International Journal of Psychophysiology 163 (2021) 35–46

Contents lists available at ScienceDirect

International Journal of Psychophysiology

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

Exploring power in response inhibition tasks using the bootstrap: The impact of number of participants, number of trials, effect magnitude, and study design☆

Curtis D. Von Gunten [1,*], Bruce D. Bartholow

Department of Psychological Sciences, University of Missouri, United States of America

ARTICLE INFO

Keywords:
Inhibitory control
Executive functioning
Power
Reliability
Reproducibility
Bootstrap

ABSTRACT

A primary psychometric concern with laboratory-based inhibition tasks has been their reliability. However, a reliable measure may not be necessary or sufficient for reliably detecting effects (statistical power). The current study used a bootstrap sampling approach to systematically examine how the number of participants, the number of trials, the magnitude of an effect, and study design (between- vs. within-subject) jointly contribute to power in five commonly used inhibition tasks. The results demonstrate the shortcomings of relying solely on measurement reliability when determining the number of trials to use in an inhibition task: high internal reliability can be accompanied with low power and low reliability can be accompanied with high power. For instance, adding additional trials once sufficient reliability has been reached can result in large gains in power. The partial dissociation between reliability and power was particularly apparent in between-subject designs where the number of participants contributed greatly to power but little to reliability, and where the number of trials contributed greatly to reliability but only modestly (depending on the task) to power. For between-subject designs, the probability of detecting small-to-medium-sized effects with 150 participants (total) was generally <55%. However, effect size was positively associated with number of trials. Thus, researchers have some control over effect size and this needs to be considered when conducting power analyses using analytic methods that take such effect sizes as an argument. Results are discussed in the context of recent claims regarding the role of inhibition tasks in experimental and individual difference designs.

1. Introduction

Inhibition-related cognitive tasks are prevalent across all areas of psychology. Deficits in inhibition-related functions feature in many psychiatric disorders, such as addiction, antisocial personality disorder and attention-deficit/hyperactivity disorder (Moeller et al., 2001). Inhibition and other executive control processes are thought to influence academic achievement during early development (Borella et al., 2010). Social psychologists treat inhibition tasks as operationalizations of self-control in mental fatigue research (Hagger et al., 2010). Personality researchers assess the association between performance on inhibition tasks and self-control and impulsivity (Duckworth and Kern, 2011), impactful traits thought to affect a wide range of life outcomes (De Ridder and Lensvelt-Mulders, 2018; Moffitt et al., 2011). Within cognitive psychology and neuroscience, laboratory-based behavioral inhibition tasks have been used to examine the structure of executive functions (Friedman et al., 2008), the mechanisms of response conflict resolution (Braver et al., 2001), and control-related frontal lobe functioning (Ridderinkhof et al., 2004), among others.

Given its importance as a construct, a thorough understanding of the psychometric properties of inhibition measures, such as reliability, is critical to advancing each of these areas of scholarship. Although the reliability of inhibition tasks has been a longstanding concern among researchers (e.g., Miyake et al., 2000; Wöstmann et al., 2013), there has been a recent influx of scholarship on the topic (Enkavi et al., 2019; Hedge et al., 2018; Rouder et al., 2019). A primary focus of this research

☆ This research was supported by grants T32 AA013526 and P60 AA011998 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). Preparation of this article was supported in part by NIAAA grant R01 AA025451.

* Corresponding author at: Duke South Clinic, Room 3524, Blue Zone, Durham, NC 27710, United States of America.
E-mail address: CurtVonGunten@gmail.com (C.D. Von Gunten).
[1] Now at Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, United States of America.

https://doi.org/10.1016/j.ijpsycho.2019.08.008
Received 12 December 2018; Received in revised form 18 June 2019; Accepted 23 August 2019
Available online 10 December 2019

has been on how reliability differentially influences experimental research designs testing group-differences vs. individual difference research designs examining correlations. As the argument goes, inhibition tasks have roots in experimental research designs (Hedge et al., 2018). For significance testing in this research paradigm, low individual difference variance (between-subject variance) is desirable because it obfuscates the experimental effect that exists between groups; for instance, it features in the variance term in the denominator of the *t*-test and in the within-treatment variance component of a one-way ANOVA. The presumably reliable and robust results of inhibition tasks in this research paradigm are thought to be due to their low degree of between-subject variance. When holding other variance components constant, lower between-subject variance results in a less reliable measure that can less easily preserve rank-order stability across participants. Because of this, it is furthermore suggested, a bit counterintuitively, that tasks used successfully in experimental designs are expected (or required) to have low reliability (Enkavi et al., 2019; Hedge et al., 2018).

It is important to recognize, however, that the above discussion equivocates on the word "reliability". When used in the technical context of discussing types of variance, it is referring to a property of a measure. When used in the context of discussing the robustness of results, it is referring to statistical power (the ability to consistently reject a null hypothesis when it is false), which is a property of the broader research context (including measurement reliability). The relationship between the reliability of a measure and the power of a research context remains an underexplored issue (Parsons et al., 2019). But it is clear that adequate measurement reliability is not sufficient for adequate power. For instance, if a sample size is too small, a statistical significance test scenario that involves values that are obtained from a reliable measure can still fail to attain adequate power. This scenario is characterized by high quality data, but not enough of it. Furthermore, recent work has found that for event-related potential measures involving inhibition tasks, increasing the number of trials can greatly improve power, even after suitable reliability has already been achieved (Boudewyn et al., 2018; Fischer et al., 2017). Therefore, studies making practical recommendations about the minimal number of trials needed in a task based solely on reaching suitable reliability thresholds may be misleading (Cohen, 1962; Fischer et al., 2017; Huffmeijer et al., 2014; Larson et al., 2010; Marco-Pallares et al., 2011; Olvet and Hajcak, 2009; Pontifex et al., 2010; Rietdijk et al., 2014; Segalowitz and Barnes, 1993; Thigpen et al., 2017). Moreover, within-subject designs typically improve power by removing systematic individual differences, making that source of variability, and the reliability estimates that depend on it, less relevant.

Part of the reason the relationship between measurement reliability and statistical power remains underexplored may be because there is little understanding of how combinations of factors like number of subjects, number of trials, and between vs. within-subject designs contribute to power. Lack of attention to these relationships has contributed to the administration of widely varying numbers of trials across studies and across tasks. For instance, Hedge et al. (2018) use a large number of trials consistently across their tasks (720 trials in two tasks; and 600 trials in two tasks), whereas Enkavi et al. (2019) use 9 to 401 trials across 36 tasks (SD = 103 trials).[2] Because there is a tradeoff between measurement quality and time when choosing the number of trials to use, it is important to know whether there is an optimal number of trials researchers should be administering for specific inhibition tasks to ensure adequate statistical power. It could be the case that researchers currently using relatively few trials should start using more or those using more should start using less. This issue is even more relevant in light of the recent claim mentioned above that designs testing group differences should have low measurement reliability (and therefore a low number of trials) in order to obtain consistent and robust results (Enkavi et al., 2019; Hedge et al., 2018). Furthermore, beyond a basic understanding that between-subject designs are less powerful than their within-subject counterparts, the degree of this difference is uncertain. Finally, it is unclear whether including additional trials in between-subject designs can compensate for their power deficiency. The present study uses bootstrap sampling of real data in order to explore these issues (Boudewyn et al., 2018; Rousselet et al., 2019).

Of course, numerous analytic methods have been developed to test the power of a given study or statistical test, and such methods generally incorporate information on the size of the effect, size of the sample, and design type (between- vs. within-subject). However, such methods have a number of limitations. For example, analytic solutions for computing the power of *t*-tests, which is the focus of the present paper, depend on assumptions, such as normality, that data may not meet (Erceg-Hurn and Mirosevich, 2008; Wilcox and Keselman, 2003). Second, such solutions do not directly factor the number of trials into the power calculation. The number of trials will influence the degree of variation, which, in turn, will influence the standardized effect size. Yet, researchers typically do not consider the influence of trial number on anticipated effect size when crafting a study, perhaps because the degree of influence is uncertain and task specific.

A third limitation of analytic solutions for computing the power of *t*-tests is that between- and within-subject designs can rely on a different variance estimate (denominator) for effect size calculations (e.g., Cohen's *d*). This can result in effect sizes in within-subject designs being more dependent on the number of trials used in the task, and therefore can result in effect sizes not generalizing across designs. For this and other reasons, some researchers regard effect sizes in within-subject designs as an overestimation of the true effect size (Delaney and Maxwell, 2004; Dunlap et al., 1996; Olejnik and Algina, 2003). The debate regarding whether effect size estimates should generalize across different designs remains unresolved (e.g., Lakens, 2013; Rouder, 2016; Westfall, 2016). The bootstrap sampling approach used in the current study is able to sidestep this issue by artificially introducing effect magnitudes in the original units of measurement (e.g., group differences in terms of accuracy percentage). Furthermore, this approach enables the examination of how number of trials influences the Cohen's *d* value of both between- and within-subject designs.

