanced across participants.

Samples 2–4

The tasks used in samples 2–4 were variations of the task used in sample 1. For each sample, only deviations from the procedure in sample 1 are described below.

Sample 2. The TNT task was programmed in E-prime (version 2, Schneider et al., 2002) and presented on a 22″ CRT monitor (Iiyama Vision Master Pro 513). A set of new cue – target pairs was constructed, consisting of 36 critical word pairs and 21 fillers (https://osf.io/me873/). Groups of 12 word pairs were rotated across instructions (baseline, respond, suppress), rendering three versions of the task.

A *drop-off phase* (Levy & Anderson, 2012) replaced the test-feedback phase of the original version. That is, rather than the 50% correct recall criterion, the drop-off version used the criterion of one correct response for each cue (i.e., 100% correct). The drop-off phase started with presenting all cues for 4 s at 800 ms intervals in a fixed random order. Response accuracy was registered by the experimenter on a serial response box (model #200a, Psychological Software Tools). Cue words that were correctly responded to were dropped from the list. Cue words with erroneous responses were randomly presented until the criterion was reached. Finally, all cue words were presented once more for 4 s in a fixed random order.

In the *warm-up TNT- and TNT-phases*, cue-words in green font denoted respond trials and cue words in red font signalled suppress trials. Thus, there was no need for a hint-training phase in this version. The warm-up TNT-phase contained 9 filler cues (5 respond and 4 suppress fillers). In the TNT phase, 12 respond and 12 suppress cues appeared 16 times, 12 baseline cues appeared 0 times. Six respond filler word pairs and 6 suppress fillers were used to make a total of 423 trials, with 52% respond trials and 48% suppress trials. The cue-words were presented in a fixed random order. Three 1-minute breaks were inserted after 108, 216, and 324 trials.

Questionnaires. A diagnostic questionnaire was administered during the warm-up TNT-phase, allowing the experimenter to detect whether the participant followed the instructions. Instructions were repeated when necessary. A TNT questionnaire was administered following the SP

and IP recall tasks. The questions assessed whether suppression during the TNT-phase had been successful and which strategies were used for target suppression. The general aim of this experiment was to replicate the basic suppression effect using the drop-off TNT and to explore its association with individual difference measures in attentional control and suppression ability.

Sample 3 (Wessel et al., 2005**).** The TNT task differed from the version in sample 1 in three ways. First, there were 0, 1, or 16 repetitions of both Respond and Suppress trials. Consequently, there were 30 cue-target pairs and 7 filler pairs in 6 versions of the task (see https://osf.io/8tcs2/). Secondly, cue words in green font denoted respond trials; cue words in red font signalled suppress trials. Thus, this version did not contain a hint-training phase. Third, participants had been told that memory suppression often results in paradoxical effects (cf. Wegner et al., 1987). The aim of the experiment was to examine differences in suppression between participants who scored high and low on a measure of dissociative tendencies. Time of testing had been varied, yet had not been included as a factor in the analyses reported in Wessel et al. (2005).

Sample 4 (Wessel et al., 2010**).** The TNT-task in sample 4 was identical to that used in sample 1, but was programmed in another programming language (Delphi). A 17 inch (1024 × 768 pixel) flatscreen monitor was used for stimulus presentation. The task was administered as part of a larger study (Wessel et al., 2010) on the impact of time of testing on several measures of cognitive control and intrusive memories in a trauma-film paradigm.

Data preparation and analysis

Overview

We analysed the data using R (Version 3.6.2; R Core Team, 2019b) and the R-packages *afex* (Version 0.26.0; Singmann et al., 2020), *ARTool* (Version 0.10.6; Kay & Wobbrock, 2019), *devtools* (Version 2.2.1; Wickham et al., 2019b), *dplyr* (Version 0.8.3; Wickham et al., 2019a), *foreign* (Version 0.8.72; R Core Team, 2019a), *ggplot2* (Version 3.2.1; Wickham, 2016), *ggpubr* (Version 0.2.4; Kassambara, 2019), *gridExtra* (Version 2.3; Auguie, 2017), *haven* (Version 2.2.0; Wickham & Miller, 2019), *knitr* (Version 1.26; Xie, 2015), *lme4* (Version 1.1.21; Bates et al., 2015), *lmerTest* (Version 3.1.0; Kuznetsova et al., 2017), *magrittr* (Version 1.5; Bache & Wickham, 2014), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *pacman* (Version 0.5.1; Rinker & Kurkiewicz, 2018), *papaja* (Version 0.1.0.9942; Aust & Barth, 2020), *reshape* (Version 0.8.8; Wickham, 2007), *summarytools* (Version 0.9.4; Comtois, 2019), *usethis* (Version 1.5.1; Wickham & Bryan, 2019), and *WRS2* (Version 1.0.0; Mair & Wilcox, 2019).

To test the hypotheses, we conducted a multiverse analysis for each sample separately. The analyses in the multiverse focused on various ways of handling violation of the normality assumption in parametric tests (i.e., outliers, skewed distributions) and the inclusion of covariates. Throughout the paper, we relate all *p*-values in the multiverse to the criterion of $\alpha = .05$, which is typically used for denoting statistical significance.

Participant exclusion at time of testing

Sample 1. Two participants in Experiment 1 failed to meet the 50% correct recall criterion in the practice phase of the TNT task. They were excluded from the experiment and replaced. Regarding Experiment 2, no information on whether additional participants were tested was retained.

Sample 2. One participant did not complete the TNT task. Another participant reported to have failed to learn four of the 12 suppress words. They were excluded from the experiment and replaced.

Sample 3. Four participants were excluded from the analyses because they were German rather than Dutch native speakers ($n = 3$) or had participated in experiment 2 ($n = 1$).

Sample 4. No information on whether additional participants were tested was retained.

Data preparation

To begin with, percentages correct recall were calculated for each combination of cue type, instruction and number of repetitions (e.g., SP – suppress – 16 times) by dividing the number of correctly recalled targets by the maximum number of items in that combination and multiplying by 100. Note that for sample 2, which used a drop-off phase ensuring that each item was recalled at least once prior to the TNT phase, the maximum items in a category reflected the number of items learned prior to suppression. Next, for each of the four samples, the percentages were exported to each of five datasets that differed in outlier treatment. We used trimming (i.e., outlier deletion) and winsorizing (i.e., substituting outliers with 1 unit higher or lower than the next highest or lowest value within the range of admissible values; i.e., non-outliers). In addition, we used two methods for determining the threshold for identifying outliers. First, we treated values below and above 1.5 times the interquartile range (IQR) as outliers (i.e., Q1- 1.5 IQR < x < Q3 + 1.5 IQR, where IQR = Q3-Q1). Second, we defined values below and above 2 Standard Deviations (SD) from the mean as outliers. Thus, the five datasets for each of the four samples were characterised by: (1) no treatment; (2) trimming based on IQR; (3) trimming based on SD; (4) winsorizing based on IQR; and (5) winsorizing based on SD.

Analytical approach

Hypothesis 1. We tested the suppression effect in each of the four samples (see Table 1 in pre-registration https://osf.io/vb78k/ for an overview). To that end, we used parametric analyses (dependent samples *t*-tests) to compare suppress-16 and baseline recall in each of the five datasets reflecting different outlier treatment. In addition, we conducted nonparametric (Wilcoxon signed-rank test) and

Table 1. Proportion of women, mean age, *N*, mean recall performance and hedges G effect sizes on same probe and independent probe tests per subsample. Standard deviations (SD) are between parentheses.

