José Proença · Raul Fervari ·
Manuel A. Martins · Reinhard Kahle ·
Graham Pluck (Eds.)

LNCS 15551

Software Engineering and Formal Methods

SEFM 2024 Collocated Workshops

ReacTS 2024 and CIFMA 2024
Aveiro, Portugal, November 4–5, 2024
Revised Selected Papers

Springer

Executive Cognitive Control of Free Choices

Graham Pluck, Fei Gu, Natasha Asawanuchit, and Suphasiree Chantavarin(✉)

Faculty of Psychology, Chulalongkorn University, Bangkok, Thailand
{graham.ch,fei.g,Suphasiree.C}@chula.ac.th,6338022238@alumni.chula.ac.th

Abstract. Originating in computer science in the 1950's, executive function is now an important concept in behavioral sciences. This Tool paper examines the core definitions of executive function, and how that relates to free, willed choices in human behavior. We contrast this with cognitive assessment methods that tend to push test takers into convergent thinking. We show how a common form of cognitive test used in behavioral sciences to measure executive functioning, the Trail Making Test, can be altered so that it requires divergent thinking. To analyze and summarize performance of multiple, individual, free choices we apply statistical methods taken from computer science to test for randomness. The tool presented, the Choice Trails Test, and the proposed analysis method, allow for novel ways to investigate top-down, executive, cognitive control using a simple paper-and-pencil test. The benefit of this approach is that it produces indices of performance that are closely aligned with the essential meaning of executive functions. Additionally, this method provides a denser data set than traditional methods that examine total performance metrics. Denser data allows for analysis that is consistent with traditional approaches to examining task performance in cognitive science that stress continuous analysis of processes across tasks.

Keywords: Executive function · Divergent thinking · Action selection · Trail making test · Task switching · Willed action · Free choices

1 Executive Cognitive Control

The concept of executive controllers, programs that oversee other programs, originated in computer science in the 1950's, but was later adopted widely by neurological and cognitive sciences [34]. The modern concept of executive control in cognitive science has thus developed from two fields. Firstly, it has been used widely in neuropsychology to functionally describe disorganized behavior seen after damage to the frontal lobes of the primate brain [38]. Secondly, the concept of a supervisory attentional controller, or central executive, has been included in highly influential cognitive models developed in experimental psychology [3,27]. Despite the concept of executive control spreading to a range

J. Proença et al. (Eds.): SEFM 2024, LNCS 15551, pp. 237–253, 2026.
https://doi.org/10.1007/978-3-031-94748-3_18

of other disciplines, the influence of neuropsychology and experimental cognitive psychology has defined how the executive functions are conceptualized and measured in quantitative cognitive research.

The vast majority of assessments of executive functions used in diverse fields including, education, linguistics, public health, human-resources management etc. have followed the experimental psychology influence in attempting to quantify executive control (essentially a latent variable) by measuring accuracy of performance on demanding, yet highly constrained tasks.

Take for example the various towers tasks, such as the Towers of Hanoi, which are widely used to measure planning as an executive function [44]. There are various Tower tasks which include Tower of London and Tower of Hanoi; however, they are quite similar in that they both require transforming an initial configuration into to a goal state, so for the purposes of this paper they will be considered as examples from the general class of Towers tasks. Towers tasks involve moving from a prespecified start state of different sized disks on any of three pegs, to a predefined finishing goal state configuration. It is cognitively demanding because there are strict rules concerning how the disks can be moved. To perform efficiently, the moves must be planned several steps in advance, with goal states decomposed into sub-goals. In fact, it is the highly constrained nature of such tasks, and limited search space, that has made them of substantial interest also to artificial intelligence [22,50].

But how well does the manifest performance on such tasks relate to the concept of an executive controller? Performance on towers tasks have, in fact, been interpreted as indicating a wide range of cognitive processes including planning [44,50], spatial working-memory [30], or resolution of sub-goal/goal conflicts [15], among others. Similar ambiguity of process issues affects other commonly used tests of executive function, such as the Stroop task and n-back tasks, raising problems of what is exactly being studied by these executive function tests [37]. At this stage it is necessary then, to consider in more detail the definition of executive control.

1.1 Defining Executive Control

Executive function is the term often used in cognitive psychology to describe the top-down processes underlying control of action. Specifically, it refers to processes underlying goal-directed action that is required to face non-routine challenges [34,45]. Furthermore, it has to be more than just goal-directed, it has to produce 'intelligent' outcomes, as some goal-directed mechanisms can nevertheless be incapable of adaption [24].

Within cognitive neuroscience, the term *cognitive control* is often preferred but refers to the same idea, defined as: "Cognitive, or executive, control refers to the ability to coordinate thought and action and direct it toward obtaining goals.... Executive control contrasts with automatic forms of brain processing." [25, p. 99]. As mentioned above, one of the historical reasons for the concept of executive controllers moving from computer sciences into cognitive neurosciences, was the application of the concept of executive cognitive control

to understanding behavioral disorders seen after experimental damage to the frontal lobes in non-human primates [38].

As an extension of that, from a clinical perspective, cognitive control is used to explain deficits seen in human patients after damage to the frontal lobes [33,34]. As an example, take this clinical description: "one straightforward difficulty common after frontal lesions is defective control of behavior in the face of choice, complexity, or ambiguity" [1, p. 1515]. The same authors also offer a further conceptual definition: "Cognitive control is required when...a stimulus is ambiguous and potentially conflicting responses might be generated". Similarly, definitions of executive control from developmental psychology emphasize that they represent top-down control of cognition when the correct responses are 'ambiguous' [1,18,41]. In this sense, ambiguous means that appropriate behavior cannot be directly driven by sensation.