Fourth, it often is not obvious how standardized effect sizes translate into effects in the original units of measurement (e.g., accuracy or response time). Because effect size estimates used in analytic power calculations are standardized by variance, the same standardized effect size can result from very different unstandardized effect sizes (hereafter "effect magnitude"). For instance, a 100-ms group difference in response time (RT) could result in the same Cohen's *d* value as a 300-ms difference, if the former has less variance within groups. Similarly, the same RT difference could produce a wide range of Cohen's *d* values depending on the variance present in the data (Boudewyn et al., 2018).

1.1. Current study

Much focus has been placed on the measurement reliability of inhibition tasks, potentially at the expense of statistical power. Focusing on independent and dependent samples *t*-tests, the current study adopts a bootstrap sampling approach to estimate the influence of number of participants, number of trials, effect magnitude, and study design (between- vs. within-subject), on the power to detect mean differences in five commonly used inhibition tasks. By crossing these four aspects of study method, the current approach provides power estimates for each combination of these method features. Furthermore, rather than treat

[2] The number of trials used is particularly important in the latter study, since it relies on Monte Carlo simulations to generate a distribution of test-retest reliability estimates for each task. These simulated distributions are then randomly sampled in order to compare their reliability to that of the literature (mean r = 0.25). This approach puts a lot of faith in the version of the task administered, particularly in the number of trials. If including more trials would have increased reliability, or decreased the range of the confidence intervals, the distributions repeatedly sampled from would have been quite different.

effect sizes as given, as required in analytic power calculations, the current study starts from the method features, introduces an effect magnitude of interest, and estimates both power and effect size to determine the feature's influence on each. This bootstrap simulation approach is also able to capture real variation and noise manifest in the tasks (Boudewyn et al., 2018; Kiesel et al., 2008; Kleinman and Huang, 2016). Reliability and power are also compared in order to determine the situations in which high measurement reliability results in low power and those in which adding more trials results in additional power even after a reliable measure has been obtained (Boudewyn et al., 2018).

2. Method

2.1. Participants and procedure

Participants were 463 undergraduates enrolled in an introductory psychology course who completed the study for partial course credit (276 male [40%]; *M* age = 18.81 years, *SD* = 1.54, range: 17–33 years). Given the nature of the color-naming Stroop task, individuals with red-green colorblindness were not eligible for the study. The data were collected as part of a larger study. Based on the aims of that study, data collection was planned to begin in the fall semester and to continue unconditionally until the end of February in the spring semester. The minimum target *N* was 300, with a desired *N* of 400 or greater.

Participants attended a single laboratory session in which a battery of inhibition tasks, described next, was administered. No experimental manipulations were introduced during data collection. As part of a larger study examining associations between inhibition task performance and self-reported self-regulation outcomes (see Von Gunten et al., 2019), participants also completed a questionnaire battery after completing the inhibition tasks, which will not be discussed in this paper. The tasks were completed in the same order for all participants (antisaccade, go/no-go, Stroop, Simon, stop-signal).

2.2. Inhibition tasks

2.2.1. Antisaccade task

In this version of the antisaccade task (adapted from Miyake et al., 2000) each trial consisted of a black fixation cross that appeared on the computer screen for a random duration between 1000 ms and 2750 ms in increments of 250 ms on a white background. During an initial pro-saccade block, the fixation point was followed by a cue (black square) appearing on one side of the screen for 200 ms, which was then replaced by a target stimulus (an arrow pointing up, down, left, or right, enclosed in an open 5/8-in. square) shown for 115 ms. The target was then masked with a four-pointed star, which remained on the screen until the participant indicated the target's direction with an arrow key press. The structure of trials in the subsequent antisaccade block was similar, except that the target stimulus appeared on the side opposite the cue. The task began with the 40-trial prosaccade block, followed by an 8-trial antisaccade practice block and then two 40-trial antisaccade blocks. The dependent measure in this task was the proportion of errors made in the antisaccade bock.

2.2.2. Go/no-go task

In this version of the go/no-go task (adapted from Newman and Kosson, 1986, and Nieuwenhuis et al., 2003), each trial consisted of a white numeral ranging from 1 through 8 that appeared randomly for 200 ms on a black background. Participants were instructed to press the space bar as quickly as possible whenever the number was not a 3 or 8 (*go trials*), and to refrain from pressing the space bar if the number was a 3 or 8 (*no-go trials*). Each of the eight numerals appeared equally often resulting in 80% *go* trials and 20% *no-go* trials. The inter-trial interval varied randomly between 500 ms, 750 ms, 1000 ms, and 1250 ms. The task began with a practice block of 15 trials followed by four blocks of 80 trials each. The dependent measure was the proportion of errors made on the no-go trials.

2.2.3. Stroop task

In this version of the Stroop task (adapted from Stroop, 1935) each trial consisted of a letter string or word that appeared on the computer screen in one of four colors (red, blue, green, yellow) on a black background. On each trial, participants were instructed to identify the color of the stimulus as quickly as possible by pressing one of four keys on a standard QWERTY keyboard ("v", "b", "n", and "m"). Trials were separated by a 75-ms inter-trial interval following the response. The task began with 24 *neutral* trials, in which participants were to identify the color of a letter string ("XXXXX"). Next, participants completed two blocks of 48 *congruent trials*, in which color words were presented in corresponding colors. Finally, participants completed a brief practice block of 16 trials followed by four blocks of 48 *incongruent trials*, in which color words were presented in non-corresponding colors (e.g., "RED" printed in green). The dependent measure for this task was the Stroop interference effect, calculated as the difference in mean reaction time (RT) between the incongruent and congruent trials (for discussion of alternative control methods, see Laird et al., 2005).

2.2.4. Simon task

In this version of the Simon task (adapted from Lu and Proctor, 1995; Simon and Rudell, 1967), each trial consisted of a white fixation cross that appeared on the computer screen for 500 ms on a black background. The fixation point was followed by the word "Left" or the word "Right" that appeared at random on the left or right side of the screen for 200 ms. Participants were instructed to identify the word as quickly as possible (within 750 ms) by pressing a left-hand key (Caps Lock) if the word was "Left" and a right-hand key (Enter) if the word was "Right." *Non-conflict* trials are those in which the word corresponds to its location (and, hence, the correct response is mapped to the word's location), whereas *conflict trials* are those in which the word and its location correspond to opposing responses (e.g., "Right" on the left side of the screen). Trials were separated by a 300-ms inter-trial interval. The task began with a practice block of eight trials, followed by four blocks of 80 trials each. The dependent measure for this task was the RT difference between the conflict and non-conflict blocks.

2.2.5. Stop-signal task

This version of the stop-signal task was taken from an open source program called Stop-It (Verbruggen et al., 2008). Each trial consisted of a white square or circle that appeared at random in the center of the computer screen on a black background. Participants were instructed to press the "z" key if the object was a square and the "/" key if the object was a circle (stickers of the shapes were placed on these keys). On 25% of the trials, the shape was followed by a beep via headphones. When this occurred, participants were instructed to withhold pressing any buttons until the next shape appeared. Shapes remained on the screen for up to 1250 ms. The amount of time between the beep and the presentation of the shape (stop-signal delay; SSD) began at 250 ms. If a participant got a beep trial incorrect the SSD was decreased by 50 ms, making the next beep trial easier. If a participant got a beep trial correct the SSD was increased by 50 ms, making the next beep trial harder. The adaptive nature of the task is intended to keep accuracy rates close to 50%. The task began with 16 practice trials, and was then followed by three blocks of 64 trials (16 signal trials per block). The dependent measure for the top-signal task was the stop-signal reaction time (SSRT). SSRT can be understood as the length of time required for a participant to react to the stop stimulus (Logan and Cowan, 1984). Although SSRT scores based on the block-based integration method are more accurate inhibition measures when gradual slowing and positive skew are present in the data (Verbruggen et al., 2013), in order to ease simulation processing time, the SSRT was calculated using the simpler mean-based method.

2.3. Data cleaning

Due to a data collection error, data from the top-signal was not recorded from 41 participants. Moreover, 100 participants had 0% accuracy on the top trials, and therefore their data had to be removed. This problem was recognized during data collection after approximately 425 participants had been run. Subsequent interviews with several participants indicated that the instructions for the top-signal task were not clear enough. The instructions were then modified to circumvent this problem for the remaining participants, which corrected the problem. An additional 17 participants were removed due to 0% accuracy during the first trial block. Because there was accuracy feedback after each block, it is likely that these participants did not understand the instructions until they saw their feedback and then corrected following the first block.