			Same Probe Recall						Independent Probe Recall					
Dataset	Women	Age	*N*	Baseline		Suppress-16		Hedges G	*N*	Baseline		Suppress-16		Hedges G
1.1	0.84	19.47	49	86.12	(17.42)	81.22	(19.33)	−0.27	49	82.04	(16.95)	82.04	(17.91)	0.00
1.2	0.86	19.55	44	88.64	(14.56)	85.91	(13.35)	−0.20	47	83.83	(14.83)	82.13	(17.81)	−0.10
1.3	0.86	19.55	44	88.64	(14.56)	85.91	(13.35)	−0.20	45	83.56	(14.95)	84.00	(15.72)	0.03
1.4	0.84	19.47	49	86.90	(15.65)	83.18	(15.05)	−0.24	49	82.82	(15.34)	82.04	(17.91)	−0.05
1.5	0.84	19.47	49	86.90	(15.65)	83.18	(15.05)	−0.24	49	82.82	(15.34)	82.82	(16.39)	0.00
2.1	0.64	21.22	36	95.04	(6.33)	90.74	(10.31)	−0.48	36	68.52	(15.32)	65.97	(15.86)	−0.16
2.2	0.63	21.20	35	95.61	(5.39)	91.90	(7.69)	−0.55	34	68.87	(15.67)	68.14	(13.37)	−0.05
2.3	0.65	21.18	34	96.08	(4.69)	92.16	(7.66)	−0.61	34	68.87	(15.67)	68.14	(13.37)	−0.05
2.4	0.64	21.22	36	95.14	(6.02)	91.41	(8.15)	−0.52	36	68.52	(15.32)	66.61	(14.47)	−0.13
2.5	0.64	21.22	36	95.31	(5.57)	91.41	(8.15)	−0.55	36	68.52	(15.32)	66.61	(14.47)	−0.13
3.1	0.93	19.69	68	85.59	(15.78)	80.88	(17.77)	−0.28	68	82.94	(18.04)	80.00	(18.28)	−0.16
3.2	0.95	19.72	64	86.25	(15.07)	83.75	(13.74)	−0.17	63	86.03	(13.74)	82.54	(15.86)	−0.24
3.3	0.95	19.72	64	86.25	(15.07)	83.75	(13.74)	−0.17	61	86.56	(13.53)	83.93	(14.06)	−0.19
3.4	0.93	19.69	68	85.87	(15.11)	82.29	(14.56)	−0.24	68	84.35	(14.70)	80.28	(17.48)	−0.25
3.5	0.93	19.69	68	85.87	(15.11)	82.29	(14.56)	−0.24	68	84.35	(14.70)	81.41	(15.27)	−0.20
4.1	NA	20.60	80	84.00	(18.93)	75.25	(20.68)	−0.44	80	78.00	(21.01)	78.75	(21.90)	0.03
4.2	NA	20.57	74	87.57	(14.69)	77.03	(18.93)	−0.62	80	78.00	(21.01)	78.75	(21.90)	0.03
4.3	NA	20.57	74	87.57	(14.69)	77.03	(18.93)	−0.62	75	79.47	(19.16)	82.13	(17.27)	0.15
4.4	NA	20.60	80	85.42	(16.03)	75.25	(20.68)	−0.55	80	78.00	(21.01)	78.75	(21.90)	0.03
4.5	NA	20.60	80	85.42	(16.03)	75.49	(20.14)	−0.54	80	78.47	(19.87)	79.71	(19.30)	0.06

Note: Dataset*:* 1rst digit = Sample Number, 2nd digit = Outlier Treatment (1 = none; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on Standard Deviation (SD); 4 = Winsorized based on IQR; 5 = Winsorized based on SD); Baseline = Suppress Instruction, 0 repetitions; Suppress-16 = Suppress Instruction, 16 repetitions in samples 1, 3, 4 and 12 repetitions in sample 2. In sample 1, the proportions of women and age are for the 32 participants in experiment 1

robust (Yuend test) analyses on the untreated but not the treated datasets because these techniques would remedy non-normality in their own right. Because of the directionality of the hypothesis, the analyses comparing suppress-16 with baseline trials were one-tailed. Furthermore, we conducted analyses including task version (reflecting different cue – target pairs in each cue-type/instruction/number of repetitions category), test order (SP/IP or IP/SP) and both task version and test order as covariates. More specifically, we conducted Repeated Measures Analyses of Variance (RM ANOVAs) controlling for these between participant factors on both the untreated and treated datasets in each sample. For the untreated datasets, we added nonparametric Aligned Rank Tranform (ART) ANOVAs and Robust RM ANOVAs. As far as we are aware, it is not possible to conduct a robust RM ANOVA with two between participants factors. Therefore Robust RM ANOVAs controlling for both version and task order were not included. See the pre-registration for a detailed list of analyses (https://osf.io/2drjg/).

All in all, we conducted 27 analyses on two dependent variables (i.e., SP and IP recall), totalling 54 analyses per sample. In addition, we included the grouping variables in samples 3 (DES group) and 4 (Time of testing) as covariates in the parametric and nonparametric RM ANOVAs, increasing the total to 90 tests in each of samples 3 and 4. All robust analyses used the default option of 20% trimming.

Hypothesis 2. To test the interaction between respond and suppress instructions and the number of repetitions we conducted a multiverse analysis in samples 1, 3 and 4. For each of the five datasets of samples 1 and 4, we conducted four 2 (Instruction: Respond, Suppress) x 4 (Repetitions: 0, 1, 8, 16) RM ANOVAs. The TNT task in sample 3 did not contain an 8-repetitions condition and 2 (instruction: respond, suppress) x 3 (repetitions; 0, 1, 16) ANOVAs were conducted. Next to a version without covariates, we conducted analyses including version, test order and both version and test order. Because it was not immediately obvious how to conduct ANOVAs with multiple within-participants factors using nonparametric and robust techniques, we concentrated on parametric analyses. In total, we conducted 5 (different datasets) x 4 (different ANOVAs) = 20 tests of the interaction effect. Analysing the SP and IP data separately resulted in 40 tests per sample.

Deviations from the Preregistered Analysis Plan. There were a few deviations from what was stated in the preregistration (see https://osf.io/2drjg/ and https://osf.io/kgj42/ for detailed lists of planned analyses for hypothesis 1 and 2, respectively). First, it was not feasible to run the Robust RM ANOVAs with more than one between participants factor. Therefore, contrary to what was planned, we do not report robust analyses controlling for dissociation level (sample 3) and time of testing (sample 4) in testing hypothesis 1.

Second, in the preregistration we stated that we would create separate datasets for SP and IP data to test hypothesis 2. For the actual analyses we combined the SP and IP data in each dataset rather than create separate datasets, resulting in 15 rather than 30 datasets. Note that the SP and IP data were analysed separately. In addition, for testing hypothesis 2 we refrained from including the additional between participants factors (dissociation group; time of day at testing) in samples 3 and 4, reducing the preregistered 80 analyses to 40 in samples 3 and 4.

Results

Descriptive statistics

Table 1 displays the descriptive statistics (e.g., sample size, gender, age), mean recall performance for the critical suppress trials, and the effect sizes (Hedges g) of the difference between baseline and suppress-16 trials.

Hypothesis 1: suppression effect

Sample 1

As can be seen in Table 2 and Figure 1 (upper panel), the multiverse of statistical results in sample 1 shows little support for the presence of a suppression effect in both the SP and IP data. For the SP data, only one of the 27 analyses showed a statistically significant difference between baseline and suppress-16 trials ($p = .045$). The remaining 25 analyses yielded *p*-values ranging from .074 to .976. It should be noted that the *p*-value of the robust repeated measures analyses controlling for version could not be produced due to too few observations per cell. For the IP data, none of the 27 analyses yielded a statistically significant difference between baseline and suppress-16 trials, with *p*-values ranging from .054 to >.999.

Sample 2

As shown in Table 3 and Figure 1 (lower panel), the multiverse of statistical results generally supports the presence of a suppression effect for the SP data in sample 2. Twenty-five of the 27 analyses showed a statistically significant difference between baseline and suppress-16 trials. The significant *p*-values ranged from 0.004–0.047 and the two non-significant *p*-values were 0.625 and 0.093. For the IP data, only one of the 27 analyses showed statistical significance ($p = .031$). The non-significant *p*-values ranged from .195 to .801.

Sample 3

As is evident in Table 4 and Figure 2 (upper panel), the multiverse of statistical results in sample 3 shows some support for the presence of a suppression effect in the SP data, but not in the IP data. For the SP data, 10 of the 27 analyses resulted in a statistically significant difference between baseline and suppress-16 trials (range of *p*-values: .014–.039), whereas the remaining 19 tests showed *p*-values ranging between .092 and .685. For the IP data, none of the 27 analyses showed a statistically significant difference between baseline and suppress-16 trials, with *p*-values ranging from .103 to .754. Controlling for dissociation group in those analyses allowing for multiple between participant factors (i.e., the parametric and nonparametric analyses; see Table 5 and Figure 2, lower panel), showed statistical significance in four of the 13 analyses of the SP data (range of *p*-values: .021–.035). None of the analyses controlling for dissociation group of the IP results showed statistical significance (with *p*-values ranging from .218 to .499).

