A Flexible Gaming Environment for Reliably Measuring Cognitive Control

Lindsay Wells
Games and Creative Technologies Research Group
The University of Tasmania
AUSTRALIA

lindsay.wells@utas.edu.au

Talira Kucina
School of Psychological Sciences
The University of Tasmania
AUSTRALIA

talira.kucina@utas.edu.au

Amelia Kohl
School of Psychological Sciences
The University of Tasmania
AUSTRALIA

amelia.kohl@utas.edu.au

Ian Lewis
Games and Creative Technologies Research Group,
The University of Tasmania
AUSTRALIA

ian.lewis@utas.edu.au

Kristy de Salas
Games and Creative Technologies Research Group
The University of Tasmania
AUSTRALIA

kristy.desalas@utas.edu.au

Eugene Aidman
Australian Department of Defence
AUSTRALIA

Eugene.Aidman@dst.defence.gov.au

Andrew Heathcote
School of Psychology, The University of Newcastle
Department of Psychology, The University of Amsterdam
AUSTRALIA & THE NETHERLANDS

andrew.heathcote@newcastle.edu.au

https://www.ampl-psych.com/

ABSTRACT

Over the last few years, it has become accepted that reliable measurement of individual cognitive abilities requires participants to complete many more trials and/or to use tasks with larger effect sizes than are typical of existing cognitive batteries. This project develops a battery of cognitive control tests enabling efficient and reliable measurement of cognitive control abilities crucial for high performance under time pressure. The test battery is implemented in the Unity game engine, and accessible online using only a web browser with no installation. Gaming mechanics (e.g., variety, feedback, rewards, and a leader board) and an integrated story line maintain engagement over extended and demanding testing sessions. The battery implements most prominent measures of cognitive control including: 1) working memory (single and dual n-back tasks), 2) response inhibition (stop-signal task), 3) conflict tasks (Simon, Flanker and Stroop tasks), 4) multi-tasking, and 5) task switching. The different measures can be flexibly combined within a coherent "room-clearing" narrative, and self-contained tutorials enable easily deployed online testing. Novel versions of the conflict tasks were developed to increase effect sizes and reliability, and they were tested in an online experiment. We develop a rigorous methodology for quantifying the ability of the tests to produce reliable measurements of individual differences and report the results of applying it to data from the

experiment. We conclude that these new conflict tasks produce much more reliable measurement than has previously been achieved.

Acknowledgements

This study was funded by Australian Army Headquarters (Land Capability Division)

1.0 INTRODUCTION

Cognitive capacities and abilities are important in occupations that require superior performance under challenging and stressful conditions, such as sports, civilian high-stakes roles, and the military [1, 2, 3]. This has led to an increasing emphasis on cognitive readiness [4, 5, 6] in order to optimize team performance in the complex missions and socio-technical systems that are an increasingly common feature of modern defence settings [7]. The move away from an emphasis on physical fitness [8] brings with it the need to identify the key psychological constructs underpinning what has been variously termed mental or cognitive fitness [9, 10]. Reliable measurement of individual differences in the key constructs that underpin cognitive fitness is particularly crucial in order to optimize selection, and to evaluate the success or failure of the burgeoning array of cognitive training methods and programs [11, 12, 13].

This paper describes a battery of cognitive tests that measure cognitive abilities crucial for high performance under time pressure. The test battery, which is called "COGMISSION", is implemented in a video-game format using the Unity game engine (https://unity.com). Access is provided via "PlayUR" (https://playur.io), a platform for managing web-based Unity experiments, making COGMISSION accessible online using only a web browser, with no installation required[1]. COGMISSION is augmented with gaming mechanics (e.g., variety, feedback, rewards, and a leader board), and an integrated story line that are designed to maintain engagement over extended and demanding testing sessions. Self-contained tutorials enable easily deployed, wide-scale testing. In the next section we describe the tasks implemented in COGMISSION, the constructs which they measure, and the rationale for their selection. We then review the "reliability paradox" [14], which causes typical methods of testing one of the central constructs for cognitive fitness, attention control, to provide inadequate measurement of individual differences. Following this we describe a rigorous new statistical methodology evaluating the ability of attention control tasks to provide sufficiently reliable measurements of individual differences [15]. Next we report the results of applying that methodology to the results of a pre-registered experiment (https://osf.io/y4sbh) run through Amazon Mechanical Turk (https://www.mturk.com/), a crowdsourcing marketplace for online workers, that evaluates a variety of attention control tasks implemented in COGMISSION.

2.0 COGMISSION

2.1 Overview

The battery implements ten measures of cognitive control: working memory (both single and dual n-back tasks [16]), response inhibition (the stop-signal task [17]), attention control (Simon [18], Flanker [19] and Stroop [20] conflict tasks), multi-tasking [21], task switching [22] and using prior information [23]. Each task, and the way in which it measures a cognitive control construct, is described in detail in the next section.

Choices in each task are made by pressing either the "z" or "/" keys, which correspond to labelled buttons (e.g., "Left" and "Right") on the screen. When performing each task, correct and error responses are

[1] Visit the temporary URL https://playur.io/HFM to register an account and try out a 10-level version of the battery moving through progressively more difficult task combinations. We recommend completing the tutorial first.

followed by different tones. When a correct response is made the corresponding key briefly remains on the screen and is highlighted by an animation. When an error is made the correct key remains on the screen with its label changed to represent the choice stimulus to which it corresponds, so participants are reminded of the correct mapping between stimulus and response. Participants start with a tutorial that provides extensive written and graphic feedback on how to perform the task correctly as part of a guided playthrough, first with prompts for the correct response then without, but requires erroneous responses to be corrected with guiding feedback. The tutorial was found to make it possible to learn to correctly perform the tasks making up COGMISSION without the need for in-person training, and the feedback helped ensure that people did not forget what they learned when later performing the tasks. Once participants understand the tasks, their performance under time-pressured conditions enables measurement of various aspects of cognitive control appropriate to the high-performance context of interest here. Performance can be measured both by accuracy and the time required to make a response (RT = response time).