All of the tasks were examined for compliance using participant mean accuracy (Hedge et al., 2017). Participants with problematic accuracy levels were removed using different thresholds for each task based on each task's distribution. For the antisaccade task, participants whose accuracy was <80% on congruent trials ($n = 14$) were removed. For the go/no-go task, participants with accuracy < 80% on congruent trials were removed ($n = 23$). For the Simon task, participants with accuracy < 60% on congruent trials ($n = 5$) were removed. An additional participant with 0% accuracy on incongruent trials was also removed. For the Stroop task, participants whose accuracy was <80% on congruent trials, or <50% on incongruent trials, were removed ($n = 35$). For the stop-signal, participants whose accuracy on no-go trials was <80%, or whose signal accuracy was <20% ($n = 37$), were removed. Two additional participants were removed who did not complete all of the blocks due to computer error.

2.4. Power estimation procedure

Participants and trials were pseudo-randomly sampled with replacement from the observed task data (bootstrap sampling; Efron and Tibshirani, 1994; Peng et al., 2005). The fact that participants were not exposed to any experimental manipulations during data collection allowed us to artificially introduce carefully considered effect magnitudes (Boudewyn et al., 2018; Kiesel et al., 2008; Kleinman and Huang, 2016). Any single combination of values representing four features of an experiment (participant number, trial number, effect magnitude, between vs. within) constituted a simulated experiment (e.g., 35 participants, 100 trials, 100 ms, within).

The simulated experiments imitated designs in which participants are exposed to either between- or within-participant manipulations (i.e., designs testing the effect of some manipulation on task performance). In what follows, p corresponds to the number of participants, t to the number of trials, and e to the effect magnitude. To imitate the structure of experiments using between-subject manipulations, two groups of $p/2$ participants were sampled with replacement (ensuring independence of the subsamples [Rice, 1995]) from the available pool of real participants. Next, t trials were sampled with replacement from each participant in the two groups. The mean of each participant was then calculated. For the two tasks that depend on a difference between congruent and incongruent trials (Stroop task and Simon task), two groups (congruent and incongruent) of $t/2$ trials were sampled with replacement for each participant from the appropriate trial type. The mean for each of the two trial groups was calculated for each participant and then a difference score was computed. For the stop-signal, one group of t not-stop trials and one group of $t \times 3$ stop trials were chosen to correspond with the proportions of stop and not-stop trials in the task as it was implemented here. The mean RT of correct not-stop trials and the mean SSD of stop trials were calculated for each participant and the mean SSD was subtracted from the mean RT, resulting in an SSRT for each participant.

To introduce the effects, the participants in one of the two participant groups (the groups do not differ except by random sampling) had value e added to their dependent variable score (e.g., Boudewyn et al., 2018). The nature of e depended on the type of score in each task. For tasks that depend on accuracy (antisaccade and go/no-go), a percentage in the form of a decimal (e.g., 0.10) was added to each participant's accuracy mean. For the tasks that depend on the RT difference between congruent and incongruent trials (Stroop and Simon), a ms value in the form of an integer (e.g., 50) was added to each participant's difference score. For the stop-signal, a ms value was added to each participant's SSRT. Cohen's d was then calculated based on the pooled variance of the two participant samples. Finally, a two-tailed, independent t-test was performed ($\alpha = 0.05$). This process was performed 500 times for each set of experimental conditions. Power was calculated as the proportion of tests that were statistically significant. Cohen's d for each experiment was calculated as the mean of the 500 Cohen's d values.

To imitate within-subject manipulations, one group of p participants was sampled with replacement from the pool of participants. Next, $t/2$ trials were sampled with replacement from each participant in the two groups. The mean of each trial group was then calculated for each participant. For the Stroop task and Simon task, $t/4$ trials were sampled with replacement from congruent trials twice and from incongruent trials twice, creating four groups of trials. The mean of the four groups was computed and two difference scores were computed for each participant. The process for the stop-signal was similar; $t/2$ trials were sampled twice from not-stop trials and $t/2 \times 3$ trials were sampled twice from stop trials. This resulted in four trial groups for each participant. The mean RT of the two not-stop trial groups and the mean SSD of the two stop trial groups were computed and two difference scores (SSRT scores) were computed.

To introduce the artificial effects, one of the two dependent variable scores for each participant had value e added to the score in a manner similar to the between-participant process. Cohen's d was then calculated based on the variability of the within participant difference between condition means (d_z; see Lakens, 2013; Rouder, 2016; Westfall, 2016 for discussion on this topic). Finally, a two-tailed, dependent t-test was performed ($\alpha = 0.05$). This process was performed 500 times for each set of experimental conditions. Power and Cohen's d were calculated in the same manner as the between-participant process.

Choosing the values to use for the method variables across tasks involved a balance between attaining a good spread of power (minimal floor and ceiling effects), keeping Cohen's d values in between-participant designs similar across tasks, and keeping the values themselves similar across tasks. Some tasks appear to have a lower number of trials. This is because those tasks include trial types (e.g., go trials) that are critical to task structure but that are not used in the calculation of the performance score. In order to facilitate design comparison, the same values were chosen for both between- and within-participant designs within each task. The number of participants and the number of trials refer to the total number of participants, and to the total number of trials for each participant, for the experiment. Therefore, for between-participant designs, p number of participants means $p/2$ participants per condition (i.e., group), and for within-participant designs, t number of trials means $t/2$ trials per condition. This facilitates practical comparison, in terms of resources, across the two design types.

2.5. Reliability analyses

Internal reliability was analyzed as a function of number of subjects and number of trials. Subjects were randomly sampled with replacement. Trials were selected in the order completed by participants. For instance, if ten trials were used in an analysis it corresponds to the first ten trials completed and twenty trials corresponds to the first twenty trials. Reliability therefore represents the reliability up to that point in the task. For tasks that did not rely on difference scores (antisaccade, go/no-go) reliability was estimated with Cronbach's alpha. For tasks that did rely on difference scores (Stroop, Simon, stop-signal) reliability was

estimated by splitting trials into even and odds trials for each trial type (e.g., congruent trials and incongruent trials; 4 sets of trials). A mean was taken for each set of trials and difference scores computed, resulting in a difference score for both even and odd trials for each participant. The correlation between the even and odd difference scores was corrected by the Spearman-Brown formula.

For the tasks not dependent on a difference score, the variance of the N (number of participants) × I (number of trials) matrix was decomposed into three parts: between-subject variance, between-item variance, and error variance. Between-subject and between-item variance were calculated via the method used in two-way analysis of variance designs. Between-subject variance was determined by calculating the mean for each participant (rows) across all trials. Next, the grand mean was subtracted from each participant mean and the results were each squared and then summed together. The resulting value was multiplied by the number of trials (which was the same for each participant) and divided by the number of participants minus one (the degrees of freedom). The same process was applied to trials (columns) exchanging "participant" with "trial" and vice versa. Error variance was calculated by subtracting the between-subject sum of squares and between-item sum of squares from the total sum of squares and dividing by the number of rows minus one multiplied by the number of subjects minus one. Total sum of squares variance was calculated by subtracting the grand mean from each element in the data matrix and summing the results. This process was performed as a function of numbers of trials in order to examine in more detail how number of trials contributes to reliability. Furthermore, the three variance components were analyzed both in mean square units and in a relative fashion by normalizing by the total amount of variance. All analyses were conducted with Python 3 (https://github.com/Curt-Von-Gunten/Power-Reliability-BootstrapSimulations).

3. Results

3.1. Reliability

Fig. 1 shows reliability as a function of number of participants and number of trials for each of the five tasks.[3] Tasks varied both in the gains provided by increasing trials and in the overall level of reliability. The antisaccade and stop-signal reached reliability levels above 0.80, though the go/no-go and Stroop task were just below this mark. The Simon task had the lowest reliability. The number of participants had no influence on reliability with the exception of the Simon task; although, this influence was small.

Analysis of the normalized decomposed variance of the antisaccade and go/no-go tasks (left panel of Fig. 2) revealed a large increase in between-subjects variance and a smaller decrease in error variance as the number of trials increased. Examining the non-normalized mean variance (right panel of Fig. 2) revealed that these relative changes in variance were due primarily to changes in between-subject variance as the number of trials increased. More specifically, between-subject variance increased with increasing trials whereas the remaining contributions to variance remained mostly stable, with the exception of a slight decrease in between-item variability in the go/no-go task.

3.2. Power

Figs. 3 through 7 depict the power simulation results for each task. Table 1 provides associations between power and each of the three method features examined (number of participants, number of trials, effect magnitude).