Sample 4

As can be seen in Table 6 and Figure 3 (upper panel), the multiverse of statistical results in sample 4 generally supports the presence of a suppression effect for the SP, but not the IP data. For the SP data, all analyses except one were statistically significant ($p < .001$–.015; non-significant $p > .999$). For the IP data, none of the 27 analyses yielded statistical significance, with *p*-values ranging between .073 and .805. Controlling for the time of testing in those analyses allowing for multiple between participants factors (i.e., the parametric and nonparametric analyses; Table 7 and Figure 3, lower panel), rendered all 13 analyses of the SP results statistically significant (ranging between $p < .001$ and .015). None of the 13 analyses of IP data were statistically significant (with *p*-values between .274 and .804).

Table 2. *P*-values resulting from testing Hypothesis 1 in Sample 1.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
T-test	.074	.201	.201	.112	.112	.500	.314	.446	.412	.500
RMv	.169	.519	.519	.261	.261	.929	.382	.921	.722	.921
RMo	.153	.401	.401	.233	.233	.978	.672	.843	.841	.979
RMvo	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Wilc	.081					.503				
ARTv	.303					.972				
ARTo	.148					.724				
ARTvo	.045					>.999				
Yuen	.079					.500				
RRMv	NA					.054				
RRMo	.976					.175				

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "T-test" = Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed;Note: should be interpreted with caution); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order; "Yuen" = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo" = Robust RM ANOVA, 20% trimming, controlling for version and order, respectively. Due to too few observations per cell, the RMvo's are Not Available (NA) and the RRMv *p*-values should be interpreted with caution.

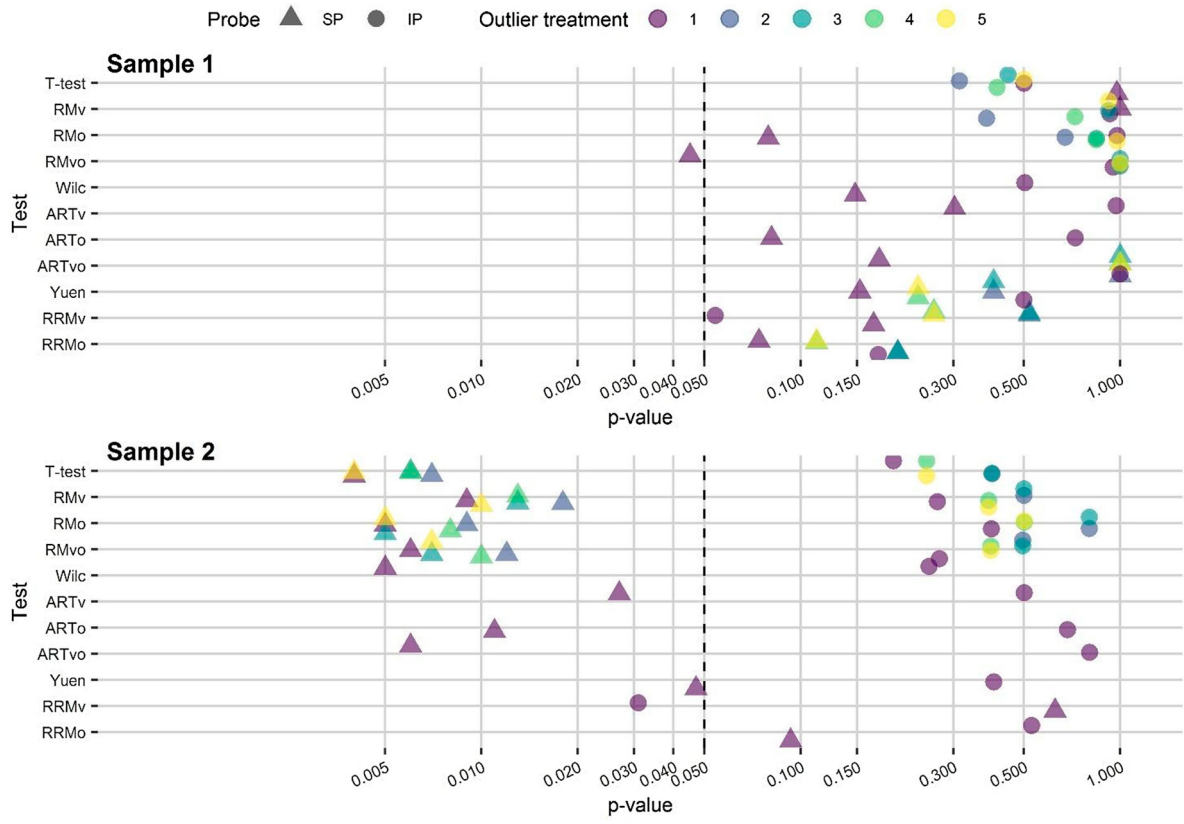

Figure 1. Distribution of *p*-values testing the suppression effect (hypothesis 1) in samples 1 (upper panel) and 2 (lower panel).

Note: SP = Same Probe; IP = Independent Probe; Outlier treatment: 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "T-test" = Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order; "Yuen" = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo"/"RRMvo" = Robust RM ANOVA, 20% trimming, controlling for version, order, and both version and order.

Hypothesis 2: instruction (respond / suppress) by number of repetitions interaction

Sample 1

The multiverse of statistical results consistently supports the presence of an instruction by repetitions interaction for the SP data, but not for the IP data. For the SP data, all 20 analyses showed a statistically significant interaction effect (p = .004–.025; see Figure 4, top row). For the IP data, none of the 20 interaction effects were statistically significant, with *p*-values ranging from .744 to .902 (see Figure 5, top row). Table 8 gives an overview of the relevant *p*-values.

Table 3. *P*-values resulting from testing Hypothesis 1 in Sample 2.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
t-test	.004	.007	.006	.006	.004	.195	.397	.397	.248	.248
RMv	.009	.018	.013	.013	.010	.268	.500	.500	.388	.388
RMo	.005	.009	.005	.008	.005	.395	.801	.801	.501	.501
RMvo	.006	.012	.007	.010	.007	.272	.495	.495	.394	.394
Wilc	.005					.252				
ARTv	.027					.501				
ARTo	.011					.684				
ARTvo	.006					.803				
Yuen	.047					.402				
RRMv	.625					.031				
RRMo	.093					.529				

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "T-test" = Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed; Note: should be interpreted with caution); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order; "Yuen" = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo"/"RRMvo" = Robust RM ANOVA, 20% trimming, controlling for version, order, and both version and order.

Table 4. *P*-values resulting from testing Hypothesis 1 in Sample 3.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
T-test	.014	.092	.092	.031	.031	.148	.103	.172	.061	.121
RMv	.020	.160	.160	.050	.050	.351	.266	.363	.147	.289
RMo	.029	.214	.214	.062	.062	.296	.222	.392	.124	.243
RMvo	.021	.192	.192	.050	.050	.352	.288	.420	.149	.290
Wilc	.016					.173				
ARTv	.020					.173				
ARTo	.183					.361				
ARTvo	.039					.537				
Yuen	.030					.114				
RRMv	.147					.754				
RRMo	.685					.365				

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. With regard to the different analyses, "T-test"= Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, respectively controlling for version, order, and both; "Wilc" = Wilcoxon signed-rank test (one-tailed; Note: should be interpreted with caution); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, respectively controlling for version, order, and both version and order; "Yuen' = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo"/"RRMvo" = Robust RM ANOVA, 20% trimming, controlling for version, order, and both version and order.

Sample 3

The multiverse of statistical results shows a robust instruction by interaction effect for the SP, but not for the IP data. For the SP test, all 20 interaction effects were statistically significant (all with $p < 0.001$; see Figure 4, middle row). By contrast, for the IP data none of the 20 interaction effects were significant, with *p*-values ranging from .524 to .921 (see Figure 5, middle row). For an overview of all *p*-values, see Table 9.