A points tally is continuously displayed, with 1000 points added for a correct response, and a 1000-point speed bonus when the response was faster than on the previous trial. Gaining points was signalled by both the total incrementing and the reward amount floating up above the total. Responses slower than 2 seconds cause the message “Hurry Up” to flash. If no response was made after 4 seconds the message “Out of Time” appeared. At the end of each game the number of correct responses and the total number of responses are displayed, and at the end of a mission (a set of games), a scoreboard is presented giving the percentage of responses completed and accurate responses and the points total. The provision of points and auditory feedback were found during piloting to be motivating, promoting continued engagement when preforming under time pressure across longer sessions. There is also a facility to provide a bonus payment based on points earned. The bonus earned in a game, along with the total pay-out accrued over all previous missions can be displayed at the end of each mission, although this was not used in the experiment reported in this paper. At the end of an experiment participants are given a Mechanical Turk code which they can use to retrieve a fixed payment for the tasks they completed plus the bonus if applicable.

The various tasks making up COGMISSION can be flexibly combined within a coherent “room-clearing” narrative, with the full suite of tasks involving repeated cycles of:

1) Choosing one of 3 doors contingent on previous choices (choices remembered over cycles constitute the n-back task). Judgments are contingent on either the location of the door or a symbol labelling each door. There are two versions of single 2-back tasks, requiring participants to choose a door which was not used on the previous task.

2) When a door is chosen it becomes the focus of the display and participants make a choice by pressing a button corresponding to the direction of an arrow displayed on the door. Occasionally an alarm sounds shortly after the arrow appears, which indicates the response must be withheld (a stop-signal task).

3) When the stop-signal task is completed (either by responding or after a fixed time with no response), a room is entered. Participants then perform two rounds of choosing the side on which an enemy is hiding based on information in a display at the top of the screen designed to induce Simon, Flanker or Stroop conflict (e.g., Figure 1).

Figure 1. Tutorial test and feedback screens for a Simon conflict trial. The blue square's colour provides decision-relevant information, and the square's location (on the right) conflicting, decision-irrelevant, information.

Repeat vs. change in conflict task over the two testing rounds assesses task-switching costs. Multi-tasking is assessed in two ways, by comparing different types of n-back tasks (see below) and by comparing task sequences with and without the n-back task. When an n-back task is included the last two doors entered must be rehearsed while performing the other tasks, and so requires multi-tasking.

2.2 Detailed Task Description

2.2.1 Working Memory 2-Back tasks

On each trial one of three doors is spotlighted, and participants indicate whether it is the same or different from the door that was highlighted two trials back (the first two trials are warmups to set this up). The current door is always different from the door on the last trial (as 1-back decisions are typically fast and easy [24]), and participants must remember whether either one or two attributes of the highlighted door occurred either two-trials back (same) or three trials back (different). There are three different versions of the working memory task:

a) Single 2-Back Location: The same vs. different decision depends on the location (left, middle, or right) of the door, so participants must hold two items in working memory. Apart from the restriction on 1-back repeats, the next location is chosen randomly, naturally leading to an equal probability of either button being correct on each trial. This is important because otherwise guessing strategies (e.g., quickly picking the most common response) can distort measurement (see [16] for more discussion).

b) Single 2-Back Symbol: The same vs. different decision depends on a symbol (+, O, or X) presented on the highlighted door, so again participants must hold two items in working memory. And again, an equal probability of symbols being chosen on each trial, apart from the 1-back repeat restriction, leads to it being equally probable that each response is correct.

c) Dual 2-Back: The same vs. different decision depends on both the location and symbol, so participants must hold four items in working memory. If the spotlighted door has the same symbol, but not the same location, or the same location, but not the same symbol as two trials back, participants should press the "One same, one different" button. If neither symbol or location is the same, or if both symbol and location are the same as the previous trial, participants should press the "Both same or different" button. This "exclusive or" response rule both makes the task appropriately demanding for a high-performance context and naturally leads to equal probabilities of each response being correct, which is not the case with simpler response rules.

Although the usual assumed capacity of working memory (3-5 items [25]) is not much strained by these

tasks, they are in fact quite difficult because of the demands they make on cognitive control. Participants cannot rely on habitual or automatic responding [26] due to the inconsistent mapping between doors and responses (i.e., a door associated with a given response on one trial can be associated with the other response on other trials). Hence, the task targets higher-level executive control processes to resolve the strong interference between stimulus-response mappings that occurs [27]. These characteristics also mean these tasks are sensitive to key aspects of working memory: flexible updating, maintenance, and interference control.

2.2.2 Stop signal task

The stop-signal task is based on an easy choice: pressing a button corresponding to the direction of an arrow that appears in the middle of the highlighted door. However, that response must be withheld on 25% of trials when a "stop signal" (a red outline around the door) occurs. The time between the appearance of the arrow and the stop signal (the stop-signal delay or SSD) is systematically varied so that withholding a response is successful on around 50% of stop trials using a staircase algorithm: the SSD is increased by 50ms if withholding was successful on the last trial (making it harder to withhold) and decreasing it by 50ms if withholding failed (making withholding easier). This allows estimation of "stop-signal reaction time" (SSRT), a measure of the speed of inhibitory processes (see [30] for a consensus guide to the design and analysis of stop-signal tasks). Because most trials do not require inhibition, participants cannot rely on habitual or automatic processes, as can be done in the other prominent measure of response inhibition, the go/no-go task [27]. This makes the stop-signal task more appropriate for the sort of flexible response inhibition required in dynamic high-performance scenarios.