It is well known that definitions of executive function and cognitive control used in behavioral sciences are vague and variable [4]. Nevertheless, key aspects are their top-down coordination of intelligent goal-directed behavior, in non-routine situations, and in contexts in which the appropriateness of stimulus-response associations are ambiguous. Indeed, the behavioral outcome of such executive control processes are often referred to as being 'willed', as opposed to being automatic and stimuli-driven [14,20,25,27,42]. How well then does such an analysis describe the task goals of common laboratory and clinical assessments of executive functions used in clinical neuroscience and psychology?

If we stay with the towers tasks as an example, we can clearly see why performance of them is often considered a measure of executive cognitive control. They are certainly goal-directed, one of the key, defining features of executive control, in that the task is to move from a start state to a goal state, and there are many choices that need to be made. And at least on the first attempts, they are non-routine and cannot be completed through automatic routines triggered by stimuli.

However, on closer examination, we can see that participants in research studies, or clinical patients, do not just perform a single tower task that is novel to them. In order to obtain scores with a wide dispersion across individuals, usually multiple trials are performed, each with different start and end goal states. Typically, between 8 and 20 different trials are performed per person, involving potentially hundreds of separate moves. Total scores are calculated based on using the fewest number of moves possible to reach the goal states. There are in fact simple routines that can be applied to efficiently achieve the end goal states, and people do spontaneously apply them [46,50]. Furthermore, substantial learning occurs during task performance. Some of the learning is procedural, but also declarative discovery of rules, which means that people can effectively identify and apply schematic routines to achieve the goal state on each trial [48]. For this reason, towers tasks, as they are typically analyzed, tend to be actually quite unreliable measures of processes supposedly under executive cognitive control [32].

The problem may be that most quantitative measures of executive cognitive control have originated in laboratory-based experimental psychology. Within that field it is very common for test procedures to constrain the response space and to classify and score all responses as correct or incorrect. The measure of performance is then simply the total accuracy. Consequently, such lab-based tasks are inherently convergent, in the sense proposed by Guilford. He defined convergent cognitive processing in this way: "In convergent thinking, there is usually one conclusion or answer that is regarded as unique, and thinking is channeled or controlled in the direction of that answer" [16, p. 274]. Reasoning using deductive logic is a classic example of convergent thought. Most cognitive tests used in experimental psychological research or clinical practice channel performance in ways that meet the definition for convergent thinking given by Guildford. Responses are essentially scored as being right or wrong, according to predefined criteria. Even in tasks in which response time is taken as a variable of interest, it is still inevitably the time taken to produce the unique response that is considered correct by the experimenter.

1.2 Divergent Thinking and Executive Control

One of us has previously argued that the conceptual definitions of executive functions are often more consistent with tasks that involve divergent thinking [34]. Divergent thought, the antipode to convergent thought, was originally defined by Guilford in terms of task-related processing in which "...there is much searching or going off in various directions. This is most clearly seen when there is no unique conclusion." [16, p. 274].

Guilford gives the task of verbal fluency as an example of a divergent thinking task [16]. This is in fact one of only a few examples of tests used with the intention of measuring executive functions that clearly involves divergent thought. In verbal fluency tasks, people are asked to think of as many words as possible that meet a criterion, within a short period, usually one minute.

Phonemic fluency involves producing words beginning with pre-specified letters, in English often the letter 'F'. Similarly, category fluency involves producing words within a predefined semantic set, in English 'animals' is the most commonly used set. Another example of a test described by Guilford as an exemplar of divergent thinking is production of alternative uses for objects, most commonly a brick is the target, and the participant is required to produce as many possible uses as possible (e.g., a door stop, to crush cans for recycling...). Together the tests as described here, and others such as gestural fluency, are well-known to be sensitive to damage to the frontal lobes of the brain and are considered tests of executive cognitive control involving voluntary generation of responses [40].

It appears that the human cognitive system finds fluency tasks, such as verbal fluency, difficult because retrieval in that way is an unusual task requirement, and we thus lack routines to do so, necessitating top-down executive control. Evidence to support this interpretation comes from the observation that in verbal fluency tasks people spontaneously cluster items that they recall, and frequently

switch cluster types. In comparison, patients with cognitive impairment caused by dementia produce fewer, and smaller clusters. This is interpreted as indicating a loss of volitional, spontaneous strategy application [5]. Furthermore, verbal fluency is impaired the most for sets with large numbers of items [11], suggesting that search strategies are the limiting factor, not availability of lexical items.

Another rare example of an executive function assessment method that invokes divergent thinking is the Hayling Sentence Completion Test [7]. The test consists of two sets of 15 sentences each having the last word missing. The second part requires participants to quickly complete the sentence with a word that does not make any sense within the sentence context. The free choice aspect of this task makes it difficult as there are so many possible words to choose from, even neurologically healthy people struggle and tend to make errors by reverting to routines, in that they give words that do in fact make sense. Mounting evidence suggests that performance on this test is much more closely associated with real-life performance in challenging environments than conventional, convergent cognitive tests [35,36], suggesting that whatever it measures conforms well to that expected from the concept of an executive controller.

We argue that tasks such as described here which promote voluntary, divergent thinking, where constraints are ambiguous because decisions can go in unforeseen directions, are better at eliciting measurable behavior that conforms to the conceptual definitions of 'executive cognitive control', at least, as opposed to the majority of tests used in cognitive research and clinical practice, which are decidedly convergent. This is because tasks that require divergent thinking generally do not allow for routine, automatic processing. In fact, they are highly executive because they measure free choices.

