Aus dem Institut für Systemische Neurowissenschaften
der Heinrich-Heine-Universität Düsseldorf

Computer Assisted Speech Analysis in Testing Executive Functions

Dissertation

zur Erlangung des Grades eines „Doctor rerum medicarum“
(Dr. rer. med.)
der Medizinischen Fakultät der Heinrich-Heine-Universität
Düsseldorf

vorgelegt von

Julia Amunts

(2021)

Als Inauguraldissertation gedruckt mit der Genehmigung der

Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:

Dekanin/Dekan:

Guchtachter/innen: Prof. Dr. Simon Eickhoff, PD Dr. Katja Biermann-Ruben

Teile dieser Arbeit wurden veröffentlicht:

Amunts, J., Camilleri, J.A., Eickhoff, S.B., Heim, S., Weis, S., (2020), Executive functions predict verbal fluency scores in healthy participants. *Scientific Reports*, (10), 11141[1]

Amunts, J., Camilleri, J.A., Eickhoff, S.B., Patil, K.R., Heim, S., von Polier, G., Weis, S. (2021), Comprehensive verbal fluency features predict executive function performance. *Scientific Reports*, (11**)**, 6929[2]

Camilleri, J.A., Eickhoff, S.B., Weis, S., Chen, J., Amunts, J., Sotiras, A., Genon, S. (2021), A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan Executive Function System. *Scientific Reports*, (11), 16896[3]

Zusammenfassung

Sprachproduktion ist ein komplexes Konstrukt und beinhaltet zahlreiche kognitive Prozesse. Exekutivfunktionen (EFs) umfassen kognitive Prozesse, die zielgerichtetes Verhalten und die Produktion von Sprache ermöglichen. Beim Wortabruf spielen verschiedene EF eine Rolle, die z.B. für die Wortauswahl, das Erinnern von bereits produzierten Wörtern oder die Inhibition von falsch aktivierten Wörtern zuständig sind. Die Wortflüssigkeits- (VF) Aufgabe ist ein Diagnostikinstrument, das in vielen neuropsychologischen Testbatterien enthalten ist und als Test für EF dient. Dennoch ist die genaue Art der Involvierung der einzelnen EF in der VF Aufgabe bisher nicht vollständig geklärt. Gründe hierfür sind vor allem kleine Stichproben und eine geringe Anzahl an EF-Tests, die dazu führen, dass Studienergebnisse nur schwer generalisiert werden können.

Das Hauptziel der Studie war es, die konkrete Beteiligung der einzelnen EF in der VF zu untersuchen. Als erstes wurden gemeinsame Strukturen der EF-Tests und der VF untersucht (Studie 1). Im nächsten Schritt wurden, unter Berücksichtigung individueller Unterschiede, die konkreten Zusammenhänge der EF und der VF untersucht (Studie 2). Zuletzt wurde das prädiktive Potential der VF Aufgabe untersucht, um zu prüfen, ob die VF zukünftig als mögliches diagnostisches Screening für EF eingesetzt werden könnte (Studie 3). In allen Studien wurden Methoden des *Maschinellen Lernens* angewendet, um eine Generalisierung der Ergebnisse zu ermöglichen.

Studie 1 zeigte eine Aufteilung der EF in zwei Faktoren auf, die einfache und komplexere EF-Tests beinhalteten. Der einfachere Faktor beinhaltete VF Aufgaben und EF-Tests, die den Bereichen der kognitiven Flexibilität, Aufmerksamkeit und Inhibition zuzuordnen sind. Der zweite Faktor beschrieb EF-Tests des abstrakten sowie des logischen Denkens. Diese Ergebnisse zeigen die Einbeziehung von einfacheren EF während der VF. Studie 2 unterstützte diese Befunde und identifizierte allgemeine Reaktionsgeschwindigkeiten als zentralen Aspekt der VF. Schließlich wurde die Beziehung von EF und VF *vice versa* untersucht (Studie 3) und einzelne EF-Test-Score mittels umfangreicher VF Informationen vorhergesagt. Während in Studie 2 nicht alle Bereiche der EF den VF-Leistungen zugeordnet werden konnte, konnte Studie 3 alle Unterbereiche der EF vorhersagen.