2.2.3 Conflict tasks

When the door opens, the participant enters a room where an enemy is lurking in the shadows at the back of the room, either on the left or right. The participant's task is to press the button corresponding to the side where the enemy is located in order to neutralize them. Information about the location of the enemy is provided in a "heads up" display at the top of the screen (see Figure 1). In a *Simon task*, a block of either blue or orange is presented on the left or right of the display. Each participant is taught one of the two mapping from colour to side (e.g., orange = right, blue = left). On *congruent* trials the side on which the colour block is presented corresponds to the location of the enemy (e.g., blue on the left or orange on the right). On *incongruent* trials the opposite correspondence holds (e.g., blue on the right or orange on the left), causing conflict between the automatic tendency to respond towards the location of a stimulus and the learned response mapping. The conflict effect, the difference in performance between incongruent and congruent conditions (typically incongruent – congruent RT), measures a participant's ability to exert attentional control among internal representations of tasks stimuli, by focusing on the relevant colour information and ignoring the irrelevant location information.

A second type of conflict is induced in the *Stroop task*, due to interference from the internal representation arising from the automatic tendency to read written words[2]. In the congruent condition either the word "ORANGE" is presented coloured orange, or "BLUE" coloured blue is presented in the middle of the heads-up display. In the incongruent condition the word "ORANGE" is presented in blue or "BLUE" in orange, and again RT in the incongruent condition minus RT in the congruent condition measures the conflict effect. The *Flanker task* requires a different type of control, that of visuo-spatial attention. The side to shoot depends on a central character in the heads-up display, either < for left or > for right. The central character is surrounded by "flankers" that can either be congruent (i.e., >>>>> or <<<<<) or incongruent (i.e., <<><< or >><>>). Control is often conceptualized through a spotlight metaphor; initially the spotlight is wide and lets in the potentially conflicting flanker information, but it is then narrowed to focus on only the relevant target

[2] We induced interference in the English language and gave participants in our experiment a test of English proficiency. This, along with presenting all instructions in English, ensured that Stroop interference would be present in our Mechanical Turk participants, who tend to mainly come from the USA and India.

character as control is exerted. Again, RT in the incongruent condition minus RT in the congruent condition measures the conflict effect. Note that in all three conflict tasks incongruent and congruent trials occur with equal probability and in a random order across trials.

The experiment reported next focuses on improving the conflict tasks, so they provide better measures of individual differences in attention control. The size of the conflict effect can be reduced by strategies that make it less likely that the irrelevant information is encoded. For example, if the Flanker display occurs at a predictable position, attention may be narrowed to the position of the target character before the display appears. To avoid this, on each trial we randomly varied the position of the Flanker display by two character widths. An example of a strategy that is harder to avoid occurs in the Stroop task when participants blur their vision, so the high spatial frequency information required for reading is not available while colour information is preserved. To explore how to generically protect against such strategies, the experiment examined the effect of requiring participants to make a "second shot" based on the irrelevant information on a randomly selected one third of trials. Once a response is made, a silhouette of the enemy appears; on the double shot trials the colour of the enemy changes indicating that it has raised a shield and a second shot is required to neutralize it. This encourages participants to encode the irrelevant information on all trials to gain a large bonus (3000 points) for a correct second shot or loss of points (-2000 points) for an incorrect second shot.

The experiment also explores whether Simon conflict can be combined with each of the two other conflict effects: in the *"Flankon" task*, Flanker displays are presented on either the left or right of the display, and the "*Stroopon" task* displays coloured words that are presented on either the left or the right. As a result, the two sources of potential interference (from flankers/reading + location) can can be congruent for both sources (CC), incongruent for one and congruent for the other (IC or CI), or incongruent for both (II). The four trial types occur equally often with interest focusing on the contrast subject to both interference components (II – CC), as it presumably will produce the largest effect. However, we included the CI and IC conditions in the experiment so that the other two contrasts (IC – CC and CI – CC) can be used to measure the size of each component to determine if they combine additively or in some other way in the double-conflict effect.

2.2.4 Task-switching, using prior information, and multi-tasking.

The task-switching effect addresses a key aspect of cognitive control, the ability to select, establish, maintain, and update task goals. It is typically measured by interleaving two different tasks in a sequence. "Local" task-switching cost is typically measured by the slowing in RT observed when the task changes relative to when it repeats. "Global" task-switching costs are measured by the slowing observed for repeat trials in an interleaved sequence compared to RT on blocks of trials that always use the same task. COGMISSION can be configured to measure both types of cost through having participants perform two conflict trials in a row once the room is entered, both of the same type (e.g., both Simon or both Flanker) or of different types (i.e., Simon followed by Flanker or vice versa). RT on the second task in the latter sequence (a "switch" trial) minus RT for the second trial with the same task in the former sequence (a "repeat" trial) measures local switch cost. Global switch cost is measured by comparing performance on the same task in a game using only task-repeat pairs to repeat trials in a game using both task-switch and task-repeat pairs.

The conflict tasks can also be used to measure another key capability for optimal performance in uncertain conditions, monitoring and adjusting performance to take account of prior experience as well as the current circumstances when making decisions. This is operationalized in COGMISSION by providing participants with the knowledge prior to the beginning of a game that one response in the conflict tasks is more likely than the other (e.g., a left response will be correct 75% of the time and a right response only 25% of the time). Utilization of prior information is reflected in gains in overall RT and accuracy relative to blocks in which each response occurs on half of the trials.

The final cognitive ability measured by COGMISSION, multi-tasking, is also sensitive to the same cognitive control factors as task-switching as well as more general attention capacity and performance monitoring capabilities. Multi-tasking ability is typically measured in a dual-task paradigm, where performance in a primary task alone is compared to performance in that task when a second task is performed simultaneously. Typically, the second task requires working memory representations to be maintained and updated without requiring overt responding (e.g., counting backwards by 3s while making a series of simple or choice responses in the primary task [29]) to avoid introducing an additional response conflict effect. In COGMISSION multi-tasking cost can be measured by comparing performance in each non-working memory task when performed with vs. without the working memory task. Because the working-memory task requires information to be maintained while other tasks are being performed (so it is available when the next set of doors is presented), it is a classic example of the dual-task paradigm.