Sample 4

The multiverse of statistical results in sample 4 shows a robust instruction by repetitions interaction effect for the SP data, but not for the IP data. For the SP data, all 20 interaction effects

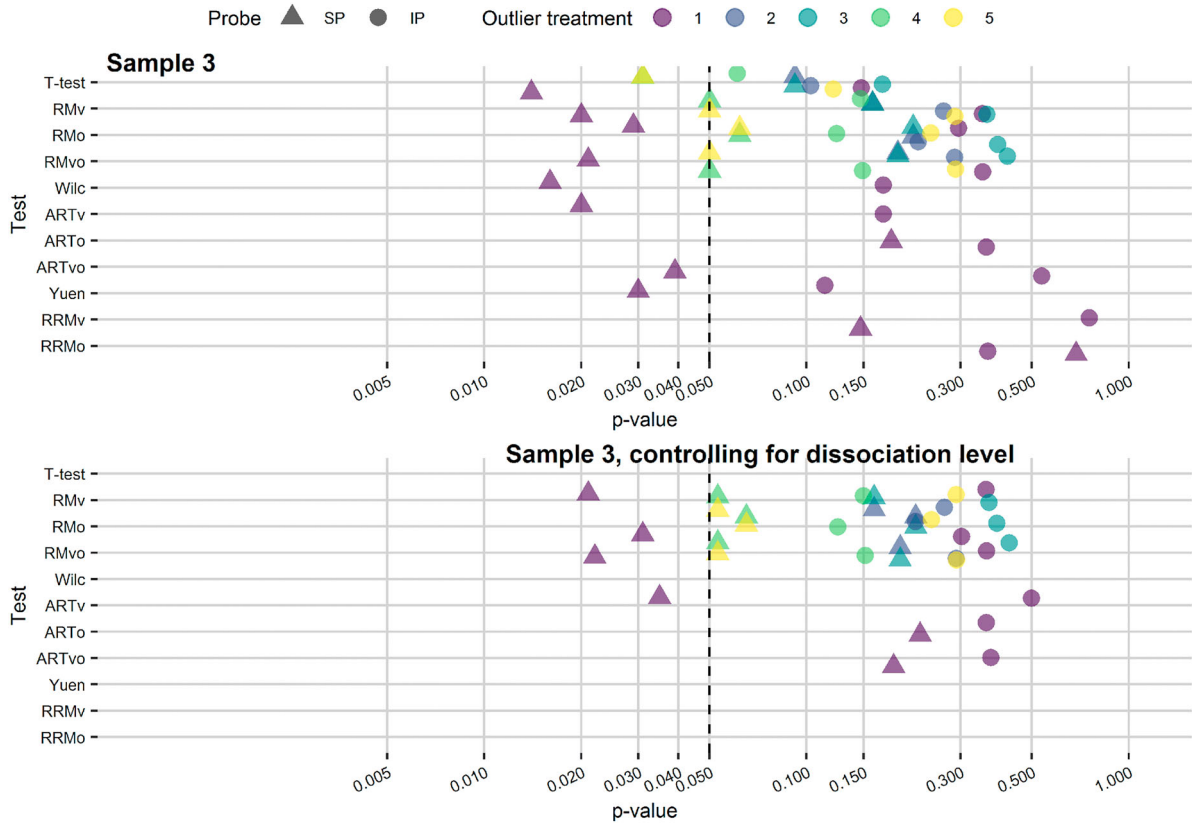

Figure 2. Distribution of *p*-values testing the suppression effect (hypothesis 1) in sample 3.

Note: SP = Same Probe; IP = Independent Probe; Outlier treatment: 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "T-test" = Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order; "Yuen" = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo"/"RRMvo" = Robust RM ANOVA, 20% trimming, controlling for version, order, and both version and order.

Table 5. *P*-values resulting from testing Hypothesis 1 in Sample 3, controlling for dissociation level.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
RMv	.021	.162	.162	.053	.053	.360	.268	.368	.150	.291
RMo	.031	.218	.218	.065	.065	.303	.218	.389	.125	.244
RMvo	.022	.195	.195	.053	.053	.362	.291	.425	.152	.292
ARTv	.035					.499				
ARTo	.225					.361				
ARTvo	.186					.373				

Note: The multiverse analysis in Sample 3, controlled for dissociation level where possible. The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed; Note: should be interpreted with caution); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order.

were statistically significant (all *p*s < 0.001; see Figure 4, bottom row). For the IP data, six of the 20 interaction effects were statistically significant, with *p*-values ranging from .018 to .038 (see Figure 5, bottom row). The remaining 14 analyses showed *p*-values ranging from .063 to .112. An overview of the relevant *p*-values is given in Table 10.

Discussion

We performed multiverse analysis on the data obtained with variations of the TNT task in four samples with two goals. The first goal was to extend the TNT literature with unpublished replications of Anderson and Green's (2001) and Levy and Anderson's (2012) methods (in sample 1 and 2, respectively). The second goal was to compare the results from systematically varying choices in analytic strategies that may be invited by these data. Specifically, we concentrated on handling outliers as well as covarying factors in the experimental design (i.e., recall test order, word list version). We tested two pre-registered hypotheses. The first hypothesis concerned the suppression effect, that is, we expected below baseline performance for recall of suppressed items in both SP and IP recall tests. As for the SP data, in two samples (samples 2 and 4) all analyses but one yielded *p*-values that were below the conventional threshold ($\alpha = .05$) for statistical significance. The observed effect sizes of the difference between baseline and suppress-16 trials were in the medium range in these samples. However, in two samples the observed effect sizes of the difference were small. Either one (sample 1) or a minority (10 out of 27; sample 3) of the analyses were statistically significant at the $\alpha = .05$ level. As for IP tests, the observed effect sizes of the difference between baseline and suppress-16 trials varied between zero and small. Only one of the analyses across the four samples displayed a suppression effect that was statistically significant by conventional standards. Thus, across samples, the evidence for a suppression effect was mixed when same probes were used. The evidence was inconclusive altogether when the final recall test contained independent cues.

The second pre-registered hypothesis concerned the pattern across varying numbers of trial repetitions that is typically reported in the literature using Anderson and Green's (2001) original method. This pattern consists of an increase in recall performance with an increasing number of respond trials and a decrease in recall performance with an increasing number of suppress trials. We expected this to occur for SP as well as IP recall tests. The results show that without exception, the instruction (i.e.,

Table 6. *P*-values resulting from testing Hypothesis 1 in Sample 4.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
T-test	.001	< .001	< .001	< .001	< .001	.403	.403	.167	.403	.328
RMv	.001	< .001	< .001	< .001	< .001	.801	.801	.252	.801	.645
RMo	.001	< .001	< .001	< .001	< .001	.805	.805	.328	.805	.655
RMvo	.001	< .001	< .001	< .001	< .001	.799	.799	.247	.799	.644
Wilc	.001					.691				
ARTv	.004					.338				
ARTo	.008					.485				
ARTvo	.001					.366				
Yuen	.001					.294				
RRMv	.015					.220				
RRMo	> .999					.073				

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. With regard to the different analyses, "T-test" = Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed; Note: should be interpreted with caution); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order; "Yuen" = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo"/"RRMvo" = Robust RM ANOVA, 20% trimming, controlling for version, order, and both version and order. RRMo should be interpreted with caution.