1.3 Free Choices in Experimental Tasks

The reason that quantitative cognitive research has generally avoided addressing free choices is that it is difficult to operationalize behavioral experiments to measure them. Modern cognitive psychology is extremely experiment based. Approximately 97% of all published cognitive psychology articles describe experiments [49]. In cognitive psychology, experimentation is viewed in terms of stimuli and response—the experimenter manipulates some variable (the stimulus) and observes the effect on behavior (the response).

But willed actions, the behaviors said to result from executive cognitive control [14,20,25,27,42], by definition are not stimuli-driven. In the cognitive psychology laboratory then, the standard stimulus-response experimental design is of little use. If an experimenter asks a research participant to make free choices, perhaps lift a finger whenever they want to, then the response cannot be readily categorized as correct or not, nor the response time from will to action calculated.

In cognitive neuroscience this is less of a problem, as physiological measures are taken as the response. This was demonstrated in one of the earliest functional brain imaging studies, in which it was shown that free choice finger movements activate the frontal lobes of the brain [14]. In fact, they activated the exact same subregion which had been identified, and is still recognized, as the neurological

hub of executive cognitive control [29]. Furthermore, which willed action will be made can be predicted at the neurophysiological level before the decision is made by executive control [42]. This is because free choices appear to be influenced, at least partly, by random noise of neuronal firing—if one set of cells are randomly more active at a particular time point, then they are more likely to influence the outcome when a decision is called for. In this sense, free choices are difficult to maintain and require top-down control.

Free choices could potentially be used as a behavioral measure of top-down executive cognitive control if the convergent paradigm is not used. Instead of accuracy of responses, one could measure the ability to override decision making that is driven by factors such as routines, stimuli-response associations, and neuronal noise. In a free-choice paradigm a research participant could be asked to respond randomly and to not plan ahead, but still within a constrained task. However, deviation from routineness in task performance is more difficult to measure than accuracy, and it is an undeveloped field of cognitive research. In the following section we describe a novel task, and a mathematical method to quantify performance, which targets how well participants can avoid patterns in their free choices. This procedure also allows for the collection of 'dense' data [47], making it more amenable to a detailed cognitive analysis.

In the remainder of this paper, we describe a novel method for collecting data on free choices, as well as a suggested statistical approach for its analysis (Sect. 2). We then describe an example of the analysis using a sample of data we collected (Sect. 3). The paper finishes with a discussion on the wider context of the research reported, including applications and implications of this novel approach (Sect. 4).

2 A Method for Eliciting Free Choices as Behavioral Data

As previously described, the majority of laboratory tests of cognitive function promote convergent thinking, encouraging research participants to produce predefined correct responses. As another example of this we could examine the Trail Making Test [39]. This has been widely in use in clinical and educational cognitive assessment since the 1950's. It is a paper-and-pencil test that involves participants being presented with a page that has 25 circles marked on it. Each of the circles contains a number (from 1 to 13) or an English alphabetic letter (from A to L). The task is to draw lines as quickly as possible to join the circles consecutively, but alternating number and letter sequences (i.e., 1-A-2-B-3- etc.).

There are numerous versions of this test [10,39] and also task modifications, one of the most common modified forms is the Color Trails Test which dispenses with the alphabetic letters and instead requires participants to switch between joining pink and yellow circles [23]. This produces a more culture-fair test, in that knowledge of the English alphabet is not needed for task completion. But it necessitates that foils be provided- each number is shown twice, once in pink and once in yellow. Both the standard version [10,39] and Color version [23] require

convergent processing, as only one of the circles is ever considered correct as the target of the line. We took this basic design but altered it to allow divergent, responding via free choices of color.

2.1 The Choice Trails Test

To allow free choices, A4 size pages were produced which contained the numbers 1 through 25. However, each number was shown four times, each time in a different color. The same pink and yellow as the original test were used, but additionally blue and violet circles were included. The four colors were selected to have different brightness levels, so that they would be distinguishable even for color-blind people. A sample task is shown in Fig. 1. The basic task requirement is that participants must join the circles in numerical sequence, starting from 1, finishing at 25, choosing a different color each time. This and other rules are described in more detail below.

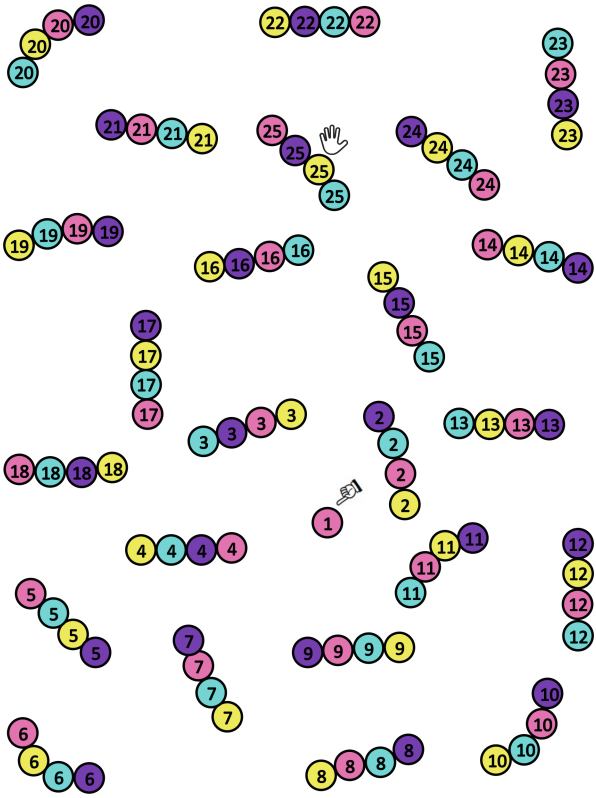

Fig. 1. The Choice Trails Test. (Color figure online)