Insgesamt erbrachten diese drei Studien neue Erkenntnisse über die Involvierung der EF in der VF-Aufgabe und untersuchten darüber hinaus das Potential dieser Sprachproduktionsaufgabe, differenzierte Einblicke in EF zu erhalten. Diese Ergebnisse stellen eine Grundlage zur Entwicklung sprachbasierter Screenings im klinischen Kontext dar.

Summary

The conceptualization and production of speech are complex processes that require multiple cognitive processes. Executive functions (EFs) encompass a set of cognitive processes allowing for goal-directed behaviour and facilitating speech production. Different subdomains of EFs are well known to play a crucial role in word retrieval e.g., for the selection of correct words, remembering already produced words or for the suppression of incorrectly activated words. An instrument to test EFs, is the verbal fluency (VF) task, a well-established speech test that is part of numerous diagnostic batteries and known to reflect EF performance. However, the concrete involvement of the different subdomains of EFs and VF performance remains controversial. Reasons for this are small sample sizes and a limited number of EF tests resulting in a lack of generalizability of the study results.

The main goal of this thesis was to investigate the involvement of EFs in VF performance. In a first step, the common structure of different EF tests (including the VF task) was investigated (study 1). In a second step, the concrete relationship of EFs and VF performance was investigated taking into account individual differences (study 2). In a last step, the predictive power of the VF task was studied to elaborate the potential of the VF task serving as a diagnostic screening for testing EFs (study 3). To examine the generalizability of results, machine learning methods were applied.

Results of study 1 revealed two factors differentiating between more simple and complex EF tests. Here, the simple factor included all VF tasks as well as tests tapping into cognitive flexibility, attention, and inhibition. The second factor contained tests referring to abstract thinking and reasoning. These study results highlighted the involvement of more simple aspects of EFs and VF performance. Results of study 2 supported findings of study 1 i.e., the close relationship of more simple EF domains including general processing speed and reaction times and VF. In a last step, study 3 investigated the relationship of the VF task and EF tests *vice versa* and applied a comprehensive set of VF information to predict EF test performance. While study 2 could not reveal relationships of all EF domains and VF performance, study 3 predicted EF test performance tapping into all subdomains of EFs.

In sum, the present three studies provided deeper insights into the involvement of EFs in the VF task and further elucidate the potential of a speech production task to gain deeper insights into EF performance. These results provide a basis for the development of language-based screening in a clinical context.

Abbreviations

ADHD	Attention deficit hyperactivity disorder
CWI	Colour Word Interference
D-KEFS	Delis-Kaplan Executive Function System
EFs	Executive functions
eNKI	Enhanced Nathan Kline Institute
ERT	Estrogen replacement therapy
ML	Machine Learning
MUC	Memory, Unification, Control
TMT	Trail-Making Test
TOL	Tower of London
VF	Verbal fluency
WCST	Wisconsin Card Sorting Test

Table of Contents

1 Introduction

Approximately eight billion people use 7.000 different languages worldwide[4] comprising different vocabularies, sounds, syntax and grammar with the common aim to communicate thoughts and needs. *Language* proficiency involves language specific abilities, such as lexical access, and language unspecific cognitive components, such as long-term memory and planning abilities[5]. The actual articulation of words, the sound and prosody of the voice is named as *speech*. Speech is induced by laryngeal structures and motoric abilities. However, articulation processes also include cognitive components to build an articulatory plan or monitor overt speech[6].

Although there are controversial discussions about the dependencies language and cognition *perse*[7], there is consensus that general as well as language specific cognitive abilities are needed for successful speech production[6–8]. A central component of speech production processes are EFs containing a set of cognitive functions to control behaviour. They play a major role in planning the general content of speech, accessing correct words[6], controlling and remembering produced speech and facilitate the ability to inhibit or correct words[9].