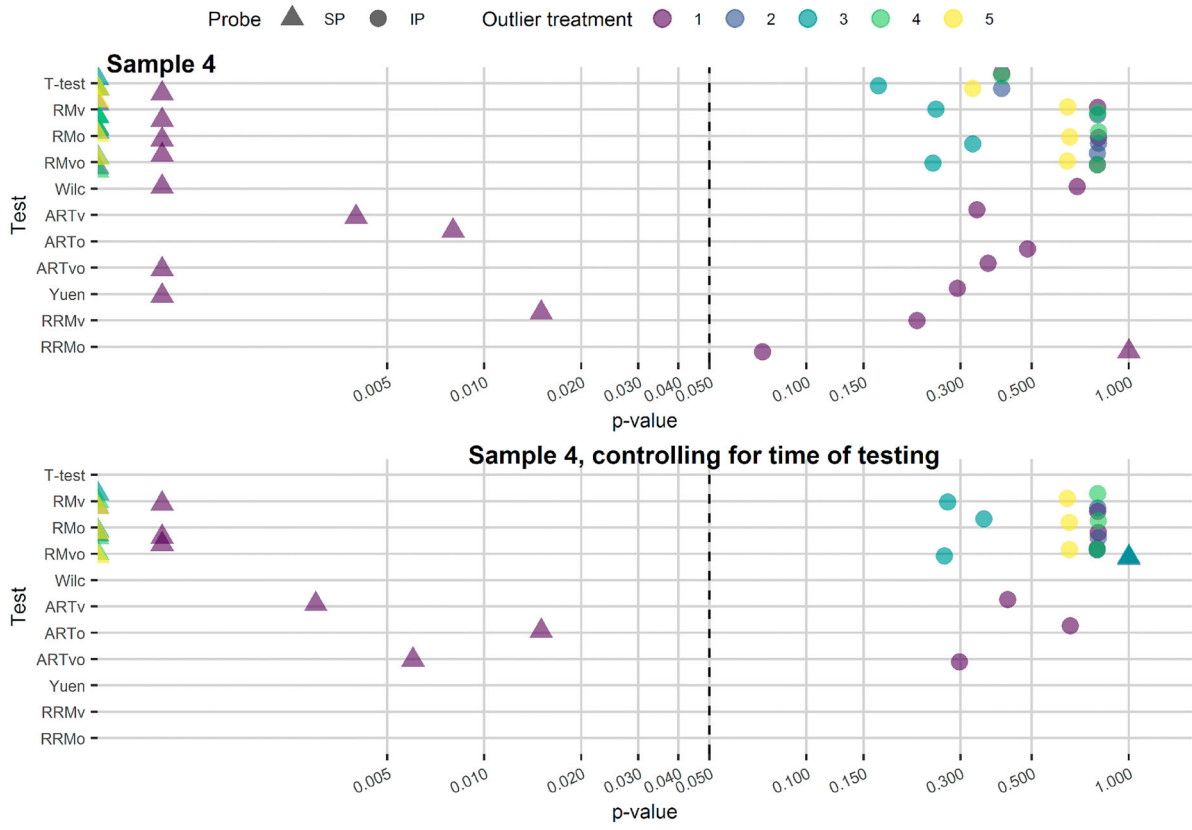

Figure 3. Distribution of *p*-values testing the suppression effect (hypothesis 1) in sample 4.

Note: SP = Same Probe; IP = Independent Probe; Outlier treatment: 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "T-test" = Dependent samples *t*-test (one-tailed); "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both version and order; "Wilc" = Wilcoxon signed-rank test (one-tailed); "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, controlling for version, order, and both version and order; "Yuen" = Yuen's 20% trimmed means test for dependent samples (one-tailed), "RRMv"/"RRMo"/"RRMvo" = Robust RM ANOVA, 20% trimming, controlling for version, order, and both version and order

respond or suppress) by repetitions (0, 1, 8, 16 times) interactions for the SP data were statistically significant. However, the interaction effects did not reach statistical significance for IP test recall in samples 1 and 3. The analyses in sample 4 deviated from that pattern in that a minority (six out of 20) of the analyses of the IP data yielded statistical significance.

Overall, within samples the multiverse analysis consistently supported an interpretation of the results in terms of either statistical significance or non-significance. There were 2 exceptions. First, the multiverse for the SP suppression effect in sample 3 showed a mixed pattern. Here, the statistically significant effects predominantly emerged in the untreated sample. It should be noted that Wessel et al. (2005) reported one of the significant outcomes in this multiverse as evidence for a suppression. However, the multiverse suggests that different statistical choices, such as trimming or winsorizing the data would have resulted in a different interpretation. This implies that Wessel et al.'s (2005) firm conclusion should be tempered. The other exception is the mixed pattern of results for the instruction by repetitions interaction in the

Table 7. *P*-values resulting from testing Hypothesis 1 in Sample 4, controlling for time of testing.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
RMv	.001	< .001	< .001	< .001	< .001	.800	.800	.274	.800	.644
RMo	.001	< .001	< .001	< .001	< .001	.804	.804	.355	.804	.654
RMvo	.001	NA	NA	< .001	< .001	.798	.798	.268	.798	.654
ARTv	.003					.421				
ARTo	.015					.658				
ARTvo	.006					.298				

Note: The multiverse analysis in Sample 4, controlled for time of testing where possible. The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "RMv"/"RMo"/"RMvo" = Repeated Measure (RM) ANOVA, controlling for version, order, and both; "ARTv"/"ARTo"/"ARTvo" = Aligned Rank Transform (ART) ANOVA, respectively controlling for version, order, and both version and order. NA = Not Available.

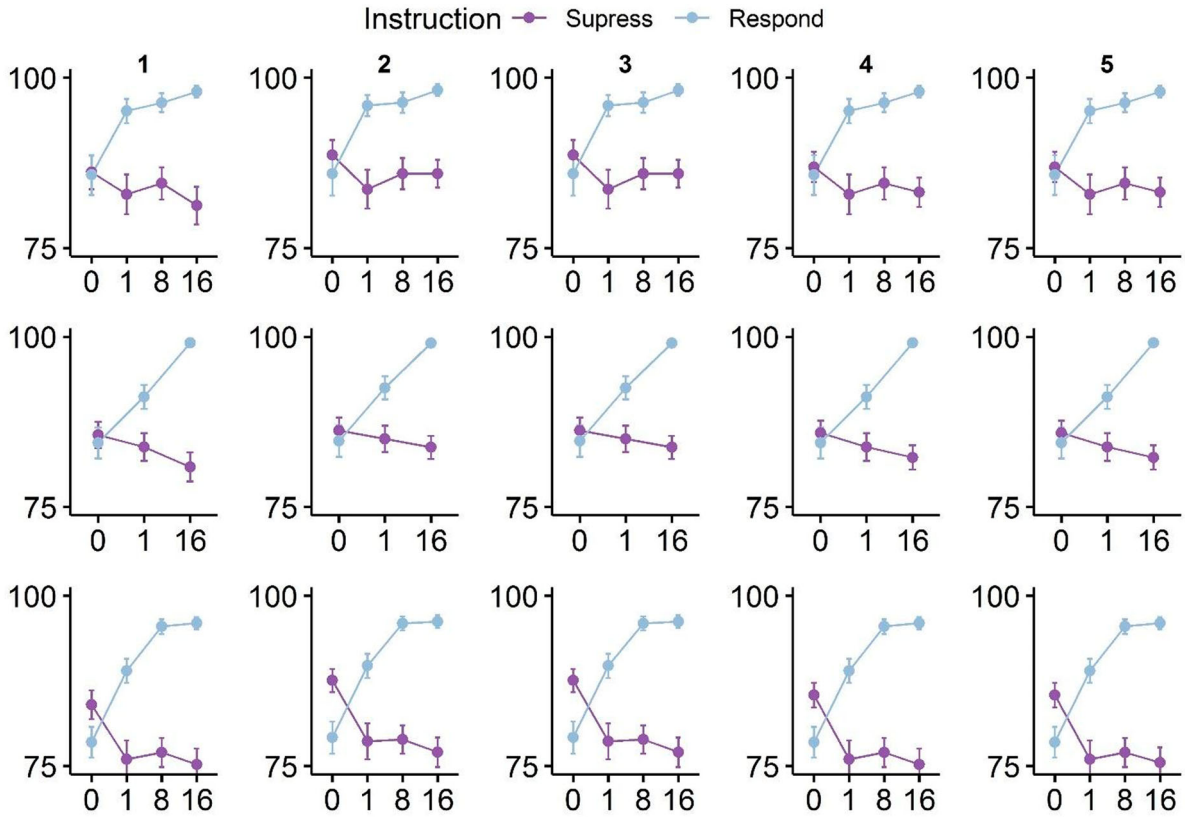

Figure 4. Plots of percentage recall in the instruction (suppress or respond) by repetition interactions (hypothesis 2) in the SP data.
Note: The graphs in the top row come from sample 1, whereas the graphs in the middle and bottom rows come from samples 3 and 4, respectively.

IP results in sample 4. Inspection of Figure 5 suggests that the significant effects were mainly due to an increase in performance under respond instructions and not so much to the decline of performance under suppress instructions.

Across samples, the SP suppression effect shows a mixed pattern (present in samples 2 and 4, inconclusive in sample 1, and mixed in sample 3). This pattern adds to the broader literature containing both positive as well as inconclusive findings, albeit the latter can be found to a lesser extent. Such a pattern is to be expected if relatively modest sample sizes are used for studying modest effects. For example, Anderson and Green's (2001) initial experiments relied on 32 participants. This sample size yields 80% power for detecting statistical significance at the .05 level for a medium effect size (Cohen's $d = 0.45$) in a one-tailed dependent *t*-test. This means that in a large series of similar experiments, 20% of the studies would yield a statistically non-significant result. The experiments in the present paper that followed the original TNT procedure most closely (samples 1 and 4) yielded more power than Anderson and Green's (2001) experiments for detecting a Cohen's *d* of 0.45. Accordingly, the multiverse for sample 4 (99% power) contained predominantly statistically significant results. Yet, the multiverse for sample 1 (93% power) yielded relatively few significant results. It should be noted that even with high power, there is a (small) probability of obtaining a statistically non-significant result. However, another and perhaps more likely possibility is that the true effect size of the SP suppression effect is small. In that case, relatively small sample sizes can be expected to yield an even higher percentage of inconclusive results. Put differently, our samples would have been too small to detect a small effect size with a high probability of finding a true effect. To illustrate, detecting a Cohen's $d = 0.2$ with 95% power and an $\alpha = .05$ in a one-tailed dependent t-test would require 272 participants.