3.2.1. Between-subject designs

Although the effect magnitudes were tailored to keep the standardized effects sizes in between-subject designs similar across the tasks, the Stroop and Simon tasks had slightly higher standardized effect sizes across experiments. For between-participant designs, small and small-to-medium effect sizes (~0.20–0.35; first panels of Figs. 3 through 7) were generally not detectable above a probability of 50%, even with 150 participants, consistent with power estimates from analytic methods. For effect sizes less than or equal to $d = 0.35$, power did not exceed 60% in any experiments for any task (max = 54%). For small-to-medium and medium-sized effects (~0.30–0.60; first and second panels) power was below 55% for all experiments with 50 participants or less for all tasks. When the number of participants was doubled to 100, the tasks generally had power < 80%, with some experiments just reaching the 80% mark at effect sizes around $d = 0.60$. With 150 participants, power results were highly variable, ranging from 24% to 95%, depending on the effect magnitude and number of trials in an experiment. The lowest effect size at which an experiment with small-to-medium effect sizes and with 150 participants attained 80% power or greater was $d = 0.48$. Panels three and four in Figs. 3 through 7 indicate that roughly 100 participants were needed to attain 80% power across tasks for medium-to-large effects (~0.60–0.80). For large effects and greater (>0.80), roughly 50 participants were needed to attain 80% power.

3.2.2. Within-subject designs

Regardless of effect magnitude, all experiments that used 150 participants and roughly 200 trials reached the power ceiling. The antisaccade, Stroop, and stop-signal all exhibited significantly larger standardized effect sizes in within- compared to between-subject designs (Figs. 3, 6, and 7). The large effect sizes for these three tasks were accompanied by large boosts in power. For the effect magnitudes depicted in the third and fourth panels for these three tasks, almost all experiments reached adequate power, and many reached ceiling. For the effect magnitudes depicted in the second panel of the figures, using the maximum number of trials (~200) results in power > 80% for all but two experiments for all five tasks, indicating that even with small effect magnitudes and a small number of participants, effects are reliably detectable. For the smallest effect magnitudes (left panel of the figures), the tasks show a large degree of variation in power, but all of the tasks attain power > 80% with 100 participants and the maximum number of trials tested (~200).

3.2.3. Comparing between-subject and within-subject designs

Because the number of participants, the number of trials, and the effect magnitudes are matched across designs for each task, the figures demonstrate how much more power mileage a researcher can get using the same resources (e.g., 50 participants each performing 100 trials when the real effect is 25 ms) in a within- vs. a between-subject design. Generally, the number of trials had a moderate influence on power in between-subject designs, with correlations ranging from 0.10 to 0.35 (Table 1). The number of trials had a large influence on power in within-subject designs, with correlations ranging from 0.46 to 0.58. The number of participants had a large influence in both between-subject (0.62. to 74) and within-subject (0.51 to 0.59) designs.

As expected, the number of trials had a much greater impact on Cohen's *d* values in within- compared to between-subject designs, with correlations ranging from 0.53 to 0.75 for within-subject designs and from 0.20 to 0.50 for between-subject designs (Table 1). Also of note, effect magnitude was highly correlated with effect size in between-participant designs, with correlations ranging from 0.85 to 0.97. This association was attenuated for within-participant designs, with correlations ranging from 0.63 to 0.82. As expected, the number of participants had no influence on effect size because although greater samples reduce the standard error of sampling means, greater samples do not reduce variance (and actually increase it by better estimating the population variance, which is underestimated by uncorrected standard deviation estimates).

3.2.4. Differences across tasks

For the two tasks that depended on accuracy (antisaccade and go/no-

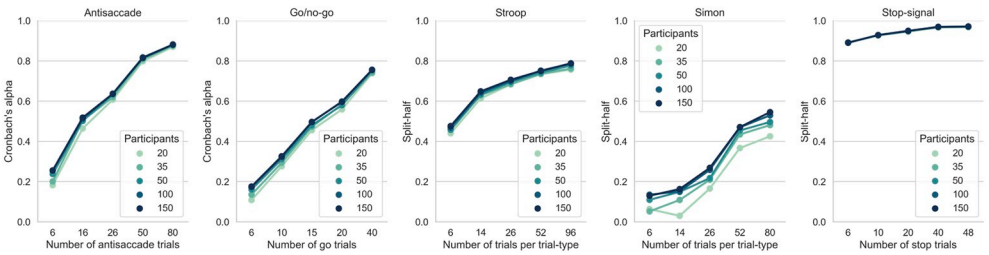

Fig. 1. Internal reliability as a function of number of trials (x-axis) and number of participants (line color) for each task (column). The same number of participants was analyzed in all tasks. Two different approaches to estimating internal reliability were used depending on the task (y-axis). For details regarding the number of trials, see footnote 3[1].

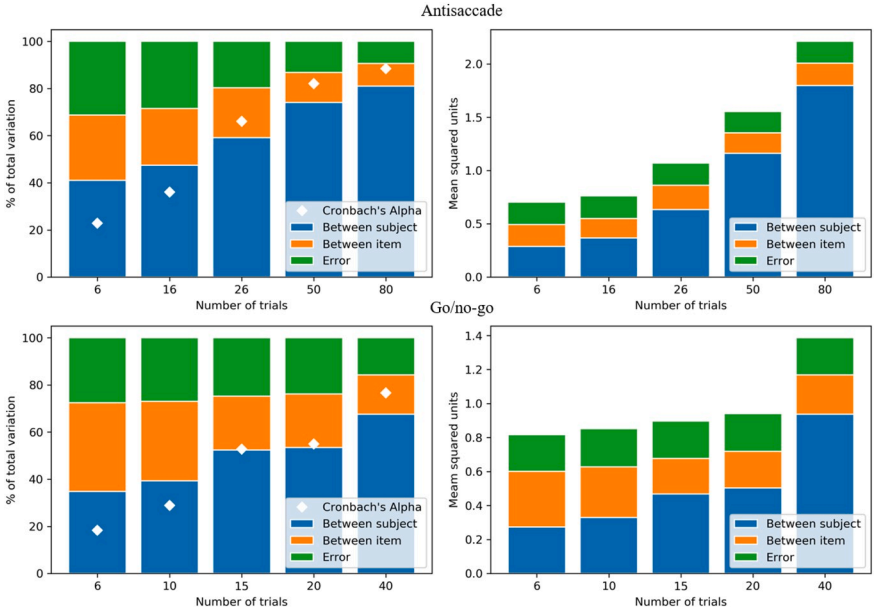

Fig. 2. Variance decomposition as a function of number of trials (x-axis) for two tasks (row). The left column presents variance normalized by the total amount of variance. The right column presents variance in mean squared units. Cronbach's alpha values are depicted in non-decimal form in the left column.

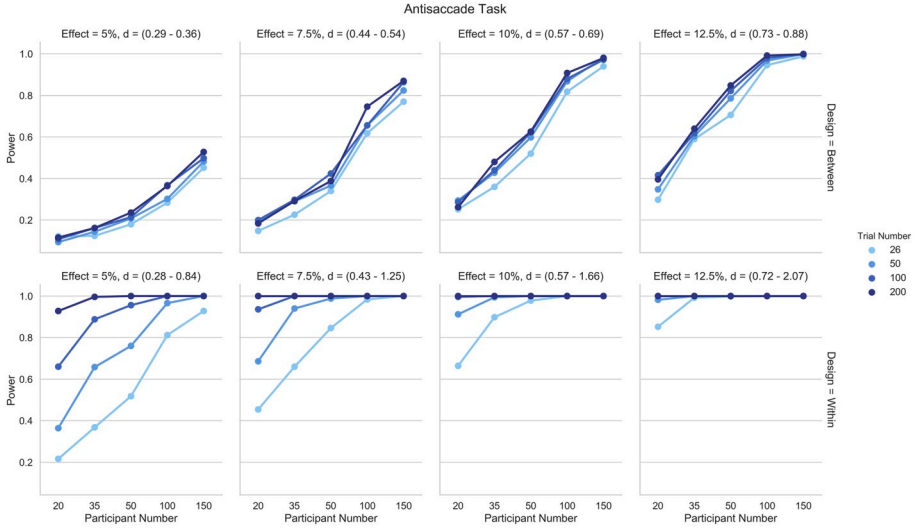

Fig. 3. Power for the antisaccade task as a function of the number of participants (x-axis), the number of trials (color), effect magnitude (column), and design type (row). Effect: the known effect magnitude in unstandardized units. d: the range of Cohen's *d* values for the 20 experiments in each graph.