When a participant performs the task correctly the lines that they draw will not transect each other. For this task the completion of each page (i.e., joining all numbers from 1–25) is considered a single trial. Within each trial 24 free choices are made. We refer to each of these as 24 movements as steps. As the starting point is a pink circle, the first step made within a trial is to choose whether to select a yellow, blue, or violet circle 2. This continues until the participant completes the 24th step (connecting circle 24 to circle 25) with their pencil. To allow for multiple trials by the same participants, multiple versions were made with the circles in different configurations on each page. All were very similar, containing the same 4 colors, and each requiring 24 steps to complete. The minimum line length to make all 24 steps was within 5% of the total distance on all versions. A set of materials for 8 trials are available to download from https://gpluck.co.uk/Tests/.

When completing a trial of the task, an individual sits at a table and the Choice Trails test is placed in front of them. The test is obscured by the experimenter's hand while the participant listens to the verbal instructions of the task. The participant is given a pencil and told that they must draw lines to connect all of the circles starting at circle 1. The specific rules used were:

1. They should join the numbers in sequence with the pencil.
2. They should perform as quickly as possible without making mistakes.
3. They should not choose the same color twice in succession.
4. They should try to choose all the colors equally often.
5. They should avoid using any plans or strategies.

On the experimenter's instruction to begin, a stopwatch is started. The experimenter watches performance and if the participant breaks a rule, such as missing out a number, they are stopped and told to continue from the last point before the error was made. The stopwatch is left running until the participant completes the final drawn line (completing step 24, terminating on circle 25). The completion time is recorded. To create the data set for the current study, three trials were completed by a sample of 30 participants. This sample size is sufficiently powered to detect large effect sizes, according to standardized criteria. The participants were all undergraduate students. In addition to task completion times all choices from all steps were tabulated in sequence. As there were 24 steps made on each trial, this totaled 72 choices per research participant, and 2,160 choices recorded in total from the sample. This is an ongoing study and the analyzed sample will be larger in future studies on this assessment tool.

2.2 A Method to Test the Randomness of Responses

In the task described above, the requirement faced by the research participants is to make choices that will use all colors equally often, without using any strategy. By this we mean the commonly understood meaning of strategy, being the application of a plan to achieve a goal. This rule was introduced as, from our experience in cognitive testing, research participants often do spontaneously

apply simple strategies, such as repeating patterns. Although, from a philosophical perspective the concept of a strategy could be interpreted in many other ways, we expected participants to understand the instruction as to not apply explicit, conscious plans to meet the task requirements. Therefore, although not explicitly instructed, the task implicitly required them to try to produce random free choices.

From a data analysis perspective we have three trials of performance, each with sequences of 24 steps/free choices, totaling sequences of 72 separate responses. The question is then, how random are their responses? Logically, it is impossible to prove that a set of numbers are truly random, it is only possible to show that statistically speaking, they do not appear to be random [19]. Much of the research on detection of randomness stems from attempts to create random number generators for commercial computing purposes, for use in industries such as gaming and online gambling. However, the largest application is cryptography. These industries require computations acting as random number generators, producing information that is highly unpredictable, unbiased, and trustworthy. As computer programs in practice use pre-established, finite lists of numbers to produce seemingly random actions, their output over time will become predictable. Thus, research on testing the efficacy of computerized random number generators has developed methods to detect non-randomness (not randomness per se).

One approach widely used in cryptography and in many other fields is the chi-squared method [2]. This creates a p value for the analyzed string to indicate the probability that it comes from a set of numbers that are unpatterned (i.e., randomly assorted). Over multiple strings analyzed the resulting p values can then be used as a data set. Although behavioral scientists are more familiar with using p values to evaluate hypotheses for individual research studies, the p values generated from analyzing data sets can themselves comprise a data set, and can be analyzed with normal inferential statistics. The most notable example of this in recent years is the Open Science Collaboration which revealed that most psychology research studies fail to be replicated [28]. Part of their analysis involved examining the p values reported in published psychology articles, and using null-hypothesis testing to decide the extent to which the p values from replications of the same protocols showed the same score distributions.

Basically, p values are random variables, technically transformed test-statistics (in our case from chi-squared tests) which puts them into a standard form, allowing interpretation independent of the particular statistical test used [26]. They have the benefit of being potentially normally distributed when they are derived from hypothesis tests in which the null-hypothesis is incorrect. For this reason, the transformed scores (i.e., the p values) will usually produce distributions more amenable to further parametric statistical analysis, compared to the raw test statistics (i.e., the chi-squared test values). This is because, for example, the chi-squared statistic is calculated from the sum of squares of both positive and negative values, the accumulation of relatively higher values will produce positively skewed distributions.

Each chi-squared calculation is done at the level of the individual participant (not group data). For each analysis performed, at the individual level a high p value will indicate that the responses made by the participant appear to be more random. Therefore, relatively higher p values can be interpreted as showing relatively better top-down, executive control of behavior. The potential range of p values is between 0 and 1.

3 An Example of the Analysis Method Using Task Performance Data

From the data acquired from the 30 participants tested, our goal is to test whether each participant appears to have chosen the colors randomly. If a participant chooses colors in a purely random fashion, given 24 choices in a single trial, the expected counts of a single color is 6 ($=\frac{24}{4}$) in a single trial. Because each participant had three trials, the expected counts of a single color is 18 in three trials combined. Similarly, we can calculate the expected counts of two-color bigrams (e.g., how often yellow-blue occur) are chosen in a single trial and in three trials combined. Particularly, given 4 colors, there are 12 possible permutations of 2 different colors. Then, the 24 choices in a single trial produce 23 permutations of 2 different colors. Thus, the expected counts of 2 different color bigrams is ($=\frac{23}{12}$) (roughly 1.92) in a single trial, and that is ($=\frac{23}{4}$) (which is 5.75) in three trials combined.