2.3 Design Rationale

2.3.1 Cognitive control constructs important for high-performance contexts.

After the initial design and implementation of COGMISSION, the two senior authors became involved in a "Delphi" study [30] seeking an expert consensus on the cognitive factors that underpin optimal performance under pressure [31]. The 68 experts who participated made up four panels from different areas: defence, competitive sports, high-stakes civilian roles and applied-cognition research. Over multiple rounds the experts rated and re-rated constructs that were mainly drawn from the National Institute of Mental Health Research domain criteria (RDoC), a neuroscience-based classification framework for research on mental disorders [32] along with additional expert-contributed constructs. Of the 10 RDoC constructs that reached consensus across all panels, 7 were drawn from its Cognitive Systems domain:

1) Attention
2) Cognitive Control—Goal Selection, Updating, Representation & Maintenance
3) Cognitive Control—Performance Monitoring
4) Cognitive Control—Response Selection & Inhibition/Suppression
5) Working memory—Flexible Updating
6) Working memory—Active Maintenance
7) Working memory—Interference Control

The full suite of tasks implemented in COGMISSION addresses all seven constructs. The prominence of working memory is to be expected given its strong association with many measures of scholastic aptitude and fluid intelligence [33, 34], but the particular emphasis on its dynamic aspects in domains 5-7 corresponds strongly to the working memory tasks used in COGMISSION. The stop-signal task maps directly to domain 4, task-switching maps to domain 2, using prior information to domain 3, and multi-tasking to both domains 2 and 3. The conflict tasks particularly map to domain 1, but given its breadth there is also a mapping to multi-tasking, in terms of attention capacity.

2.3.2 Approaches to measuring cognitive control.

Although not pursued here, the mappings identified in the last section could be further examined by exploring the relationship between performance on COGMISSION and on the tests traditionally used to measure performance in these domains in the clinical context. However, it is important to note that although the latter tests often use similar tasks, each task is typically completed in isolation and is usually quite short (i.e., requires the participant to perform a few trials over a short period of time). The latter characteristic is convenient when administering a large set of tests, and a practical constraint required by the limits on sustained engagement associated with many clinical populations. However, as we discuss in detail in the next section with respect to conflict tasks, there is a growing realization that this approach affords results that

are too unreliable (i.e., subject to larger variations over repeated measurement occasions) to be of much use in quantifying stable individual differences.

COGMISSION was designed to address these issues in several ways. Naturally, sustained engagement is less of an issue for populations involved in high-performance activities, but nevertheless, as discussed previously, COGMISSION includes a variety of gaming mechanics that promote engagement. Second, COGMISSION avoids the inefficiencies associated with running a series of separate tests of each construct and leveraging the same relatively small set of tasks in various combinations in a single integrated framework, enabling simultaneous measurement of many different constructs. Further, we argue that high-performance applications typically require the integrated and simultaneous use of many cognitive domains, so measuring these domains in a simultaneous and integrated manner has greater face validity than traditional approaches to testing. This is likely to be particularly important for constructs related to limits on attention capacity, which are notoriously difficult to measure [35], principally because the addition of extra load may only have a measurable effect when done in the context of a high baseline load that limits the amount of spare capacity. Indeed, attention capacity effects absent in simple laboratory tasks have been shown to emerge in more complex real-world settings [36] that are more likely mimicked in COGMISSON that traditional tests.

However, it is important to acknowledge that the approach taken by COGMISSION raises several challenges. First, participants need to have a clear understanding of what is required of them so that variations in performance are related to individual differences in the cognitive domains of interest rather than misapprehensions about how the tasks work. Achieving this understanding can be challenging given the complexity of the combined tasks, particularly in the online environment. We have found well-designed tutorials and the feedback described above to be effective and are presently exploring other avenues that are typically found to be effective, such as part-task training and gradually increasing difficulty [37]. Second, even with the efficiencies gained through COGMISSION's integrated task design, the number of trials required for sufficiently reliable measurement may be problematic in some applications. To provide a realistic assessment of what is possible, a rigorous methodology is required to determine the relationship between the number of trials performed and measurement reliability. We provide and use such a methodology in the next section and apply it to our experiment that aimed to refine COGMISSION's conflict tasks.

3.0 EXPERIMENT

3.1 Test trials and reliability

Tasks used to measure cognitive control, such as the Simon, Stroop and Flanker were developed in the context of experimental psychology where the emphasis is on measuring group differences in very specific constructs while minimizing individual differences in performance. Specificity is aided by performing tasks under homogenous conditions so that variations in performance are mostly due to the factors manipulated in the task. It is also aided by using appropriate measures, for example, measuring conflict by taking the difference between incongruent and congruent RT controls for individual differences in the overall speed of performance[3]. However, this measurement specificity comes at a cost: the variance of a difference is the sum of the variances of each of its components (assuming independence), which inflates measurement error [38]. More generally, the very control that affords the homogeneity among participants that is desirable for experimental research causes range-restrictions that weaken correlations that underpin individual-differences applications, such as selecting suitable candidates for particular roles, or identifying personal characteristics that are correlated with excellence in cognitive abilities. This "reliability-paradox" [14] might be addressed by reducing experimental and measurement controls, but at the cost of making the results more ambiguous.

[3] Other sources of extraneous variability may not be so well controlled by taking differences. For example, [41] shows that it does not control for differences in speed-accuracy tradeoff settings.

Fortunately, a straightforward solution that increases reliability without sacrificing specificity simply involves recording performance over a larger number of trials. This is because fluctuations in performance from trial to trial are the source of the measurement noise that causes unreliability and attenuation of individual-differences correlations, and the effects of measurement noise reduce (and hence reliability and correlations increase) in proportion to the square root of the number of trials. In this sense it is non-sensical to talk about the reliability of a conflict task without specifying the number of trials performed, just as reliability is quantified separately for short and long forms of psychometric tests addressing the same construct. Note that increasing the number of individuals tested does not provide a solution if measurement noise remains high for each person because they perform very few trials. In short, *individual differences cannot be understood if individuals are measured imprecisely*.