Although the true effect size of the SP suppression effect is unknown, there are reasons to believe that Anderson and Green's (2001) original procedure yields small rather than medium effect sizes. Since the initial publication, the TNT procedure has evolved. It is plausible that such procedural advancement results in less error variance. Some changes in the TNT task appeared relatively early in the literature and were applied in some of the experiments in the present paper. These changes were using a 100% rather than 50% correct criterion for the initial study phase (Levy & Anderson, 2012; sample 2) and employing cue word colour to denote the type of trial rather than instructing participants to memorise the No-Think cues (Anderson et al., 2004; Levy & Anderson, 2012; samples 2 and 3). It is noteworthy that the multiverse in the sample in which both changes were adopted (sample 2) contained

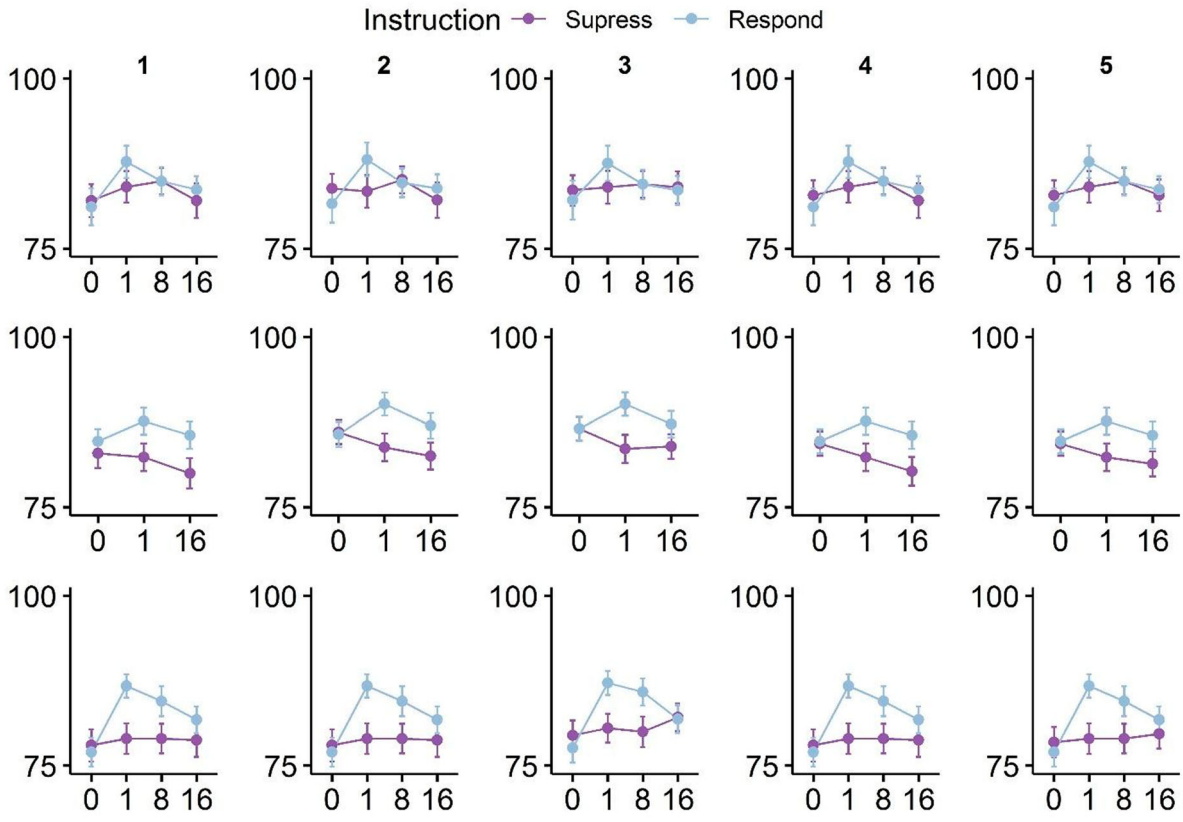

Figure 5. Plots of percentage recall in the instruction (suppress or respond) by repetition interactions (hypothesis 2) in the IP data.
Note: The graphs in the top row come from sample 1, whereas the graphs in the middle and bottom rows come from samples 3 and 4, respectively.

Table 8. *P*-values resulting from testing Hypothesis 2 in Sample 1.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
TwRM	.001	< .001	< .001	< .001	< .001	.764	.490	.726	.677	.681
TwRMv	< .001	< .001	< .001	< .001	< .001	.470	.138	.219	.374	.403
TwRMo	.001	< .001	< .001	< .001	< .001	.773	.496	.738	.684	.689
TwRMvo	< .001	NA	NA	< .001	< .001	.485	NA	NA	.384	.413

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment of the dataset for that specific outcome variable, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. With regard to the different analyses, "TwRM" = 2 (Instruction: Respond vs Suppress) x 4 (Repetitions: 0, 1, 8, 16) Repeated Measures ANOVAs, with no controls; "TwRMv" = TwRM controlling for version; "TwRMo" = TwRM controlling for order; "TwRMvo" = TwRM controlling for both version and order. NA = Not Available.

Table 9. *P*-values resulting from testing Hypothesis 2 in Sample 3.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
TwRM	< .001	< .001	< .001	< .001	< .001	.498	.164	.190	.265	.324
TwRMv	< .001	< .001	< .001	< .001	< .001	.532	.172	.163	.261	.324
TwRMo	< .001	< .001	< .001	< .001	< .001	.494	.128	.120	.258	.315
TwRMvo	< .001	< .001	< .001	< .001	< .001	.525	.116	.077	.250	.312

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment of the dataset for that specific outcome variable, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. With regard to the different analyses, "TwRM" = 2 (Instruction: Respond vs Suppress) x 4 (Repetitions: 0, 1, 8, 16) Repeated Measures ANOVAs, with no controls; "TwRMv" = TwRM controlling for version; "TwRMo" = TwRM controlling for order; "TwRMvo" = TwRM controlling for both version and order.

Table 10. *P*-values resulting from testing Hypothesis 2 in Sample 4.

	Same Probe					Independent Probe				
Test	1	2	3	4	5	1	2	3	4	5
TwRM	< .001	< .001	< .001	< .001	< .001	.110	.110	.037	.110	.063
TwRMv	< .001	< .001	< .001	< .001	< .001	.072	.072	.018	.072	.037
TwRMo	< .001	< .001	< .001	< .001	< .001	.112	.112	.038	.112	.064
TwRMvo	< .001	< .001	< .001	< .001	< .001	.074	.074	.018	.074	.038

Note: The numbers 1, 2, 3, 4, and 5 below "Same Probe" and "Independent Probe" refer to the type of outlier treatment of the dataset for that specific outcome variable, where 1 = No outlier treatment; 2 = Trimmed based on Inter Quartile Range (IQR); 3 = Trimmed based on SD; 4 = Winsorized based on IQR; 5 = Winsorized based on SD. "TwRM" = 2 (Instruction: Respond vs Suppress) x 4 (Repetitions: 0, 1, 8, 16) Repeated Measures ANOVAs, with no controls; "TwRMv" = TwRM controlling for version; "TwRMo" = TwRM controlling for order; "TwRMvo" = TwRM controlling for both version and order.

predominantly statistically significant results, whereas the multiverse based on sample 3 did not. The magnitude of the standard deviations in sample 2 was about half of those in the other samples, suggesting more precise measurement. Thus, especially for sample 1, the number of participants may have been too small for detecting a small effect size.

A procedural refinement that was not incorporated in the present experiments concerns specifying suppression strategies in the participant instructions (see Benoit & Anderson, 2012; Bergström et al., 2009). Similar to other initial TNT studies, our participants were instructed to prevent target words from entering awareness, but these instructions were silent on what to do or not to do in order to achieve this. The assumption was that such unspecified instructions would prompt participants to directly stop the retrieval of target words. However, Levy and Anderson (2008) scrutinised participant reports and found a wide variety of suppression strategies, many of which would classify as "thinking of an alternative thought" (p. 632; see also Hertel & Calcaterra, 2005, on the role of thought substitution). Accordingly, researchers (e.g., Benoit & Anderson, 2012; Bergström et al., 2009) began instructing participants more explicitly to either block targets without thinking of alternatives (i.e., direct suppression instructions) or avoid targets by using other thoughts (i.e., thought substitution). Stramaccia et al.'s (2019, preprint) unpublished meta-analysis suggests that studies using specific instructions yield larger effect sizes than a general instruction such as used in the present studies.