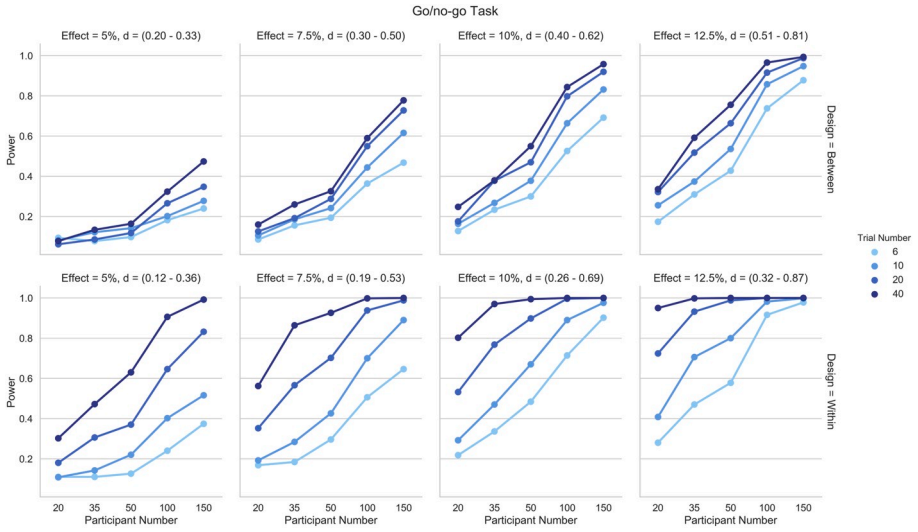

Fig. 4. Power for the go/no-go task.

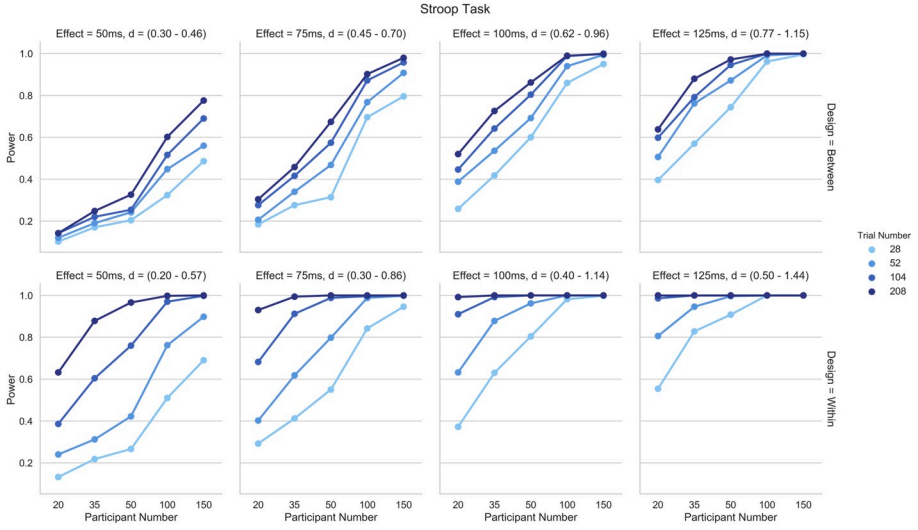

Fig. 5. Power for the Stroop task.

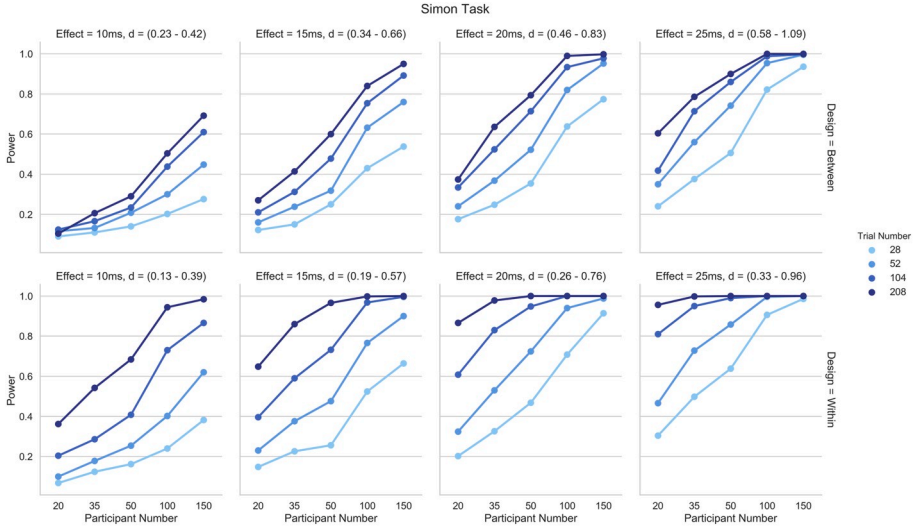

Fig. 6. Power for the Simon task.

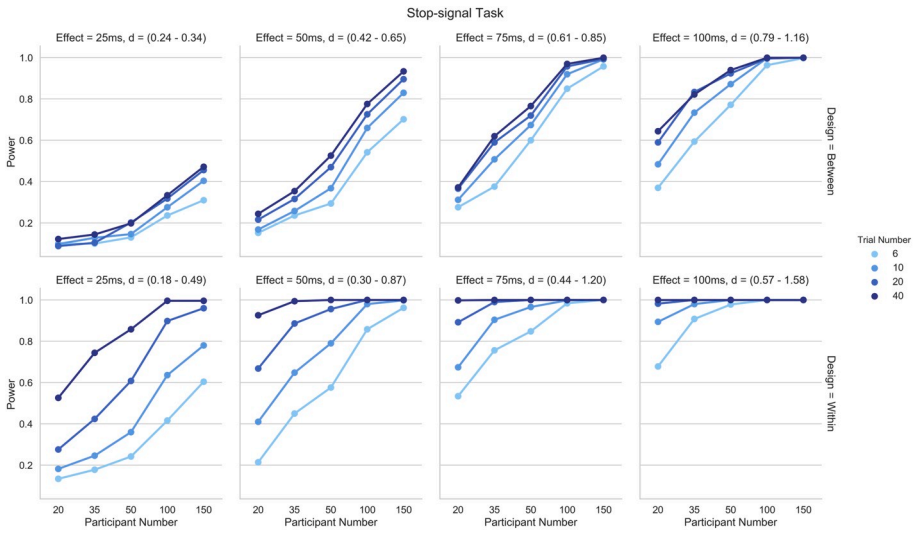

Fig. 7. Power for the stop-signal task.

Table 1
Spearman's rank correlations between experiment features, power, and effect size (Cohen's *d*) for between- and within-subject designs in each task.

Task and design	Correlation with power				Correlation with effect size		
	Number of participants	Number of trials	Effect magnitude	Effect size	Number of participants	Number of trials	Effect magnitude
Antisaccade							
Between-subjects	0.74	0.10	0.63	0.58	−0.08	0.20	0.97
Within-subjects	0.52	0.52	0.43	0.66	−0.05	0.75	0.63
Go/no-go							
Between-subjects	0.71	0.24	0.63	0.67	−0.01	0.38	0.91
Within-subjects	0.59	0.57	0.53	0.75	−0.05	0.73	0.65
Stroop							
Between-subjects	0.71	0.24	0.63	0.64	−0.06	0.36	0.92
Within-subjects	0.56	0.55	0.53	0.73	−0.06	0.74	0.64
Simon							
Between-subjects	0.69	0.35	0.6	0.66	−0.06	0.50	0.85
Within-subjects	0.58	0.58	0.52	0.75	−0.05	0.74	0.64
Stop signal							
Between-subjects	0.62	0.17	0.74	0.75	−0.06	0.28	0.95
Within-subjects	0.51	0.46	0.64	0.74	−0.08	0.63	0.75

Note. Numbers in each cell are Spearman's rank correlation coefficients. The values of the three method variables were experimentally crossed and therefore are correlated r = 0. Moreover, each method variable was restricted to a minimal number of values (five values for number of participants; four values for number of trials) and effect magnitude (total of 80 combinations [5 × 4 × 4]). Thus, the correlations should only be used as rough estimates of associations.

go), the same effect magnitudes were used for the simulations (5%, 7.5%, 10%, 12.5%; Figs. 3 and 4) because they gave a similar spread of power for between-subject designs. For those designs, power was slightly greater for the antisaccade than for the go/no-go task. However, keep in mind that although the two tasks were equated on the effect size in the original unit of measurement, the standardized effect size was slightly larger for the antisaccade tasks (Figs. 3 and 4). There was also a difference in the contribution of the number of trials to power across the two tasks. For the antisaccade, the correlation between the number of trials and power was only r = 0.17, whereas for the go/no-go task the correlation was r = 0.24 (Table 1). A bigger overall power difference was found for within-subject designs, with greater overall power for the antisaccade task than for the go/no-go. This was the case despite the number of participants and the number of trials being more highly correlated with power for the go/no-go task than for the antisaccade task (Table 1). This is likely because power reached ceiling on more experiments for the antisaccade task compared to the go/no-go task.