The second key ingredient for reliability is the true level of individual differences in the target population. This cannot simply be measured by taking the average of performance over trials for each person and looking at the variance of the average between people, because that confounds individual differences and variance due to measurement noise, unless the number of trials is extremely large. Fortunately, standard linear mixed-model analyses can separately identify these two variance components (individual differences and measurement noise), enabling the contribution of each to reliability to be quantified. Here we a adopt a Bayesian version of this approach using the BayesFactor package in R [38], as proposed by [39]. One of the many advantages of a Bayesian approach is that it affords uncertainty in these and other estimates to be quantified in terms of credible intervals, that is, intervals within which the true value lies with some probability (here we adopt 95% credible intervals, see [40] for a discussion of why confidence intervals produced by frequentist analyses often do not have this simple characterization).

3.1.1 Determining the number of trials required for reliable conflict measurement.

We use the approach developed by [15], where the key quantity is the ratio, η, of the standard deviation of individual differences in the RT conflict effect (CE), CE = RT(incongruent) – RT(congruent), which we denote SD_{ID} to the standard deviation of the measurement noise, SD_N: $\eta = SD_{ID}/SD_N$. The authors [15] applied this analysis to data from 15 standard Stroop, Simon and Flanker tasks (which they refer to as inhibition tasks), and their results were aptly captured in the title of the paper: "Why most studies of individual differences with inhibition tasks are bound to fail". The average value of the ratio was 0.16, with averages with each of the three tasks having similar values, meaning that measurement noise was over six times larger than individual differences. They concluded that the number of trials required in these experiments for adequate measurement of individual differences is an order of magnitude larger than the typical level of 50-100 trials, which is already much larger than the typical number of trials in clinical settings. To provide a concrete comparison for our results, we applied a slightly modified version of these techniques to Flanker and Simon data that combine the 53 participants in Experiments 1 and 2 reported by [14] and Simon data from 102 participants collected in [41], with results shown in Tables 1-3.

Table 1. Effect size and reliability analysis of Flanker data from [14]. See text for definitions of the column headings. The rows give the 2.5%, 50% and 97.5% quantiles of the posterior values (i.e., the middle row is the median estimate, and the top and bottom rows give the associated 95% credible interval). All results are displayed in seconds.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.421	0.455	0.441	0.035	0.032	0.013	0.101	0.139	1.924	2.076	827
50%	0.422	0.456	0.442	0.037	0.034	0.015	0.102	0.156	2.388	2.605	654
97.5%	0.424	0.458	0.443	0.038	0.036	0.017	0.102	0.174	2.867	3.161	528

Table 2. Effect size and reliability analysis of Simon data from [14]. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.397	0.410	0.407	0.039	0.010	0.013	0.131	0.113	0.592	0.706	1262
50%	0.398	0.411	0.409	0.041	0.013	0.015	0.131	0.127	0.868	1.041	993
97.5%	0.400	0.413	0.410	0.042	0.015	0.017	0.131	0.142	1.145	1.391	799

Table 3. Effect size and reliability analysis of Stroop data from [41]. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.591	0.651	0.625	0.046	0.056	0.026	0.188	0.141	1.621	1.750	801
50%	0.594	0.654	0.627	0.048	0.061	0.030	0.188	0.159	2.072	2.255	635
97.5%	0.596	0.658	0.629	0.050	0.065	0.035	0.188	0.177	2.543	2.793	512

Our modification was to apply the analysis to log(RT- 0.25s)[4], as on this scale the RT data was much closer to the Gaussian distribution assumed by the model of noise variability assumed by the mixed-model analysis. We used the BayesFactor *lmBF* function to fit a model with fixed effects for the intercept and conflict effect, random participant intercepts and slopes, where the latter term allows for individual differences in CE. We then obtained 10^4 samples from the posterior distribution of the model's parameters and used them to calculate the median values and 95% credible intervals for the average RT in the congruent and incongruent conditions, the intercept (I) and its standard deviation (SD_I), the conflict effect (CE) and its standard deviation (SD_{ID}), measurement noise (SD_N), η, effect size relative to measurement noise (i.e., CE/SD_N) and individual differences (i.e., CE/SD_{ID}) and an estimate of the combined number of congruent and incongruent trials required for adequate measurement (n). We defined the latter by requiring that SD_{ID} be at least twice the standard error of the conflict effect, $SD_N (2/n)^{1/2}$. As can be seen, although the effect sizes are substantial in all cases, and consistent with [15], the values of η are quite small because measurement noise is much larger than the individual differences in the conflict effect, and so very large number of trials are required for adequate measurement. We now report analogous results for COGMISSION.

3.2 Methods

3.2.1 Design and Sampling Plan

All participants first performed a tutorial requiring them to complete guided trials (with feedback) to familiarise themselves with the task they would subsequently undertake. This included both prompted and unprompted trials, with all providing feedback regarding choice accuracy. Participants completed this procedure for the basic task alone before introducing any additional components for the double shot conditions (e.g., Flanker alone, followed by Flanker with double shot trials). Participants then moved on from the tutorial and played 9 games of 12 trials. One set of participants performed Flanker-based tasks, the standard Flanker task (F), the Flanker task with double shot trials (F2) and the Flankon task with double shot trials (FO). The other group performed the Simon-based task, the standard Simon task (SI), the Simon task with double shot trials (SI2), and the Stroopon task with double shot trials (SO). We do not report any analyses of the double shot responses except to note that accuracy was well above chance (~70%). Each set initially contained three games in the fixed order F, F2, FO or SI, SI2, SO; this data was discarded as practice. Participants then played two sequential games of each of the three tasks. The order of the task pairs was counterbalanced over 6 groups of participants in each set, so in total there were 12 between-subject conditions. The counterbalancing ensured that the three conditions could be compared without confounding from order effects.