Based on these calculations, we can apply the chi-squared goodness of fit test [8] to check if the observed counts follow the same distribution as the expected counts. We conduct this test five times to the data of each participant. The first three chi-squared tests each check the randomness of single-color choices in each of the three trials. The fourth chi-squared analysis checks the randomness of a single color over three trials combined (i.e. 72 choices). The fifth and final chi-squared checks the randomness of permutations of different color bigrams over all three trials. For all of these analyses the expected counts were greater than 5 (which is a precondition of analysis with chi-squared). The FREQ procedure in SAS 9.4 [43] was used to implement the chi-squared tests, and a SAS macro was written to automate the analysis procedure across the 30 participants.

Across the group of 30 participants who provided the test data, the mean p values for the randomness of individual choices (i.e., did they select all colors equally often) were all around 0.8, these p values were derived from single trials, as well as the calculation using performance across all three trials. As higher p values indicate greater appearance of randomness of responses, it appears that participants were quite good at this, and the measure may therefore not sufficiently challenge executive processes in this type of participant.

More challenging appears to be randomness of responding when measured by bigram frequencies (totals over three trials). The mean p value for this measure was 0.61 (range of scores = 0.02–0.99). The lower p value indicates that in general, the scores appeared to be less like a random set. For the proposed methods

of response-by-response analysis with chi-squared to be useful as a method of quantifying executive control, certain qualities of the data distributions are desirable. One is that there should be a normal distribution of scores. This appears to be the case of the p values of bigram frequencies, as shown by a Shapiro-Wilk test: $W = 0.95$, $df = 30$, $p = 0.20$. In addition, skew values ($z = 1.12$) and kurtosis values ($z = 0.62$) were both within limits for assumption of normal distribution of data [21]. In contrast, equivalent analyses for the frequencies of single-color choices indicated that all had statistically non-normal distributions. Consequently, the p values of bigram frequencies appear to be more appropriate measures of the ability of participants to deliberately avoid patterned responding (i.e., give responses that appear random). For this reason, only the bigram frequencies were further analyzed.

The score distribution also has to be sufficiently broad that it can distinguish different levels of performance, that is, it contains sufficient variance. The coefficient of variance was found to be 0.45. This is somewhat higher than the coefficient of variance for the total task completion time (the conventionally used measure of performance on trail-making tasks), which was 0.21. As both the distributions for total time and the p values of bigram frequencies were normally distributed, we examined the Pearson zero-order correlation between the two measures of performance. That revealed a significant negative correlation, $r = -0.49$, $p = 0.006$. This suggests that, across participants, relatively poor task performance as measured by completion time is associated with relatively poor performance as measured by bigram frequencies.

These preliminary analyses therefore suggest that randomness of single choices is not a good way to measure top-down executive control in this novel task. However, bigram frequencies, represented as p values of how much the responses appear random, may function better as a summary measure of executive control. The potential implications and applications of this are described in the final section.

4 Discussion and Conclusions

4.1 Summary

In this preliminary report, we provided first details of a novel cognitive tool, that is nevertheless similar in many respects to other paper-and-pencil 'trails tasks' used widely in behavioral sciences to measure executive cognitive control [10,23,39]. The principal difference being that this new task requires participants to make free choices, rather than to perform the task in a predefined way, which is the format of previous trails tasks. Moreover, we provide a method to analyze how well participants who perform the task can resist tendencies to pattern their responses.

The concept of free choice here is that the participants can, at each of 24 steps within each trial, choose between any of three colors without violating any rules. Admittedly, they are told to not choose the same color twice (hence limiting them from 4 to 3 options at each step), and they are instructed not to use any plans

or strategies. So, they are constrained at the overall task level within a trial, but not at the individual choice level at each step. Although not told to respond randomly, to attempt to do so is the only remaining approach they have to guide their choices. This is why we consider them free choices. Nevertheless, following discussions from the workshop, additional studies are being run excluding Rule 4 (Try to choose all the colors equally often) as it may contradict the goal of randomness of single choices.

In addition, we provide a statistical method to describe how well individual participants were at avoiding patterned responding and effectively responding randomly. This approach uses p values derived from chi-squared analyses, calculated at the level of the individual.

However, the wider context is that we show how data collection methods in behavioral sciences can be approached differently, to allow measures that more closely align to the concept of executive cognitive control. We have previously argued that definitions of executive control, which emphasize processing in nonroutine or ambiguous situations to produce appropriate responses are best considered as divergent thinking [34]. Divergent and convergent thinking are concepts in the classification of cognitive processes that have been popular since the 1950's [16]. Despite this, the vast majority of tests used in experimental and clinical practice attempt to measure executive processes that are substantially convergent in their structure and analysis methods. Because divergent processes, as we have attempted to elicit in our Choice Trails Test via free choices, do not have unique right answers, novel ways of deriving a performance measure have to be explored, necessitating meditations on what exactly is meant by top-down executive cognitive control.