1.1 The relevance of executive functions

The concept of EFs encompasses higher and lower level cognitive functions that refer to a set of top-down mental processes which control and modulate goal-directed behaviour[10]. While higher level processes include abilities such as reasoning, planning and problem-solving, they influence lower-level processes such as cognitive flexibility, inhibition and working memory[11]. Previous work has highlighted the overall importance of EFs for mental[12] and physical health[13], school success[14], cognitive and social development[15]. In clinical context, EFs play an important role in a high number of psychiatric and neurologic diseases and influence the patient's behaviour e.g., in dementia[16] or attention deficit hyperactivity disorder (ADHD)[17].

1.1.1 The measurement and conceptualization of executive functions

In research and clinical contexts, complex test batteries are used to assess EF performance. These test batteries contain a high number of different EF tests to cover the wide spectrum of the different EF subdomains. The *Delis-Kaplan Executive Function System* (D-KEFS)[18] is an example for a normalized test battery consisting of nine verbal and non-verbal EF tests in English. Since these test batteries are standardized and normalized for a specific population, test batteries in other languages, like the *Wiener Testsystem*[19] used for German speakers, were created.

Commonly used instruments for assessing problem-solving and reasoning abilities are the *Wisconsin Card Sorting Test*[20] (WCST) and *Trail-Making Test*[21] (TMT). In the WCST the participant is asked to identify the right card that fits into the pattern of four other cards. Here, the challenge is to recognize the change of different patterns and adapt the rules. The TMT consists of two parts: In the first part the participant is asked to click on random-sorted numbers in a sequential order, and the second more complex part requires clicking on letters and numbers in an alternately order. However, studies found heterogenous task performances in patients suffering from similar neurological impairments, indicating that similar constructed EF tests do not automatically assess exactly identical EFs due to the subtle involvement of additional EFs[22].

EF batteries also include tests tapping into working memory performance. Examples are *span tasks* which are mainly related to numbers or letters. However, these tasks are known to also require attention components and additional executive processes[23]. Therefore, the *n-back task* was developed to capture specialized working memory performance[24]. This task is used to recognize whether the presented item matches the item *n* turns back.

The *go/no-go tasks*[25] and *stop-signal tasks*[26] are often used in clinical practice and research to measure response inhibition. Additionally, the *Stroop task* is a commonly used diagnostic instrument which is based on naming and reading abilities, and consists of congruent and incongruent conditions which are related to colours and written words[27].

Measuring EFs in clinical context is accompanied by multiple clinically and conceptually challenges. Firstly, EF test batteries are extremely time consuming. Assessing the full variety of EF tests occupies much time and often patients are not able to perform such time and energy consuming tasks due to the symptoms of the current disease[28]. Thus, a detailed insight into the patient's EF performance is rarely possible. Here, disease-specific diagnostic screenings, such as the *Mini-Mental-Status-Test* [29], provide a solution gaining first insights into the patient's EF performance within a short period of time. Secondly, commonly used tests in clinical routine are usually *pen-and-paper tests*. Even though it is well known that digitalized tests provide

higher objectivity, efficiency and accuracy, *pen-and-paper* versions such as the TMT[30] are applied in clinical context with the aim to simultaneously assess the motor skills of the patient and cognition.

From a conceptual perspective, there are mainly two challenging reasons interpreting EFs: Firstly, it is difficult to capture pure EF performance since EFs are closely related to additional cognitive processes such as intelligence[31]. Secondly, due to the specific instructions and the laboratory task design, researchers criticize the missing link to real world situations and life-like activities (ecological validity)[32].