[4] Following the analysis by [39] of the Flanker and Stroop data (and discarding data from the neutral condition as they did) we removed all RTs less than 0.25s and greater than 1.5s from all three data sets, but we did not remove error RTs as they did.

The pre-registered sampling plan was determined using the sequential Bayes factor method [41] with a minimum sample size of 72 in each set and a maximal sample size of 216. Stopping data collection was based on a Bayesian repeated-measures *t*-test comparing RT conflict effect in the standard and double shot tasks performed with the default settings of the *ttestBF* function with the stopping criteria being a Bayes Factor in favor of a difference (BF_{01}) being greater than 10 or less than 1/6. For both sets this criterion was immediately fulfilled at 72 participants in favor of the null (BF_{01} = 0.134 for the Flanker-based tasks and BF_{01} = 0.141 for the Simon base tasks).

3.2.2 Participants

The total time for participation being around 20 minutes, for which participants were paid US$2. 242 participants attempted the task. Following the pre-registered procedure participants were discarded if a) Their overall accuracy was less than 60%; b) More than 10% of responses were fast guesses (RT < 0.1s) or more than 10% of RTs were greater than 1.5s. On these criteria 60, 1, and 37 participants, respectively, were discarded before the final sample of 144 participants was obtained.

3.3 Results

Figure 2 shows the results for RT and accuracy. Although accuracy was not of primary interest here, Figure 2 shows that incongruent accuracy was less than congruent accuracy, and so the conflict effect in RT cannot be attributed to a speed-accuracy tradeoff. The conflict effect is clearly much bigger for the Flanker than Simon tasks and although the double shot slows overall RT, it does not affect the size of the conflict effect. In the double conflict task, the Simon component of the conflict effect is weakest, followed by the Stroop component and then the Flanker component. In the Flankon task, the double conflict effect is sub-additive so that it is only slightly larger than the Flanker effect alone. In the Stroopon task the opposite pattern appears, with the double conflict effect being at least as large as the sum of its component Stroop and Simon effects.

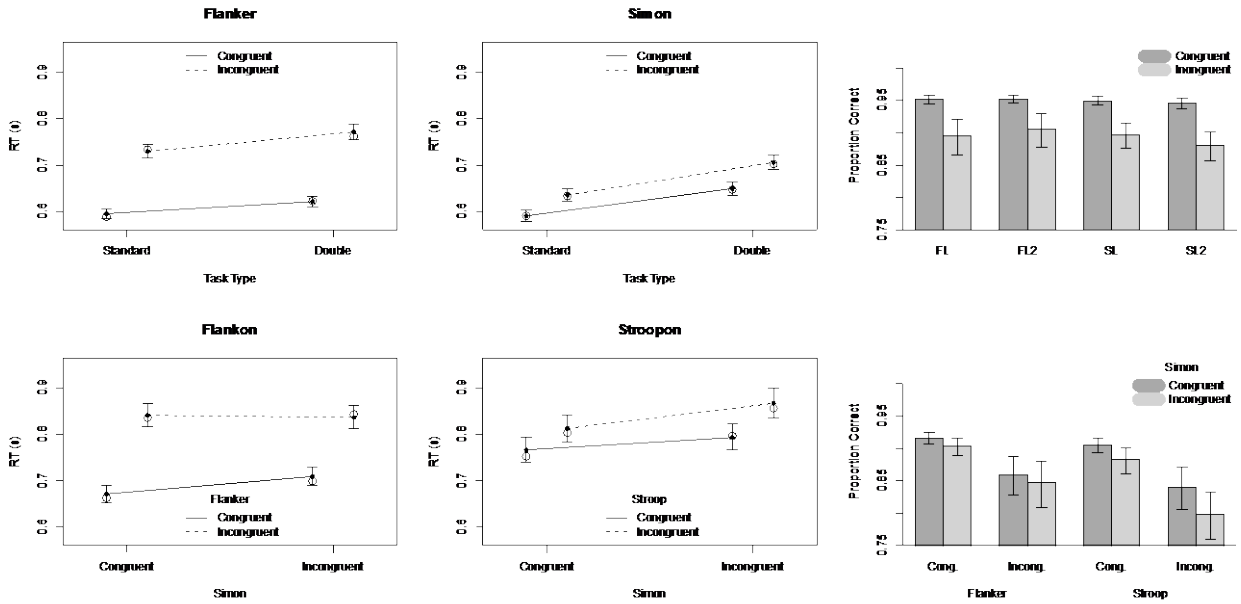

Figure 2. RT and accuracy results. For RT the data are represented by open circles and the fit of the linear mixed model by lines joining solid points with 95% credible intervals. The error bars for the accuracy analysis are 95% confidence intervals calculated on the probit scale applied to edge-corrected accuracy estimates [42].

Table 4. Effect size and reliability analysis of the COGMISSION Flanker task. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.585	0.715	0.676	0.130	0.123	0.080	0.145	0.392	1.337	1.658	104
50%	0.595	0.730	0.686	0.140	0.142	0.097	0.148	0.498	1.641	2.133	65
97.5%	0.606	0.744	0.696	0.151	0.161	0.116	0.152	0.600	1.969	2.786	44

Table 5. Effect size and reliability analysis of the COGMISSION Flanker task with second shots. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.610	0.756	0.709	0.130	0.138	0.089	0.155	0.365	1.380	1.753	120
50%	0.622	0.771	0.719	0.142	0.160	0.108	0.159	0.464	1.698	2.281	74
97.5%	0.633	0.788	0.731	0.155	0.181	0.131	0.163	0.569	2.017	2.969	49

Table 6. Effect size and reliability analysis of the COGMISSION Flankon task with second shots comparing double congruent and double incongruent conditions. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.652	0.812	0.767	0.149	0.143	0.116	0.170	0.389	1.065	1.474	106
50%	0.670	0.836	0.784	0.165	0.178	0.143	0.177	0.531	1.358	2.063	57
97.5%	0.689	0.861	0.801	0.183	0.212	0.174	0.186	0.692	1.659	2.898	33

Tables 4-6 display the results of the effect size and reliability analysis for the Flanker-based tasks. For the Flankon the effect comparing the double congruent to double incongruent conditions is reported. All conflict effects and η values are much larger than in Table 1 for the data from [14]. Consequently, the trial numbers required are much lower. There is only a modest increase in conflict between the Flanker and double shot Flanker, and, if anything, a decrease in reliability, due to a larger increase in the measurement noise than individual differences. The increase in the conflict effect caused by combining Simon and Flanker interference is also modest but is associated with a clearer increase in reliability due to a larger increase in individual differences than in measurement noise.