As for IP recall, the results were predominantly inconclusive, within as well as across samples. Thus, our data are not in line with the idea that suppression impairs recall triggered by cues that are unrelated to the study context. The consistent failure in the present samples to replicate earlier findings of IP suppression raises problems for the assumption that blocking memories from awareness results in a deactivation of the memory representation itself (Anderson & Huddleston, 2012). Showing suppression with an IP test is important as the test was devised to separate non-inhibitory (e.g., interference, unlearning) from inhibitory accounts (Anderson & Green, 2001). The several mechanisms that potentially underlie forgetting cannot be inferred from a SP effect alone. Thus, especially the combination of finding statistically significant suppression effects in SP but not IP recall (samples 2 and 4) is problematic for inhibition theory. In addition, inconclusive IP suppression effects cast doubts on the idea that the TNT paradigm is an appropriate model for repression (Anderson & Green, 2001; Conway, 2001; Lambert et al., 2010).

There are some methodological issues that deserve attention. To begin with, it should be noted that the present multiverses were not exhaustive in terms of analytic choices. For example, we did not add data transformations (e.g., logarithmic transformation) to deal with skewness. In addition, some choices may have been less optimal, such as performing the 20% trimmed robust tests. These analyses discard the extremes of both ends of the distribution, without taking the distances of these values to the mean or median of the distribution into account. Nevertheless, these analyses should be regarded within the whole of the multiverse analysis. The idea is that there is no single best approach.

Relatedly, one might wonder about the value of multiverse analysis. Overall, our analyses showed quite some variety in *p*-values across the parallel "universes" that we defined. This suggests that caution is warranted. That is, it is advisable for future researchers to be aware that a different methodological or analytical approach could lead to different outcomes. Conclusions that are based on one specific set of choices may not be robust, in that they may not hold across the majority of plausible ways to detect a suppression effect. Therefore, rather than merely justifying one specific set of modelling choices, researchers might want to consider adding a multiverse analysis as supplementary material.

Furthermore, especially with regard to the negative findings for an IP suppression effect, we emphasise that statistical non-significance cannot be interpreted as evidence that a true effect does not exist. Although we endeavoured to replicate Anderson and Green's (2001) and Levy and Anderson's (2012) methods as closely as possible, it remains possible that some unknown methodological issues might have been responsible for our failure to find IP suppression. Alternatively, it may be that the true effect size of IP suppression is much smaller than that of SP suppression. Perhaps procedural refinements such as specific direct suppression instructions will yield more precise estimates of the IP suppression effect in the long run. Clearly, more research is needed. Judging from the

published literature, it seems that IP tests are less often administered than SP tests (Stramaccia et al., 2019, preprint). In addition, we note that sometimes IP and SP data are lumped together in the analyses, obscuring the separate effects of either method (e.g., Mecklinger et al., 2009; van Schie & Anderson, 2017). All in all, the literature on inhibition theory would benefit from more reports of TNT studies in which IP suppression effects are consistently assessed and reported. In particular, large-scale replication studies, designed to detect small effect sizes with a low probability of false negative findings (e.g., 95% power) are warranted.

It is important to consider the implications of confirming a small effect size for IP suppression. From a theoretical perspective, such an outcome would support an inhibitory account. Whether that would speak to the repression of real-life traumatic memories is a different matter. It has been pointed out (Barnier, 2012; Kihlstrom, 2002) that TNT experiments conducted in the laboratory are far removed from the forgetting of complex traumatic experiences they seek to model. Whereas the debate on repressed memories focused on having no memory of trauma whatsoever, forgetting in TNT research represents only a small percentage of previously studied stimuli (Kihlstrom, 2002). In addition, it is questionable whether the innocuous (often word-) stimuli in TNT studies are ecologically valid, that is, whether they represent real-life traumatic experiences adequately (Barnier, 2012). Even though SP suppression effects have been reported for more ecologically valid stimuli such as aversive pictures (e.g., Catarino et al., 2015; Depue et al., 2013) or autobiographical memories (e.g., Noreen & MacLeod, 2013), those studies typically do not report on IP suppression effects. Indeed, devising independent probes for such targets would be challenging, if not impossible. This may imply that the more successful authors are in building ecological validity into their TNT studies, the less informative their results may be for theories on inhibition and repression.

Conclusion

Across four samples, our multiverse analysis yielded a mixed pattern of results for the forgetting of targets when recall was tested with familiar cues (i.e., SP test). Such a pattern fits with the common practice of conducting studies with less than perfect power to detect true effects. In contrast, the results regarding suppression effects with independent probes were inconclusive overall. This is problematic for inhibition theory and its implications for repression as a mechanism of forgetting. Multiple studies using methods allowing for more precise measurement than in the present studies are warranted. These should establish whether estimates of the true effect size for IP suppression effects are large enough to be of interest. Until then, it seems prudent to refrain from interpreting suppression induced forgetting as evidence for inhibition.

Notes

1. There were 8 studies that overlapped between Anderson and Huddleston (2012) and Stramaccia et al. (2019, preprint).
2. No data were collected on gender and age in experiment 2.
3. In Anderson and Green's (2001) original procedure, cue-presentation times had been 4 s. However, additional work in their lab had indicated that 3 s yielded similar results, with the advantage of a shorter TNT phase (B. Levy, personal communication, October 30, 2001). In addition, rather than 400 ms intertrial intervals (Anderson & Green, 2001) we used 800 ms. Pilot testing suggested that the pronunciation of some targets took longer than 400 ms and inadvertently triggered the offset of the subsequent cue.

Acknowledgements

Data collection in samples 2 and 4 was supported by a VIDI grant (452-03-329) of the Foundation for Behavioural and Educational Sciences of the Netherlands Organization for Scientific Research (NWO) awarded to Ineke Wessel. We thank the research support staff at Maastricht University and Eize Hoekstra and Bert Hoekzema (Research support BSS, University of Groningen) for programming the TNT tasks. We are grateful to Wolf-Gero Lange, Sippie Overwijk, Marjel Buiter, Jorge Tendeiro and Gert Stulp for their involvement in various stages of the project and to all students involved in data collection for their help. We are indebted to Michael C. Anderson and Benjamin J. Levy for sharing their original materials. IW and AREZ designed the initial research questions. AREZ collected the data in sample 2. VEH and IW wrote the preregistration, with feedback from AREZ and CJA. The data analyses were conducted by VEH with feedback from CJA. The sections in the paper were written by IW (introduction, discussion); AREZ (method) and VEH (results, with feedback from CJA). All authors provided feedback on draft versions and all approved of the final version of the manuscript. *Author Information*: Data for samples 1 and 3 were collected when IW was at the Department of Experimental Psychology, Maastricht University, The Netherlands; data for samples 2 and 4 were collected after IW moved to the University of Groningen. AREZ conducted the study in sample 2 for her thesis submitted to the University of Groningen as part of the requirements for her MSc degree in Clinical Psychology.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Nederlandse Organisatie voor Wetenschappelijk Onderzoek [grant number VIDI 452-03-329].

Data availability statement

All study materials, pre-registration, data, and code can be found at https://osf.io/qgcy5/.

ORCID

Ineke Wessel http://orcid.org/0000-0001-6312-376X
Casper J. Albers http://orcid.org/0000-0002-9213-6743
Anna Roos E. Zandstra http://orcid.org/0000-0002-0583-1829
Vera E. Heininga http://orcid.org/0000-0003-0889-8524

References

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, *410*(6826), 366–369. https://doi.org/10.1038/35066572

Anderson, M. C., & Huddleston, E. (2012). Towards a cognitive and neurobiological model of motivated forgetting. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, *58*, 53–120. https://doi.org/10.1007/978-1-4614-1195-6_3

Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., Glover, G. H., & Gabrieli, J. D. E. (2004). Neural systems underlying the suppression of unwanted memories. *Science*, *303*(5655), 232–235. https://doi.org/10.1126/science.1089504

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics*. https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r*. https://CRAN.R-project.org/package=magrittr

Barnier, A. J. (2012). Memory, ecological validity and a barking dog. *Memory Studies*, *5*(4), 351–359. https://doi.org/10.1177/1750698012461243

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. https://CRAN.R-project.org/package=Matrix

Benoit, R. G., & Anderson, M. C. (2012). Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron*, *76*(2), 450–460. https://doi.org/10.1016/j.neuron.2012.07.025

Bergström, Z. M., de Fockert, J. W., & Richardson-Klavehn, A. (2009). ERP and behavioural evidence for direct suppression of unwanted memories. *NeuroImage*, *48*(4), 726–737. https://doi.org/10.1016/j.neuroimage.2009.06.051

Beringer, J. (1996). *Experimental run time system (erts)*. [Computer software]. Berisoft.