For the two tasks that depended on RT difference scores (Stroop and Simon), unlike the accuracy-based tasks, different effect magnitudes were chosen for the simulations in order to get adequate power spreads for between-subject designs. The Simon required an RT effect magnitude range of 10–25 ms, whereas the Stroop required a much larger RT range of 50–125 ms. This suggests that the variance of the mean RT difference scores (from which the standard error is based in between-subject designs) was smaller in the Simon task. For the between-subject experiments, power was slightly higher for the Stroop task. However, recognize that both the effect size in terms of measurement units and in terms of standardized units differed for the four panels across the two tasks, with the Stroop exhibiting slightly higher standardized effect sizes (Table 1). There was also a difference in the contribution of the number of trials to power across the two tasks. The correlation between the number of trials and power was r = 0.35 for the Simon task and r = 0.24 for the Stroop task. A bigger overall power difference was found for within-subject designs, with greater overall power for the Stroop task than for the Simon task. Like the accuracy-based tasks, this was the case even though the number of participants and the number of trials were

[3] For each no-go trial there were 4 go trials and for each not-stop trial there were 3 stop trials. For the stop-signal, triple the number of stop trials to get the number of not-stop trials used for the reliability calculation. For the Stroop and Simon tasks, the same number of congruent and non-congruent trials was used. It should be noted as well that the congruent and non-congruent trials were blocked for the Stroop task but were not blocked for the Simon task.

more highly correlated with power in the Simon task than in the Stroop task (Table 1). Again, this difference in correlations is likely due to the fact that power reached ceiling on more experiments for the Stroop task compared to the Simon task.

4. Discussion

Researchers interested in measuring response inhibition with behavioral tasks have long relied on intuition and precedent to determine the number of trials to administer and the number of participants to test in order to achieve adequate power. Few specific recommendations concerning the proper combination of these factors have been available in the published literature, which has likely contributed to the variability across studies in the numbers of trials administered in these tasks. Furthermore, the contribution of measurement reliability to experimental vs. correlational research designs using inhibition tasks has been the focus of recent scholarship (Enkavi et al., 2019; Hedge et al., 2018), with the current understanding being that experimental designs testing group differences require low measurement reliability to have high power. The current study examined the influence of number of participants, number of trials, effect magnitude, and study design (between- vs. within-subject) on reliability and power in five commonly used inhibition tasks.

4.1. Reliability and power

The current results demonstrate the shortcomings of relying solely on measurement reliability in order to determine the number of trials to use in an inhibition task. Adequate measurement reliability is neither necessary nor sufficient in order to achieve adequate power. Comparing Fig. 1 to Figs. 3 through 7, it is clear that high internal reliability can be accompanied by low power and that low reliability can be accompanied by high power. It is also the case that adding additional trials once sufficient reliability has been reached can result in large increases in power (see, for instance, the antisaccade task). Part of the reason for this disconnect between internal reliability and power is due to their differential reliance on number of trials and participants, particularly for between-subject designs. The number of trials, as expected, had a large influence on reliability, whereas the number of participants had no influence on reliability for majority of the tasks. Contrary to this, both the number of trials and participants contributed to power. These associations with power were qualified by interactions with design type. The number of trials had small to moderate influence in between-subject designs, whereas in within-participant designs the number of trials had a large influence, contributing to power to roughly the same extent as the number of participants. The interaction between the number of participants and design-type on power was less apparent, with number of participants having a large influence on power for both design types, albeit a larger influence for between-subject designs.

4.2. Between vs. within subjects designs

Although it is a rudimentary fact that within-subject designs are typically more highly powered than between-subject designs, it is still illuminating to point out how large the difference is in light of study resources (i.e., number of trials and participants). In within-participant designs, power was above 80% for three of the tasks in experiments with the smallest effect magnitudes and with 50 participants when using the largest number of trials (~200). For the remaining two tasks (Simon and stop-signal), 100 participants were required at the lowest effect magnitudes. This reveals just how important the number of trials administered can be in within-subject designs. Since the gains are roughly similar to the gains from adding more participants to the study, increasing the number of trials could be an effective method for conserving resources. However, it should be noted that the figure trends indicate that successively doubling the number of trials results in roughly equal increases in power at each step, suggesting that the association with trial number and power is not linear and that, therefore, continuing to increase the number of trials results in diminishing returns (similar to reliability).

Unfortunately, many studies are not suited for within-subject designs. The current study finds that if a design requires a between-subject manipulation and the effect of interest is small or small-to-medium in size, then an experiment may not be worth running with fewer than 150 participants. None of the five inhibition tasks were able to detect an effect of this size above a probability of 54% when 150 or fewer participants were included. One pressing question is whether increasing the number of trials could rectify this situation. This is difficult to answer inasmuch as the current paper finds a positive association between number of trials and effect size. It can be seen from examining the top left graph of the power figures, that two of the tasks achieved power in the 0.60 to 0.70 range for experiments with the lowest effect magnitude and with the greatest number of trials and participants examined. However, because these experiments contained more trials, the effects sizes were pushed outside of the effect size range under consideration (as can be seen from the Cohen's *d* ranges in those graphs). Therefore, it can be misleading to use effect size as the reference for power when designing a study with these tasks, since design decisions actually alter it. This is contrary to customary (frequentist) thinking insofar as effect size is presumed to exist in the population independently of design decisions. This needs to be recognized when conducting power analyses using analytic methods, such as when using G*power (Faul et al., 2007).

Thus, although it is true that for small or small-to-medium standardized effects, experiments with 150 or fewer participants (and likely with many more, judging by the figures) will be drastically underpowered, it is also the case that adding trials can result in a greater effect size and greater power. For between-subjects designs, the power gained from doubling trials was 10% or less; although, this varied by task, with some showing a gain of 5% or less. This means that the experiments using 200 trials would need to increase to 800 trials in order to get at most a 20% increase in power. This number of trials is, however, pushing the limits of feasibility. This line of thought is assuming the same rate of increase as a function of number of trials. Yet, just as internal reliability increases will eventually stabilize as more trials are included, it is possible that the same may occur for power gains. Because all of the tasks but the stop-signal still had room for their internal reliability to stabilize or reach ceiling (Fig. 1), no firm statement can be made regarding this suggestion. What is clear is that power gains are markedly higher when increasing the number of participants for between-subject designs, suggesting that researchers should possibly decrease the number of trials (and thereby decrease reliability) in order to have more time for collecting additional participants. All of these considerations imply grim prospects for resource-intensive studies, such as those using psychophysiological methods, when effect magnitudes are small.

4.3. Reliability and power in experimental and correlational designs

It has recently been suggested that, in order to consistently detect effects (power), tasks used in experimental designs should have low reliability and low between-subject variability (Enkavi et al., 2019; Hedge et al., 2018). The motivation for this claim is that such variability features in the denominator of typical tests of statistical significance (i.e., *t*-tests and ANOVA). Nevertheless, the current study finds that increasing the number of trials increases the reliability of inhibition tasks (by increasing the relative amount of between-subject variance [Fig. 2]) *and* the power of *t*-tests, even if the power gains are often only modest in between-subject designs.

It is critical to note some terminological differences in the use of "between-subject variance". The argument from Hedge and colleagues (2018) equates between-subject variance with the variance in the denominator of a (presumably) between-subject *t*-test. This source of variance is typically called within-treatment variance in the context of

ANOVA tests. The important point is that this variance is derived from participant *averages* across trials. These averages are used across separate task administrations in Enkavi at al. (2019) and Hedge et al. (2018) to assess test-retest reliability. In the current study, the finding that increasing trials increases between-subject variance relies on variance decomposition at the individual item level and analyzes internal (within-session) reliability. The right panel of Table 1 shows that increasing trials increases standardized effect size, even when the effect magnitude remains constant. Because standardized effect size is a function of only effect magnitude and within-treatment variance (what the papers under discussion call between-subject variance), the increased number of trials and resulting increased between-subject variance (at the item level) is resulting in less variance within-treatments. This resulted in increased power.

This still leaves a discrepancy with the papers under discussion is that this decrease in within-treatment variance also coincided with an increase in internal reliability. The papers under consideration rely on the assumption that decreasing this variance will result in decreased (test-retest) reliability. In fact, this assumption underlies the claim that inhibition tasks are unsuitable for correlational research designs involving the examination of individual differences (Enkavi et al., 2019; Hedge et al., 2018). It is argued that between-subjects variance (what we call within-treatment variance), and the reliability scores that partly depend on such variability, is not large enough to consistently preserve the rank ordering of participant's inhibition scores. Nevertheless, as described above, the current study finds decreases in within-treatment variance and *increases* in reliability as more trials are included, with all tasks but one reaching internal reliability of 0.80 (or just below). We used the task data from the current study in an additional study that compares the variance decomposition of the task data to three self-report measures of self-control (Von Gunten et al., 2019). The results highlight that even though the reliability of the self-reports and tasks are similar, the variance structure is quite different, with the report measures actually containing less item-level between-subject variance. However, the lower between-subject variance is compensated for by lower error variance, relative to the tasks.