Tables 7 shows that the Simon task produces a much weaker conflict effect, and has lower reliability, than all Flanker-based tasks. Adding a second shot produces only a small increase in the conflict effect, but a larger increase in reliability.

Table 7. Effect size and reliability analysis of the COGMISSION Simon task. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.578	0.621	0.625	0.123	0.026	0.066	0.166	0.253	0.356	0.554	250
50%	0.591	0.635	0.636	0.134	0.046	0.081	0.169	0.324	0.604	0.977	152
97.5%	0.603	0.650	0.646	0.147	0.068	0.099	0.173	0.407	0.854	1.472	97

Table 8. Effect size and reliability analysis of the COGMISSION Simon task with second shots. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.635	0.690	0.688	0.127	0.036	0.076	0.174	0.295	0.396	0.567	184
50%	0.649	0.706	0.700	0.139	0.059	0.094	0.177	0.383	0.647	0.950	109
97.5%	0.664	0.722	0.712	0.151	0.081	0.114	0.181	0.480	0.904	1.431	69

Tables 9 displays results for the double interference effect in the Stroopon task. The conflict effect is much larger in this case and reliability falls between that of the Simon and Flanker tasks. To better understand this result Tables 10 and 11 report results for the two components of the double interference Stroopon effect. The Simon component is weaker than in Table 7, but individual differences and, hence, reliability are greater. The Stroop component is greater both in terms of effect size and reliability. Comparison to Table 9 suggests that, in contrast to the Flankon task, the components have a more than additive effect in the Stroopon, but this does not result in any improvement in reliability.

Table 9. Effect size and reliability analysis of the COGMISSION Stroopon task with second shots comparing double congruent and double incongruent conditions. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.739	0.835	0.824	0.161	0.059	0.135	0.208	0.327	0.412	0.652	149
50%	0.765	0.867	0.847	0.185	0.105	0.169	0.217	0.441	0.672	1.149	82
97.5%	0.791	0.900	0.871	0.213	0.152	0.209	0.229	0.585	0.928	1.772	47

Table 10. Effect size and reliability analysis of the Stroop component of the COGMISSION Stroopon task with second shots. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.740	0.784	0.804	0.169	0.008	0.139	0.210	0.375	0.049	0.074	114
50%	0.766	0.813	0.826	0.191	0.053	0.173	0.219	0.509	0.296	0.456	62
97.5%	0.794	0.842	0.848	0.217	0.097	0.215	0.231	0.659	0.540	0.878	37

Table 11. Effect size and reliability analysis of the Simon component of the COGMISSION Stroopon task with second shots. See Table 1 for definitions and units.

	Congruent	Incongruent	I	SD_I	CE	SD_{ID}	SD_N	η	ES	ES_{ID}	n
2.5%	0.741	0.766	0.789	0.152	-0.016	0.126	0.202	0.331	-0.054	-0.091	146
50%	0.766	0.793	0.809	0.172	0.026	0.156	0.210	0.452	0.186	0.310	78
97.5%	0.793	0.822	0.831	0.195	0.068	0.192	0.220	0.593	0.426	0.749	46

3.4 Discussion

Whether with a second shot or not, both the COGMISSION Simon task and particularly the Flanker task produced larger conflict effects than those in [14]. The Stroop component of the COGMISSION Stroopon task had a similar conflict effect to [41]. However, every COGMISSION task displayed a much greater level of individual differences with at most a modest increase in measurement noise, so that the η ratio was increased by a factor of 2-3 times and the required number of trials decreased by almost an order of magnitude. These results support the conclusion that the video-game instantiation of conflict tasks in COGMISSION is much more effective for individual differences research than traditional versions of these tasks. However, further research is required to better understand which factors in the task design are responsible for these improvements.

There was only modest evidence that adding a second shot or a double conflict effect improved effect sizes and increased reliability. One possibility is that the within-subject nature of the manipulation caused carryover effects, so that the presence of double shots during practice caused participants to avoid strategies for reducing the encoding of irrelevant information in later games even when there were no double shots. A follow up experiment is being performed to check this possibility by making all manipulations between-subjects.

The results for the Stroopon task were more encouraging than for the Flankon task. Similarly, the results for

the Stroop component of the Stroopon task were clearly better than for the purely Simon tasks. As COGMISSION requires two conflict tasks to enable measurement of other cognitive control constructs, future experiments will investigate replacing the Simon task with versions of either the Stroop or Stroopon task that are more reliable.

One limitation of the current results is the high attrition rate of participants due to low accuracy and slow responding. We have now implemented a tutorial which affords participants the opportunity to have multiple attempts to achieve a criterion number of correct responses before they can engage in the full experiment. Preliminary results suggest a much-reduced attrition rate, suggesting that many of the problems in the experiment reported where were due to participants not engaging sufficiently with the tutorial to clearly understand what they had to do before commencing the more demanding main task.

4.0 REFERENCES

[1] Baker, H. K., and Phillips, A. L. (2000). Knowledge, skills and attributes needed to succeed in financial management: Evidence from entry-level and mid-level practitioners. *Financial Practice Education,* 10, 189–200.