Consequently, our suggested method uses a procedure adopted from computer science, one that is frequently employed to test the abilities of random number generators [2,19]. Although this approach is a relatively novel application within cognitive sciences, similar approaches have been used to measure behavior in clinical neuroscience. For example, a statistical measure of randomness of responses was used to examine stereotypical responding in patients with schizophrenia when asked to guess the color of playing cards presented sequentially in a random order [13]. Similarly, patients with Alzheimer's disease have been shown to overproduce ascending counting patterns (e.g., 3-4-5) when asked to imagine repeatedly throwing a normal six-sided die and orally reporting the outcomes [6]. These and other similar divergent thinking studies of neuropsychiatric patients have mainly used the Random Number Generation Index of Evans (1978) [12]. However, that calculation appears to be very similar to chi-squared anyway. The benefits of using chi-squared-derived p values are that they are more easily computed in standard statistical software packages, and are well understood from their use in null-hypothesis testing.

4.2 Implications and Applications

The task described here may be useful as an alternative way to measure the ability of people to make free choices, in a nevertheless constrained task that

allows for the individual to make choices that appear random, or which follow predictable, routine patterns. Much evidence from cognitive and brain sciences suggests that the human neurocognitive system tends to revert to routine patterns, as the alternative, top-down executive control, is resource demanding and subjectively effortful [34]. Moreover, the proposed methodology, of requesting that study participants avoid routine response biases and then estimating the randomness of their free choices, can potentially be applied to many other existing cognitive laboratory and clinical assessment methods.

In the specific task presented here, we found that that the analysis method produced results that overlapped with the traditional methods of measurement in similar tasks (i.e., time taken to task completion). Both measures were correlated, suggesting that both are measuring some aspect of executive control. One observation made was that our approach involving a response-by-response analysis produced greater between-individual variance in performance scores than the traditional overall time-based method. This may have some practical application. There has recently been concern within cognitive sciences involving behavioral studies that task measures are often statistically unreliable, producing many Type II statistical errors when used for hypothesis testing [9,17]. This is because cognitive tasks have generally been developed for laboratory-based experimental studies to elicit effects which are more apparent and easier to detect when between-individual variance is low. However, that reduces their reliability and makes them poor measures of how people differ in their abilities. That reliable variability is often needed when, for example, making brain-behavior associations by linking cognitive test scores to functional or structural neuroanatomy, genetic and biomarkers etc. The methods proposed here may therefore function better as indices of individual differences in executive processing than they will in tests of experimental manipulations on processing.

However, this need not be a limitation. We argue that the method of analysis described here, which focuses on a more microanalysis of response-by-response data, can be performed in tandem with traditional analyses which focus on overall task performance. This can be done whether the study paradigm is experimental or individual-differences based. This is a wise approach anyway in that the current methods which focus on overall task performance can obscure real differences in cognitive processes that underlie performance. It is known that multiple different processes can produce the same behavior. This is known as functional equivalence in traditional cognitive science [46] and degeneracy in clinical and cognitive neurosciences [31]. Multiple analyses of task performance can help to delineate those different underlying processes.

In fact, one of us has previously argued that there is a need for clinical assessments of cognitive abilities to learn from traditional cognitive sciences [33]. Paper-and-pencil based cognitive test methods, such as described in this paper, are widespread in clinical cognitive assessment, due to their simplicity and portability. That allows them to be used in bedside testing. This contrasts with often highly-technical methods used in experimental cognitive psychology that are difficult to transpose from the laboratory setting. However, even bedside-derived

cognitive data can benefit from the process-based analyses used in cognitive sciences. Traditionally, cognitive science analyses on behavioral data have used methods to produce data with 'temporal density' that can be used to track processing over short-time periods [47]. Although this is now common in laboratory-based cognitive studies (e.g., eye tracking), clinical testing tends to rely on overall performance measures. In this paper we show how dense data can still be elicited using the traditional paper-and-pencil tests typical of clinical cognitive assessments.

4.3 Conclusions

Executive functions, by definition, deal with ambiguous stimulus-response associations and require that willed choices be made [14,20,25,27,42]. This conceptually aligns closely with the idea of divergent thinking—a broad definition that invokes cognitive processes that are creative and result in free choice of responses [16]. However, there has long been a disjunction between conceptualization of executive functions, and methods of measurement used in behavioral sciences. Here we show that common testing methods, such as the paper-and-pencil trails tests [10,23,39] can be altered to change them from evoking convergent, schematic action selections, to evoking divergent, free choices. This necessarily requires a different approach to how performance is quantified. We suggest a method using p values. We argue that this approach allows for new ways to operationalize and measure top-down cognitive control in human behavior. And these new ways may allow fresh insights into these high-level cognitive processes. Future research will ultimately support, or challenge, the utility of this approach.

Acknowledgments. This research was supported by a grant from Research Affairs at the Faculty of Psychology, Chulalongkorn University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alexander, M.P., Stuss, D.T., Picton, T., Shallice, T., Gillingham, S.: Regional frontal injuries cause distinct impairments in cognitive control. Neurology **68**(18), 1515–23 (2007). https://doi.org/10.1212/01.wnl.0000261482.99569.fb
2. Almaraz Luengo, E., Alaña Olivares, B., García Villalba, L.J., Hernandez-Castro, J., Hurley-Smith, D.: Stringent test suite: Ent battery revisited for efficient p value computation. J. Cryptogr. Eng. **13**(2), 235–249 (2023). https://doi.org/10.1007/s13389-023-00313-5
3. Baddeley, A.: Is working memory working? The fifteenth Bartlett lecture. Q. J. Exp. Psychol. A Hum. Exp. Psychol. **44**(1), 1–31 (1992). https://doi.org/10.1080/14640749208401281