It is important to stress that the current study examines internal (within-session) reliability whereas the papers under discussion (Enkavi et al., 2019; Hedge et al., 2018) examine test-retest (across-session) reliability. The extent to which number of trials contributes to test-retest reliability, and whether estimated internal reliability during a single task administration is a good proxy for test-retest reliability, remains an open question. The studies under consideration provide some insight into this question by reporting internal reliabilities in their supplemental materials. With the exception of the stop signal, which had very high internal reliability in both the current study and in Hedge and colleagues (2018), the tasks in the current study (Stroop, go/no-go) reached higher internal reliabilities. Hedge et al. (2018) note that sub-optimal test-retest reliability could be due to substantial changes in performances over time or contexts, or to problematic task construction and measurement. If the former, one might expect higher within-session reliabilities (i.e., internal reliability) than test-retest reliabilities. They do not find evidence for this, since both reliabilities were low. However, the internal reliability found in the present paper was good for four of the five tasks, suggesting either that genuine change in performance over time would result in lower test-retest reliability or that the tasks used in the present study would result in higher test-retest reliability. Future research could examine the relationship between internal reliability and test-retest reliability (Parsons et al., 2019) given the discrepancy between the tasks used in the present paper and those used in Enkavi et al. (2019) and Hedge et al. (2018).

4.4. Limitations and future directions

The approach we used to create artificial effects involved adding a constant value to participant-level means in one group or condition for all tasks (Boudewyn et al., 2018; Kiesel et al., 2008; Kleinman and Huang, 2016). One potential shortcoming of this approach is that it does not accurately represent the nature of effects—in particular, it does not account for the heterogeneity of effects across people (Kenny and Judd, 2019). Furthermore, the study does not account for another potential source of effect variability that lies within participants. Research has shown that inhibition strategies and the effects of experimental manipulations can change over the course of cognitive tasks (e.g., Volpert-Esmond et al., 2018; Von Gunten et al., 2018). Relatedly, the effect could interact with practice and fatigue effects. If these small-scale temporal fluctuations of the true effect are present, the true effect across the task may be better understood in terms of the central tendency of true effects across different periods of the task. In the current study, none of this potential effect variation is modeled, and therefore power is likely overestimated. Additionally, within-subject designs inherently involve sequential administrations of a task. The current design subsamples from the same collection of trials for each of the two treatment conditions. Thus the simulations are not capturing potential sequential effects like practice or fatigue that can systematically occur across task administrations in within-subject designs. This may have resulted in inflated power estimates for within-subject designs in the current paper. Nevertheless, table S1 in the supplementary materials shows minimal time-on-task changes.

The bootstrap sampling approach used to estimate power in the current paper subsamples trials across each task. Given that participant's responses may differ throughout the course of a task, it is possible that these time-on-task effects could have influenced the power estimates. For instance, if responses differ in the first half of a task compared to the second half of a task, subsampling from all trials to estimate power for experiments using only half the number of trials in a task could result in inaccurate power estimates. The supplemental materials include an alternative bootstrap sampling approach that aims to minimize the influence of potential time-on-task effects. The results reveal little influence of time-on-task on the power estimates.

It is also worthy of mentioning that the power estimates were derived from data that had been cleaned. No trial-level trimming was performed but participants were removed based on very poor performance. The power estimates from the current study will likely be overestimates for studies that collect the number of participants designated in the figures and that do not clean the data in this manner. Thus, researchers should aim higher than the number of participants reported in the figures.

Next, the results only apply to the specific task designs used in the current study. These tasks vary widely across labs and areas (Elson, 2017 [FlexibleMeasures.com]; Wessel, 2017), and we have already noted differences in reliability as a function of number of trials between the present study and other recent studies (Enkavi et al., 2019; Hedge et al., 2018). On top of this, there are often several ways to score each task, including model-based approaches such as drift-diffusion models (Enkavi et al., 2019; White et al., 2014). Given recent concerns with the use of mean-based scoring (Davis-Stober et al., 2018; Rousselet and Wilcox, 2019), future research could investigate which task versions and scoring procedures result in the greatest power while balancing other obvious needs like construct validity. Furthermore, the presentation order of the tasks was the same for all participants, thus it is possible that tasks that were administered later in the experiment, like the Simon task and stop-signal, were susceptible to task order effects. Also, it should be noted that two-tailed tests were used in the current study.

The present study only examined power in the context of a single analytic strategy—the *t*-test. Future research could examine more complex designs. Mixed models are becoming more commonplace for modeling inhibition task data (Volpert-Esmond et al., 2018; Von Gunten et al., 2018), and for tasks that involve repeated trials more generally (Barr et al., 2013). They can also capture additional sources of effect variation mentioned above that is ignored in mean-based *t*-tests and accompanying analytic power calculations (Page-Gould, 2017). Furthermore, multivariate techniques such as factor analysis are

commonplace within this literature (e.g., Friedman et al., 2006; Korucuoglu et al., 2017). The bootstrap sampling technique used in the current study is particularly useful for complex statistical procedures where no tractable mathematical methods exist for determining power (Kleinman and Huang, 2016).

5. Conclusion

Inhibition tasks are widespread across psychology. The current study can aid researchers in making various method decisions when trying to design an adequately powered study using these tasks. It is important to not confuse the reliability of an inhibition measure with the ability to reliably reject false null hypotheses (i.e., with statistical power). Although they are associated, method features like number of participants and number of trials can differentially impact the two. Furthermore, one should use caution when choosing standardized effect sizes in analytic power calculations, since those effects sizes are not independent of the number of trials administered and of design type. Finally, small-to-moderate effects are difficult to detect using between-subject designs for resource-intensive studies, which often cannot achieve large numbers of participants. This may be the case even taking into account the small to modest influence increasing the number of trials can have on effect size and power.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpsycho.2019.08.008.

References

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278.

Borella, E., Carretti, B., Pelegrina, S., 2010. The specific role of inhibition in reading comprehension in good and poor comprehenders. J. Learn. Disabil. 43 (6), 541–552.

Boudewyn, M.A., Luck, S.J., Farrens, J.L., Kappenman, E.S., 2018. How many trials does it take to get a significant ERP effect? It depends. Psychophysiology 55 (6), e13049.

Braver, T.S., Barch, D.M., Gray, J.R., Molfese, D.L., Snyder, A., 2001. Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. Cereb. Cortex 11 (9), 825–836.

Cohen, J., 1962. The statistical power of abnormal-social psychological research: a review. J. Abnorm. Soc. Psychol. 65 (3), 145–153.

Davis-Stober, C.P., Dana, J., Rouder, J.N., 2018. Estimation accuracy in the psychological sciences. PLoS One 13 (11), e0207239.

De Ridder, D.T., Lensvelt-Mulders, G., 2018. Taking stock of self-control: a meta-analysis of how trait self-control relates to a wide range of behaviors. In: Self-regulation and Self-control. Routledge, pp. 221–274.

Delaney, H.D., Maxwell, S.E., 2004. Designing Experiments and Analyzing Data (London, England).

Duckworth, A.L., Kern, M.L., 2011. A meta-analysis of the convergent validity of self-control measures. J. Res. Pers. 45 (3), 259–268.

Dunlap, W.P., Cortina, J.M., Vaslow, J.B., Burke, M.J., 1996. Meta-analysis of experiments with matched groups or repeated measures designs. Psychol. Methods 1 (2), 170–177.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.

Elson, M. (2017). FlexibleMeasures.com: Go/No-Go Task. doi:10.17605/OSF.IO/GSX52.

Enkavi, A.Z., Eisenberg, I.W., Bissett, P.G., Mazza, G.L., MacKinnon, D.P., Marsch, L.A., Poldrack, R.A., 2019. Large-scale analysis of test–retest reliabilities of self-regulation measures. Proc. Natl. Acad. Sci. 116 (12), 5472–5477.

Erceg-Hurn, D.M., Mirosevich, V.M., 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. Am. Psychol. 63 (7), 591–601.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191.

Fischer, A.G., Klein, T.A., Ullsperger, M., 2017. Comparing the error-related negativity across groups: the impact of error-and trial-number differences. Psychophysiology 54 (7), 998–1009.