[2] Fletcher, J. D., and Wind, A. P. (2014). 'The evolving definition of cognitive readiness for military operations, pp. 24–52, in *Teaching and Measuring Cognitive Readiness*, H. F. O'Neil, R. S. Perez and E. L. Baker. New York, NY: Springer.

[3] Herzog, T. P., and Deuster, P. A. (2014). Performance psychology as a key component of human performance optimization. *Journal of Special Operation Medicine, 14*, 99–105.

[4] Morrison, J. E., and Fletcher, J. D. (2002). *Cognitive Readiness (P-3735).* Alexandria, VA: Institute for Defense Analyses.

[5] Grier, R. A. (2011). Cognitive readiness at the tactical level: A review of measures. *Proceedings of the Human Factors Ergonomics Society*, 55, 404–408.

[6] O'Neil, H. F., Perez, R. S., and Baker, E. L. (2014). *Teaching and Measuring Cognitive Readiness.* New York, NY: Springer.

[7] Bowers, C., and Cannon-Bowers, J. (2014). Cognitive readiness for complex team performance, in *Teaching and Measuring Cognitive Readiness*, pp. 301–323, H. F. O'Neil, R. S. Perez, and E. L. Baker New York, NY: Springer.

[8] Sotos, J. G. (2019). Physical fitness programs don't fit today's fight. *U.S. Naval Institute Proceedings*, 145:1396.

[9] Cornum, R., Matthews, M. D., and Seligman, M. E. P. (2011). Comprehensive soldier fitness: Building resilience in a challenging institutional context. *American Psychologist*, 66, 4–9.

[10] Aidman, E. (2020). Cognitive Fitness Framework: Towards assessing, training and augmenting individual-difference factors underpinning high-performance cognition. *Frontiers in Human Neuroscience, 13*, 1752–9.

[11] Adler, A. B., Bliese, P. D., Pickering, M. A., Hammermeister, J., Williams, J., Harada, C., et al. (2015). Mental skills training with basic combat training soldiers: A group-randomized trial. *Journal of Applied Psychology, 100*, 1752–1764.

[12] Blacker, K. J., Hamilton, J., Roush, G., Pettijohn, K. A., and Biggs, A. T. (2019). Cognitive training for military application: A review of the literature and practical guide. *Journal of Cognitive Enhancement, 3*, 30–51.

[13] Redick, T. S. (2019). The hype cycle of working memory training. *Current Directions in Psychological Science, 28*, 423–429.

[14] Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Psychonomic Bulletin & Review, 50*, 1166–1186.

[15] Rouder, J. N., Kumar, A., & Haaf, J. M. (submitted). Why most studies of individual differences with inhibition tasks are bound to fail. https://psyarxiv.com/3cjr5/

[16] Heathcote, A., Coleman, J. R., Eidels, A., Watson, J. M., Houpt, J., & Strayer, D. L. (2015). Working memory's workload capacity. *Memory & Cognition, 43*, 973–989.

[17] Matzke, D., Verbruggen, F., & Logan, G. D. (2016). The Stop-Signal Paradigm. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (4 ed., Vol. 5, pp. 1–64). John Wiley & Sons, Inc.

[18] Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica, 136*, 189–202.

[19] Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition, 2*, 101–118.

[20] MacLeod, C. M. (1991). Half a Century of Research on the Stroop Effect: An Integrative Review. *Psychological Bulletin, 109*, 163–203.

[21] Watson, J. M., & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review, 17*, 479–485.

[22] Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*, 134–140.

[23] Garton, R., Reynolds, A., Hinder, M. R., & Heathcote, A. (2019). Equally flexible and optimal response bias in older compared to younger adults. *Psychology and Aging, 34*(6), 821–835.

[24] McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 817–835.

[25] Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*, 51–57.

[26] Schneider, W., Shiffrin, R., 1977. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1–66.

[27] Shiffrin, R. M., Schneider, W., 1977. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*, 127–190.

[28] Verbruggen, F., Aron, A. R., Band, G. P., Beste, C., Bissett, P. G., Brockett, A. T., et al. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *Elife, 8*, e55–26.

[29] Castro, S. C., Strayer, D. L., Matzke, D., & Heathcote, A. (2019). Cognitive workload measurement and

modeling under divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, 45, 826–839.

[30] Brown, B. B. (1968). Delphi process: A methodology used for the elicitation of opinions of experts. Rand Corporation, Santa Monica, CA.

[31] Albertella, L., Kirkham, R. Advisory Group, Aidman, E. & Yücel, M. (in preparation). Building a transdisciplinary expert consensus on the neurocognitive drivers of performance under pressure: An international Delphi study.

[32] Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders, *American Journal of Psychiatry*, *167*, 748-51.

[33] Engle, R.W., Tuholski, S.W., Laughlin, J., & Conway, A.R.A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable model approach. *Journal of Experimental Psychology: General*, *128*, 309–331

[34] Cowan N., Fristoe, N.M., Elliott, E.M., Brunner, R.P., Saults, J.S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition*. *34*, 1754–1768.

[35] Navon, D. (1984). Resources--a theoretical soup stone? *Psychological Review*, *91*(2), 216–234.

[36] Boag, R. J., Strickland, L., Loft, S., & Heathcote, A. (2019). Strategic attention and decision control support prospective memory in a complex dual-task environment. *Cognition*, *191*, 103974.

[37] Wickens, C. D., Hutchins, S., Carolan, T., Cumming, J. (2013), Effectiveness of part-task training and increasing-difficulty training strategies: A meta-analysis approach. *Human Factors*, *55*, 461–470.

[38] Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*, 508–535.

[38] Richard D. Morey and Jeffrey N. Rouder (2018). BayesFactor: Computation of Bayes Factors for Common Designs. R package, version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

[39] Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*, 452–467.

[40] Morey, R. D., Hoekstra, R., Rouder, J. N., & Lee, M. D. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

[41] Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

[42] Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50.