4. Baggetta, P., Alexander, P.A.: Conceptualization and operationalization of executive function. Mind Brain Educ. **10**(1), 10–33 (2016). https://doi.org/10.1111/mbe.12100
5. Beatty, W.W., Testa, J.A., English, S., Winn, P.: Influences of clustering and switching on the verbal fluency performance of patients with Alzheimer's disease. Neuropsychol. Dev. Cogn. B Aging Neuropsychol. Cogn. **4**(4), 273–279 (1997). https://doi.org/10.1080/13825589708256652
6. Brugger, P., Monsch, A.U., Salmon, D.P., Butters, N.: Random number generation in dementia of the Alzheimer type: a test of frontal executive functions. Neuropsychologia **34**(2), 97–103 (1996). https://doi.org/10.1016/0028-3932(95)00066-6
7. Burgess, P.W., Shallice, T.: Response suppression, initiation and strategy use following frontal lobe lesions. Neuropsychologia **34**(4), 263–72 (1996). https://doi.org/10.1016/0028-3932(95)00104-2
8. Cochran, W.G.: The x2 test of goodness of fit. Ann. Math. Stat. **23**, 315–345 (1952). https://doi.org/10.1214/aoms/1177729380
9. Dang, J., King, K.M., Inzlicht, M.: Why are self-report and behavioral measures weakly correlated? Trends Cogn. Sci. **24**(4), 267–269 (2020). https://doi.org/10.1016/j.tics.2020.01.007
10. Delis, D., Kaplan, E., Kramer, J.: Delis-Kaplan Executive Function System Technical Manual. The Psychological Corporation, San Antonio, TX (2001)
11. Diaz, M., Sailor, K., Cheung, D., Kuslansky, G.: Category size effects in semantic and letter fluency in Alzheimer's patients. Brain Lang. **89**(1), 108–14 (2004). https://doi.org/10.1016/S0093-934X(03)00307-9
12. Evans, F.J.: Monitoring attention deployment by random number generation: an index to measure subjective randomness. Bull. Psychon. Soc. **12**(1), 35–38 (1978). https://doi.org/10.3758/BF03329617
13. Frith, C.D., Done, D.J.: Stereotyped responding by schizophrenic patients on a two-choice guessing task. Psychol. Med. **13**(4), 779–86 (1983). https://doi.org/10.1017/s0033291700051485
14. Frith, C.D., Friston, K., Liddle, P.F., Frackowiak, R.S.: Willed action and the prefrontal cortex in man: a study with pet. Proc. R. Soc. Lond. B Biol. Sci. **244**(1311), 241–246 (1991). https://doi.org/10.1098/rspb.1991.0077
15. Goel, V., Grafman, J.: Are the frontal lobes implicated in "planning" functions? Interpreting data from the tower of Hanoi. Neuropsychologia **33**(5), 623–42 (1995). https://doi.org/10.1016/0028-3932(95)90866-p
16. Guilford, J.: The structure of intellect. Psychol. Bull. **53**(4), 267–293 (1956). https://doi.org/10.1037/h0040755
17. Hedge, C., Powell, G., Sumner, P.: The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods **50**(3), 1166–1186 (2017). https://doi.org/10.3758/s13428-017-0935-1
18. Hughes, C., Ensor, R.: Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and self-perceived academic success at 6 years of age. J. Exp. Child Psychol. **108**(3), 663–76 (2011). https://doi.org/10.1016/j.jecp.2010.06.005
19. Hurley-Smith, D., Patsakis, C., Hernandez-Castro, J.: On the unbearable lightness of fips 140–2 randomness tests. IEEE Trans. Inf. Forensics Secur. **17**, 3946–3958 (2020). https://doi.org/10.1109/TIFS.2020.2988505
20. Jahanshahi, M.: Willed action and its impairments. Cogn. Neuropsychol. **15**(6–8), 483–533 (1998). https://doi.org/10.1080/026432998381005

21. Kim, H.Y.: Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restor. Dent. Endod. **38**(1), 52–4 (2013). https://doi.org/10.5395/rde.2013.38.1.52
22. Knoblock, C.A.: Abstracting the tower of Hanoi. In: Working Notes of AAAI-90 Workshop on Automatic Generation of Approximations and Abstractions, Boston, MA, no. 4976, pp. 1–11 (1990)
23. Maj, M., et al.: Evaluation of two new neuropsychological tests designed to minimize cultural bias in the assessment of HIV-1 seropositive persons: a who study. Arch. Clin. Neuropsychol. **8**(2), 123–35 (1993). https://doi.org/10.1016/0887-6177(93)90030-5
24. McMillen, P., Levin, M.: Collective intelligence: a unifying concept for integrating biology across scales and substrates. Commun. Biol. **7**(1), 378 (2024). https://doi.org/10.1038/s42003-024-06037-4
25. Miller, E., Wallis, J.: Executive function and higher-order cognition: definition and neural substrates. In: Squire, L.R. (ed.) Encyclopedia of Neuroscience, vol. 4. Academic Press, Oxford (2009)
26. Murdoch, D.J., Tsai, Y.L., Adcock, J.: P-values are random variables. Amer. Statist. **62**(3), 242–245 (2008). https://doi.org/10.1198/000313008X332421
27. Norman, D.A., Shallice, T.: Attention to action: willed and automatic control of behavior. In: Davidson, R.J., Schwartz, G.E., Shapiro, D. (eds.) Consciousness and Self-regulation: Advances in Research and Theory, vol. 4, pp. 1–18. Plenum, New York (1986)
28. Open Science Collaboration: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015). https://doi.org/10.1126/science.aac4716
29. Panikratova, Y.R., Vlasova, R.M., Akhutina, T.V., Korneev, A.A., Sinitsyn, V.E., Pechenkova, E.V.: Functional connectivity of the dorsolateral prefrontal cortex contributes to different components of executive functions. Int. J. Psychophysiol. **151**, 70–79 (2020). https://doi.org/10.1016/j.ijpsycho.2020.02.013
30. Phillips, L.H., Wynn, V., Gilhooly, K.J., Della Sala, S., Logie, R.H.: The role of memory in the tower of London task. Memory **7**(2), 209–31 (1999). https://doi.org/10.1080/741944066
31. Pluck, G.: The misguided veneration of averageness in clinical neuroscience: a call to value diversity over typicality. Brain Sci. **13**(6), 860 (2023). https://doi.org/10.3390/brainsci13060860
32. Pluck, G., Amraoui, D., Fornell-Villalobos, I.: Brief communication: Reliability of the D-KEFS tower test in samples of children and adolescents in Ecuador. Appl. Neuropsychol. Child **10**(2), 158–164 (2021). https://doi.org/10.1080/21622965.2019.1629922
33. Pluck, G., Ariyabuddhiphongs, K.: Clinical cognitive sciences. In: Aldini, A. (ed.) Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops. SEFM 2023. Lecture Notes in Computer Science, vol. 14568. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-66021-4_9
34. Pluck, G., Cerone, A., Villagomez-Pacheco, D.: Executive function and intelligent goal-directed behavior: perspectives from psychology, neurology, and computer science. In: Masci, P., Bernardeschi, C., Graziani, P., Koddenbrock, M., Palmieri, M. (eds.) Software Engineering and Formal Methods. SEFM 2022 Collocated Workshops. SEFM 2022. Lecture Notes in Computer Science, vol. 13765, pp. 324–350. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-26236-4_27
35. Pluck, G., Crespo-Andrade, C., Parreño, P., Haro, K.I., Martínez, M.A., Pontón, S.C.: Executive functions and intelligent goal-directed behavior: a neuropsychologi-