Besides the challenges in testing EFs in clinical context, the general concept of EFs is discussed controversially. Here, the contribution of the different aspects of EFs to the overall concept and the intertwining of the different EF domains contribute to the complexity of EFs[33]. Additionally, EFs do not represent single processes but rather include a complex macro-construct of various cognitive functions[34] and there is still a lack of a formal definition of EFs. Consequently, there is constant interest in the investigation of common structures of EFs and its relationship to other cognitive traits in healthy controls as well as in patients. The concept of EFs has been investigated in multiple studies by applying different statistical approaches, such as factorization approaches, to better capture the unity and diversity of EFs[22]. Studies described different models and subdivisions of EFs based on the analysis of EF test results. While some groups suggested two- and three-factor models, others proposed more elaborated options including up to eight different EF sub-domains[18]. For the sake of simplicity and a consistent naming of EFs, many studies rely on a three-factor model as suggested by *Diamond*[11] including the subdomains of cognitive flexibility, working memory and inhibition.

Cognitive flexibility enables people to adapt and change perspectives on spatial and interpersonal levels[11]. Moreover, it involves the ability of pursuing complex tasks and facilitate creative thinking. Successful cognitive flexibility performance is built on another EF domain, namely working memory. *Working memory* allows people to hold information in mind temporarily and manipulate it with the aim to achieve short-term goals[35]. Literature distinguishes between non-verbal and verbal working memory highlighting the importance of working memory for speech production. The third EF subdomain is *inhibition*, also known as *inhibitory control*. This term embraces the ability to inhibit or control prepotent, automatic and dominant actions, thoughts and feelings when they are not appropriate for the current context[36]. It also includes the ability to suppress interfering information[37]. As mentioned previously, inhibitory performance does not act in isolation but is rather intertwined with attention, cognitive flexibility and working memory capacities.

To summarize, testing EFs is quite complex and requires a broad range of verbal and non-verbal tests examining the different subdomains of EFs. However, the high time consumption as well as a lack of ecological validity and missing digitalized tools, are still a problem in clinical context.

1.1.2 The relationship of speech and executive functions in the clinical context

EF test batteries are used in the clinical context for diagnosing various neurological and psychiatric diseases. Depending on the specific disease and the underlying impaired neural processes, different symptom profiles occur. Although the different EF subdomains are highly related to each other, they are distinctly severely impaired in each disease and lead to various clinical presentations. These symptoms are observed in the patient's behaviour as well as in their speech. In detail, EFs play an essential role in speech production processes. A high number of speech production models describe e.g. single word production processes (*Levelt* model)[6], the influence of EFs on visual and verbal subsystems (*Baddeley's* working memory model)[9] and neurobiological models of language (MUC)[8]. However, the respective components of these speech production models vary greatly, and it is difficult to compare the exact relationship of speech parameters and EFs.

The close relationship of EFs and speech production is verifiable in many neurological and psychiatric diseases. Depending on the specific disease different domains of EF are impaired. For example, in Alzheimer's disease working memory is particularly impaired resulting in an increase of *misbinding errors*[38]. In these errors, the patient misremembers features of a specific object which leads to difficulties in caring for themselves and perform daily living activities[39]. The lack of working memory capacities also affects the semantic system of the patients. This involves sematic paraphasias and the use of improper words approximating those intended[40]. Another example of a disease with impaired EFs is schizophrenia. Studies found that cognitive flexibility and working memory processing speed are especially impaired in schizophrenic patients[41]. On the behavioural level, this leads to impaired self-monitoring and formal thought disorders. On a linguistic level these symptoms result in reduced syntactic complexity[42], neologisms, derailment and poverty of speech[43].

Due to the occurrence of speech abnormalities, and the close relationship of EFs and speech production, EF test batteries also contain verbal components. Some EF tests are provided in verbal and non-verbal versions to ensure that the patient's test performance does not mainly

rely on language specific components. The same EF test construct provides different stimuli e.g., within the *n-back* tasks, assessing working memory performance. Here, one option is based on non-verbal abstract figures, while the verbal option of the *n-back* task includes letters. While some tests are provided in verbal and non-verbal alternatives, some EF tests mainly rely on verbal components, as it can be seen in the *Stroop* test. This test assesses naming and reading interference to provide insights into inhibitory processes.

1.2 Diagnostic value of speech production parameters

Speech production is affected in many neurological and psychiatric diseases and provides various insights into cognitive functions[44,45]. Due to the strong involvement of EFs in speech production processes[6], the analysis of speech parameters provides a time-saving alternative to complex diagnostic EF batteries to gain first insights into EF performance, e.g. in severely impaired patients.