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. https://doi.org/10.1097/00005053-198612000-00004

Brewin, C. R., & Andrews, B. (2014). Why it is scientifically respectable to believe in repression: A response to Patihis, Ho, Tingen, Lilienfeld, and Loftus (2014). *Psychological Science*, *25*(10), 1964–1966. https://doi.org/10.1177/0956797614541856.

Bulevich, J. B., Roediger, H. L., Balota, D. A., & Butler, A. C. (2006). Failures to find suppression of episodic memories in the Think/No-think paradigm. *Memory & Cognition*, *34*(8), 1569–1577. https://doi.org/10.3758/BF03195920.

Catarino, A., Küpper, C. S., Werner-Seidler, A., Dalgleish, T., & Anderson, M. C. (2015). Failing to forget: Inhibitory-control deficits compromise memory suppression in Posttraumatic Stress Disorder. *Psychological Science*, *26*(5), 604–616. https://doi.org/10.1177/095679761556988

Comtois, D. (2019). *Summarytools: Tools to quickly and neatly summarize data*. https://CRAN.R-project.org/package=summarytools

Conway, M. A. (2001). Repression revisited. *Nature*, *410*(6826), 319–320. https://doi.org/10.1038/35066672

Depue, B. E., Ketz, N., Mollison, M. V., Nyhus, E., Banich, M. T., & Curran, T. (2013). ERPs and neural oscillations during volitional suppression of memory retrieval. *Journal of Cognitive Neuroscience*, *25*(10), 1624–1633. https://doi.org/10.1162/jocn_a_00418

Dutta, A. (1995). Experimental runtime system: Software for developing and running reaction time experiments on IBM-compatible pcs. *Behavior Research Methods, Instruments, & Computers*, *27*(4), 516–519. https://doi.org/10.3758/BF03200453

Erdelyi, M. H. (2006). The unified theory of repression. *Behavioral and Brain Sciences*, *29*(5), 499–511. https://doi.org/10.1017/S0140525X06009113

Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis–a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466. https://doi.org/10.1511/2014.111.460

Heininga, V. E., Oldehinkel, A. J., Veenstra, R., & Nederhof, E. (2015). I just ran a thousand analyses: Benefits of multiple testing in understanding equivocal evidence on gene-environment interactions. *Plos One*, *10*(5), e0125383. https://doi.org/10.1371/journal.pone.0125383

Hendriks, J., Hofstee, W., & Raad, B. (1999). The five-factor Personality Inventory (FFPI). *Personality and Individual Differences*, *27*(2), 307–325. https://doi.org/10.1016/S0191-8869(98)00245-1

Hertel, P. T., & Calcaterra, G. (2005). Intentional forgetting benefits from thought substitution. *Psychonomic Bulletin & Review*, *12*(3), 484–489. https://doi.org/10.3758/BF03193792

Holmes, D. S. (1990). The evidence for repression: An examination of sixty years of research. In J. L. Singer (Ed.), *Repression and dissociation* (pp. 85–102). University of Chicago Press.

Horne, J. A., & Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, *4*, 97–110.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Kassambara, A. (2019). *Ggpubr: 'Ggplot2' based publication ready plots*. https://CRAN.R-project.org/package=ggpubr

Kay, M., & Wobbrock, J. O. (2019). *ARTool: Aligned rank transform for nonparametric factorial anovas*. https://doi.org/10.5281/zenodo.594511

Kihlstrom, J. F. (2002). No need for repression. *Trends in Cognitive Sciences*, *6*(12), 502. https://doi.org/10.1016/S1364-6613(02)02006-5

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lambert, A. J., Good, K. S., & Kirk, I. J. (2010). Testing the repression hypothesis: Effects of emotional valence on memory suppression in the Think–No think task. *Consciousness and Cognition*, *19*(1), 281–293. https://doi.org/10.1016/j.concog.2009.09.004.

Levy, B. J., & Anderson, M. C. (2008). Individual differences in the suppression of unwanted memories: The executive deficit hypothesis. *Acta Psychologica*, *127*(3), 623–635. https://doi.org/10.1016/j.actpsy.2007.12.004

Levy, B. J., & Anderson, M. C. (2012). Purging of memories from conscious awareness tracked in the human brain. *Journal of Neuroscience*, *32*(47), 16785–16794. https://doi.org/10.1523/JNEUROSCI.2640-12.2012

Loftus, E. F. (1993). The reality of repressed memories. *American Psychologist*, *48*(5), 518–537. https://doi.org/10.1037/0003-066X.48.5.518

Mair, P., & Wilcox, R. (2019). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01246-w

Mecklinger, A., Parra, M., & Waldhauser, G. T. (2009). ERP correlates of intentional forgetting. *Brain Research*, *1255*, 132–147. https://doi.org/10.1016/j.brainres.2008.11.073

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Noreen, S., & MacLeod, M. D. (2013). It's all in the detail: Intentional forgetting of autobiographical memories using the autobiographical

think/no-think task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 375–393. https://doi.org/10.1037/a0028888

Otgaar, H., Howe, M. L., Patihis, L., Merckelbach, H., Lynn, S. J., Lilienfeld, S. O., & Loftus, E. F. (2019). The return of the repressed: The persistent and problematic claims of long-forgotten trauma. *Perspectives on Psychological Science*, *14*(6), 1072–1095. https://doi.org/10.1177/1745691619862306

R Core Team. (2019a). *Foreign: Read data stored by 'minitab', 's', 'sas', 'spss', 'stata', 'systat', 'weka', 'dBase', …* https://CRAN.R-project.org/package=foreign

R Core Team. (2019b). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rinker, T. W., & Kurkiewicz, D. (2018). *pacman: Package management for R*. http://github.com/trinker/pacman

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schie, K. v., & Anderson, M. C. (2017). Successfully controlling intrusive memories is harder when control must be sustained. *Memory*, *25*(9), 1201–1216. https://doi.org/10.1080/02699931.2013.765387

Schneider, W. E., Eschman, A., & Zuccolotto, A. (2002). *E-prime user's guide*. Pittsburgh, PA, USA: Psychology software tools Inc.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *Afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Stramaccia, D., Rischer, K. M., Fawcett, J. M., & Benoit, R. G. (2019). Memory suppression and its deficiency in psychological disorders: A focused meta-analysis. Preprint at *PsychArxiv*. https://doi.org/10.31234/osf.io/5wynm

Wegner, D. M., Schneider, D. J., Carter, S. R., & White, T. L. (1987). Paradoxical effects of thought suppression. *Journal of Personality and Social Psychology*, *53*(1), 5–13. https://doi.org/10.1037/0022-3514.53.1.5

Wessel, I., Huntjens, R. J. C., & Verwoerd, J. R. (2010). Cognitive control and suppression of memories of an emotional film. *Journal of Behavior Therapy and Experimental Psychiatry*, *41*(2), 83–89. https://doi.org/10.1016/j.jbtep.2009.10.005

Wessel, I., Wetzels, S., Jelicic, M., & Merckelbach, H. (2005). Dissociation and memory suppression: A comparison of high and low dissociative individuals' performance on the Think–No think task. *Personality and Individual Differences*, *39*(8), 1461–1470. https://doi.org/10.1016/j.paid.2005.05.009.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), http://www.jstatsoft.org/v21/i12/paper. https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org.

Wickham, H., & Bryan, J. (2019). *Usethis: Automate package and project setup*. https://CRAN.R-project.org/package=usethis

Wickham, H., François, R., Henry, L., & Müller, K. (2019a). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Chang, W. (2019b). *Devtools: Tools to make developing r packages easier*. https://CRAN.R-project.org/package=devtools

Wickham, H., & Miller, E. (2019). *Haven: Import and export 'spss', 'stata' and 'sas' files*. https://CRAN.R-project.org/package=haven

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.name/knitr/