cal approach to understanding success using professional sales as a real-life measure. Psychol. Neurosci. **13**(2), 158–175 (2020). https://doi.org/10.1037/pne0000195
36. Pluck, G., Villagomez-Pacheco, D., Karolys, M.I., Montano-Cordova, M.E., Almeida-Meza, P.: Response suppression, strategy application, and working memory in the prediction of academic performance and classroom misbehavior: a neuropsychological approach. Trends Neurosci. Educ. **17**, 100121 (2019). https://doi.org/10.1016/j.tine.2019.100121
37. Poldrack, R.A., et al.: The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Front. Neuroinform. **5**, 17 (2011). https://doi.org/10.3389/fninf.2011.00017
38. Pribram, K.: The primate frontal cortex– executive of the brain. In: Pribram, K., Luria, A. (eds.) Psychophysiology of the Frontal Lobes, pp. 293–314. Academic Press, New York (1973). https://doi.org/10.1016/B978-0-12-564340-5.50019-6
39. Reitan, R.M.: The relation of the trail making test to organic brain damage. J. Consult. Clin. Psychol. **19**(5), 393–394 (1955). https://doi.org/10.1037/h0044509
40. Robinson, G., Shallice, T., Bozzali, M., Cipolotti, L.: The differing roles of the frontal cortex in fluency tests. Brain **135**(7), 2202–14 (2012). https://doi.org/10.1093/brain/aws142
41. Roebers, C.M., Rothlisberger, M., Cimeli, P., Michel, E., Neuenschwander, R.: School enrollment and executive functioning: a longitudinal perspective on developmental changes, the influence of learning context, and the prediction of preacademic skills. Eur. J. Dev. Psychol. **8**(5), 526–540 (2011). https://doi.org/10.1080/17405629.2011.571841
42. Rolls, E.T.: Willed action, free will, and the stochastic neurodynamics of decision-making. Front. Integr. Neurosci. **6**, 68 (2012). https://doi.org/10.3389/fnint.2012.00068
43. SAS Institute Inc: SAS/STAT® 15.3 user's guide. SAS Institute Inc., Cary, NC (2023)
44. Shallice, T.: Specific impairments of planning. Philos. Trans. R. Soc. Lond. B Biol. Sci. **298**(1089), 199–209 (1982). https://doi.org/10.1098/rstb.1982.0082
45. Shallice, T., Burgess, P.: The domain of supervisory processes and temporal organization of behaviour. Philos. Trans. R. Soc. Lond. B Biol. Sci. **351**(1346), 1405–1411 (1996). https://doi.org/10.1098/rstb.1996.0124
46. Simon, H.: The functional equivalence of problem solving skills. Cogn. Psychol. **7**(2), 268–288 (1975). https://doi.org/10.1016/0010-0285(75)90012-2
47. Simon, H.A.: Information processing models of cognition. Annu. Rev. Psychol. **30**(1), 363–396 (1979). https://doi.org/10.1146/annurev.ps.30.020179.002051
48. Winter, W.E., Broman, M., Rose, A.L., Reber, A.S.: The assessment of cognitive procedural learning in amnesia: why the tower of Hanoi has fallen down. Brain Cogn. **45**(1), 79–96 (2001). https://doi.org/10.1006/brcg.2000.1257
49. Youyou, W., Yang, Y., Uzzi, B.: A discipline-wide investigation of the replicability of psychology papers over the past two decades. Proc. Natl. Acad. Sci. U.S.A. **120**(6), e2208863120 (2023). https://doi.org/10.1073/pnas.2208863120
50. Zhang, C., Lipovetzky, N., Kemp, C.: Comparing AI planning algorithms with humans on the tower of London task. In: Goldwater, M., Anggoro, F.K., Hayes, B.K., Ong, D.C. (eds.) Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 45 (2023). https://escholarship.org/uc/item/5164p0rz