Different types of speech production parameters can be elicited within different types of speech tasks. While highly operationalized tasks, such as reading out a scripted text, are known to capture prosodic and motoric information[46], less structured tasks (e.g. *picture descriptions*) and *spontaneous speech* are needed to capture more complex levels of speech, i.e. syntactic or semantic information[47].

1.2.1 Verbal fluency and its relation to executive functions

The VF task is part of many diagnostic batteries and is a wide-spread tool assessing verbal ability and providing first insights into executive control in clinical and non-clinical groups. Due to the brevity and simplicity of this task it is for example suitable as a bedside test to gain first insights into general EF performance[48]. Moreover, it is implemented as an assessment tool for schizophrenia[49], stroke severity[50] and in a screening for mild Alzheimer´s disease[51]. Two different types of the VF task are used. In the lexical VF task, the participant is asked to name as many words as possible related to a specific initial letter (e.g. *L* or *R*). The semantic VF task requires the production of words belonging to a specific category (e.g. *sports*, *animals*). Here, depending on the type of the VF task, different restrictions are given to limit options, e.g. names, repetitions or words with the same stem. Usually, in clinical context, the VF task is assessed for 1 min and the number of correct words produced is evaluated. Within the last decades,

research groups investigated the validity of the VF task, tested in specific populations (e.g. healthy German participants[52]) and defined different levels of difficulty such as the difficulty level of a certain category or the frequency of specific initial letters. Hence, these levels are used to adapt the cognitive demand to clinical populations or healthy controls. To increase the involvement of working memory capacities, a switching component was created. Here, participants switch between two different categories or letters in an alternating order within the same task (e.g. *L-S-L*). Comparing the cognitive demands, required in the lexical VF and semantic VF tasks, researchers suggested an increased demand of strategic search in the lexical VF task, whereas the semantic VF tasks mainly relies on the organization of semantic knowledge[53,54].

Within a short timeframe of 1-2 minutes participants are asked to access their mental lexicon to produce words from a certain category or with an initial letter. They also need to concentrate on the given task, select correct words and focus on certain rules. Looking at the cognitive requirements more in detail, *Rosen* and *Engle*[55] described four steps needed for word retrieval in the VF task. Firstly, words that are related to the cue get activated. Secondly, already produced items are monitored to prevent errors like item repetitions. Thirdly, previously produced words are suppressed and lastly, new cues are generated to find new words. An essential role for successful VF performance plays working memory performance. It is needed to monitor the already produced output and to suppress already produced or incorrect items[55]. Also, *Hirshorn* and *Thompson-Schill* described the suppression of activated but not-matching or irrelevant responses[56]. The suppression of repetitive or irrelevant items demonstrates the intertwined performance of working memory abilities and inhibitory performance. Additionally, cognitive flexibility and planning abilities are involved in VF performance. Previous literature particularly emphasizes its involvement in generating new words[57]. This searching process of new items can be subdivided in clustering and switching processes[57]. While clustering describes the ability to search words within a particular subcategory or access phonemically related words, switching is needed to change the current category as soon as it is exhausted [58].

Although there is consensus that EFs are essential to perform the VF task, study results reporting the concrete involvement of the different domains of EFs and VF performance are ambiguous. While some studies identified a relationship between cognitive flexibility[59], working memory[60,61], inhibition[60] and VF performance, other studies failed to find a link between specific EF domains and VF performance[62–64]. Moreover, many studies investigating the relationship of VF and EFs, just use a few number of EF tests to assess domain specific or

general EF performance[59,65]. Last but not least, the sample size is rather small in many studies. Hence, despite the high number of studies investigating the relationship of EFs and VF performance, study results do not tend to be generalizable, and the concrete relationship is still an open question.