Friedman, N.P., Miyake, A., Corley, R.P., Young, S.E., DeFries, J.C., Hewitt, J.K., 2006. Not all executive functions are related to intelligence. Psychol. Sci. 17 (2), 172–179.

Friedman, N.P., Miyake, A., Young, S.E., DeFries, J.C., Corley, R.P., Hewitt, J.K., 2008. Individual differences in executive functions are almost entirely genetic in origin. J. Exp. Psychol. Gen. 137 (2), 201–225.

Hagger, M.S., Wood, C., Stiff, C., Chatzisarantis, N.L., 2010. Ego depletion and the strength model of self-control: a meta-analysis. Psychol. Bull. 136 (4), 495–525.

Hedge, C., Powell, G., Sumner, P., 2017. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 1–21.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50 (3), 1166–1186.

Huffmeijer, R., Bakermans-Kranenburg, M.J., Alink, L.R., van IJzendoorn, M.H., 2014. Reliability of event-related potentials: the influence of number of trials and electrodes. Physiol. Behav. 130, 13–22.

Kenny, D.A., Judd, C.M., 2019. The unappreciated heterogeneity of effect sizes: implications for power, precision, planning of research, and replication. Psychol. Methods 24 (5), 578–589.

Kiesel, A., Miller, J., Jolicø eur, P., Brisson, B., 2008. Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods. Psychophysiology 45 (2), 250–274.

Kleinman, K., Huang, S.S., 2016. Calculating power by bootstrap, with an application to cluster-randomized trials. EGEMs (1), 4.

Korucuoglu, O., Sher, K.J., Wood, P.K., Saults, J.S., Altamirano, L., Miyake, A., Bartholow, B.D., 2017. Acute alcohol effects on set-shifting and its moderation by baseline individual differences: a latent variable analysis. Addiction 112 (3), 442–453.

Laird, A.R., McMillan, K.M., Lancaster, J.L., Kochunov, P., Turkeltaub, P.E., Pardo, J.V., Fox, P.T., 2005. A comparison of label-based review and ALE meta-analysis in the Stroop task. Hum. Brain Mapp. 25 (1), 6–21.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863.

Larson, M.J., Baldwin, S.A., Good, D.A., Fair, J.E., 2010. Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): the role of number of trials. Psychophysiology 47 (6), 1167–1171.

Logan, G.D., Cowan, W.B., 1984. On the ability to inhibit thought and action: a theory of an act of control. Psychol. Rev. 91 (3), 295–327.

Lu, C.-H., Proctor, R.W., 1995. The influence of irrelevant location information on performance: a review of the Simon and spatial Stroop effects. Psychon. Bull. Rev. 2 (2), 174–207.

Marco-Pallares, J., Cucurell, D., Münte, T.F., Strien, N., Rodriguez-Fornells, A., 2011. On the number of trials needed for a stable feedback-related negativity. Psychophysiology 48 (6), 852–860.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., Wager, T.D., 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cogn. Psychol. 41 (1), 49–100.

Moeller, F.G., Barratt, E.S., Dougherty, D.M., Schmitz, J.M., Swann, A.C., 2001. Psychiatric aspects of impulsivity. Am. J. Psychiatr. 158 (11), 1783–1793.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H., Sears, M.R., 2011. A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences 108 (7), 2693–2698.

Newman, J.P., Kosson, D.S., 1986. Passive avoidance learning in psychopathic and nonpsychopathic offenders. J. Abnorm. Psychol. 95 (3), 252–256.

Nieuwenhuis, S., Yeung, N., Van Den Wildenberg, W., Ridderinkhof, K.R., 2003. Electrophysiological correlates of anterior cingulate function in a go/no-go task: effects of response conflict and trial type frequency. Cogn. Affect. Behav. Neurosci. 3 (1), 17–26.

Olejnik, S., Algina, J., 2003. Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol. Methods 8 (4), 434–447.

Olvet, D.M., Hajcak, G., 2009. The stability of error-related brain activity with increasing trials. Psychophysiology 46 (5), 957–961.

Page-Gould, E., 2017. Multilevel modeling. In: Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (Eds.), The Handbook of Psychophysiology. Cambridge University Press, Cambridge, UK, pp. 662–678.

Parsons, S., Kruijt, A.W., Fox, E., 2019. Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. Adv. Methods Pract. Psychol. Sci. 2 (4), 378–395.

Peng, X., Peng, G., Gonzales, C., 2005. Power Analysis and Sample Size Estimation Using Bootstrap. Phoenix: Paper presented at PharmaSUG 2005.

Pontifex, M.B., Scudder, M.R., Brown, M.L., O'Leary, K.C., Wu, C.-T., Themanson, J.R., Hillman, C.H., 2010. On the number of trials necessary for stabilization of error-related brain activity across the life span. Psychophysiology 47 (4), 767–773.

Rice, J.A., 1995. Mathematical Statistics and Data Analysis. Duxbury, Belmont, CA.

Ridderinkhof, K.R., Van Den Wildenberg, W.P., Segalowitz, S.J., Carter, C.S., 2004. Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. Brain Cogn. 56 (2), 129–140.

Rietdijk, W.J., Franken, I.H., Thurik, A.R., 2014. Internal consistency of event-related potentials associated with cognitive control: N2/P3 and ERN/Pe. PLoS One 9 (7), e102672.

Rouder, J.N., 2016, March 28. The effect-size puzzler, the answer. Retrieved from. http://jeffrouder.blogspot.com/2016/03/the-effect-size-puzzler-answer.html.

Rouder, J., Kumar, A., Haaf, J.M., 2019. Why Most Studies of Individual Differences With Inhibition Tasks Are Bound to Fail.

Rousselet, G.A., Wilcox, R.R., 2019, January 17. Reaction Times and Other Skewed Distributions: Problems With the Mean and the Median. https://doi.org/10.31234/osf.io/3y54r.

Rousselet, G.A., Pernet, C.R., Wilcox, R.R., 2019, May 27. A Practical Introduction to the Bootstrap: A Versatile Method to Make Inferences by Using Data-driven Simulations. https://doi.org/10.31234/osf.io/h8ft7.

Segalowitz, S.J., Barnes, K.L., 1993. The reliability of ERP components in the auditory oddball paradigm. Psychophysiology 30 (5), 451–459.

Simon, J.R., Rudell, A.P., 1967. Auditory SR compatibility: the effect of an irrelevant cue on information processing. J. Appl. Psychol. 51 (3), 300.

Stroop, J.R., 1935. Studies of interference in serial verbal reactions. J. Exp. Psychol. 18 (6), 643.
Thigpen, N.N., Kappenman, E.S., Keil, A., 2017. Assessing the internal consistency of the event-related potential: an example analysis. Psychophysiology 54 (1), 123–138.
Verbruggen, F., Logan, G.D., Stevens, M.A., 2008. STOP-IT: Windows executable software for the stop-signal paradigm. Behav. Res. Methods 40 (2), 479–483.
Verbruggen, F., Chambers, C.D., Logan, G.D., 2013. Fictitious inhibitory differences: how skewness and slowing distort the estimation of stopping latencies. Psychol. Sci. 24 (3), 352–362.
Volpert-Esmond, H.I., Merkle, E.C., Levsen, M.P., Ito, T.A., Bartholow, B.D., 2018. Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. Psychophysiology 55 (5), e13044.
Von Gunten, C.D., Volpert-Esmond, H.I., Bartholow, B.D., 2018. Temporal dynamics of reactive cognitive control as revealed by event-related brain potentials. Psychophysiology 55 (3), e13007.
Von Gunten, C.D., Bartholow, B.D., Martins, J., 2019, July 10. Inhibition Is Not Associated With Self-regulation Outcomes in Healthy College Students. https://doi.org/10.31234/osf.io/uwbdg.
Wessel, J.R., 2017. Prepotent motor activity and inhibitory control demands in different variants of the Go/No-go paradigm. Psychophysiology 55 (3), e12871. https://doi.org/10.1111/psyp.12871.
Westfall, J., 2016, March 25. Five different “Cohen’s d” statistics for within-subject designs. Retrieved from. http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/.
White, C.N., Congdon, E., Mumford, J.A., Karlsgodt, K.H., Sabb, F.W., Freimer, N.B., Poldrack, R.A., 2014. Decomposing decision components in the stop-signal task: a model-based approach to individual differences in inhibitory control. Journal of Cognitive Neuroscience 26 (8), 1601–1614.
Wilcox, R.R., Keselman, H.J., 2003. Modern robust data analysis methods: measures of central tendency. Psychol. Methods 8 (3), 254.
Wöstmann, N.M., Aichert, D.S., Costa, A., Rubia, K., Möller, H.-J., Ettinger, U., 2013. Reliability and plasticity of response inhibition and interference control. Brain Cogn. 81 (1), 82–94.