1.2.2 Challenges in interpreting verbal fluency performance

In general, there is consensus that the VF task is a valid tool to gain first insights into executive processes[60]. Several studies suggested the validity of the VF task as a tool to measure EF performance. For example, studies demonstrated poorer performance in the VF task of children with ADHD compared to healthy controls[66]. Also, patients suffering from brain damage in frontal areas, which are associated with EF performance, produced less words in the VF tasks than healthy controls[67].

However, the investigation of the involvement of the concrete subdomains of EFs, i.e., cognitive flexibility, working memory and inhibition in clinical populations and healthy controls shows differences among the studies. For example, some studies reported a link between fewer perseveration errors and a higher number of correct words produced within the VF task and better working memory performance[68]. In contrast, other studies failed to confirm these findings[62]. Similarly controversial results were reported for the concrete involvement of inhibitory processes in VF performance[60,69].

Such divergent study findings might be related to the influence of inter- and intra-individual differences of participants. Inter-individual differences are e.g., described in the context of fluid intelligence. Studies found that the domains *planning* and *reasoning* are influenced by the fluid intelligence of the participants[31], whereas the impact of inhibitory performance on EFs remains ambiguous[70]. Inter-individual differences in VF performance are also affected by age, education, and sex. Previous studies stressed that both a higher educational level[8] and an increasing age[72] are associated with better VF performance. Moreover, studies demonstrated different searching strategies between men and women in the VF task[73]. Moreover, fluctuating hormonal levels influence the results of studies investigating the concrete relationship of VF and EFs. Depending on the levels of hormones, e.g. progesterone[74] or cortisol[75], participants perform faster or provide more correct answers. As well, intra-individual fluctuating hormonal levels, e.g. influenced by the menstrual cycle in women, cause varying results in EF and VF performance within the same participant[76]. Thus, there is a wide range of intra- and inter-individual differences influencing VF performance. However, due to limited time and costs in

clinical and research context, the assessment of these parameters is not always possible and a comprehensive interpretation of VF performance taking into account individual differences is not a standard procedure.

1.2.3 Advanced analysis of verbal fluency performance

As discussed in the previous chapters, the construct of EFs *perse* as well as the relationship of the different subdomains of EFs and VF performance is very complex. Additionally, intra- and inter individual differences influence EF performance and language-specific abilities. To gain concrete insights into the participant's EF performance, researchers attempt to assess various EF tests tapping into the different domains of EFs and evaluate different aspects of the performance (e.g. reaction times, differentiation between error types). Here, digitalized and automatic analyses of EF testing contribute to a reduced time effort and increase objectivity[28]. In contrast, the VF tasks is still assessed and evaluated manually. The total sum of correct produced words, and occasionally the error types serve as the basis for further interpretations, e.g. in patients suffering from primary progressive aphasia[77]. For many years, the evaluation of sum scores and errors was the only parameter analysed in clinical and research context to draw conclusions on EF performance[78,79]. However, in recent years studies started to investigate the value of additional linguistic features with the aim to provide deeper insights into the cognitive performance in VF tasks. Beside the use of extended VF features *per se*, objective analysis tools were considered, and the field of *computational linguistics* gained more interests to contribute to a more fine-grained analysis of VF performance.

While the sum of correctly produced items is the commonly used parameter to evaluate VF performance, error types, semantic distances and speech breaks have been investigated within the last years to serve as further indicators of EF performance. In detail, perseveration errors (repetition of an already produced item) or category errors provide additional information about working memory capacities[68]. The evaluation of semantic relatedness between different words was shown to quantify different searching strategies within the word retrieval process and reflect switching and clustering strategies e.g., in patients suffering from schizophrenia[80]. Particularly, thought-disordered speech of schizophrenic patients can be quantified by comparing the semantic distances of patients to those from healthy controls[81]. Additionally, the analysis of speech breaks, also known as *latencies*, was shown to convey information about the lexical access speed and planning abilities within the VF task[82].

However, these studies only investigated the potential of single parameters of the VF task and did not address the full potential of the combined analysis of sum scores, errors, sematic relatedness, and latencies to gain deeper insights into EF performance. Moreover, some studies still do not exploit the maximum potential of automated and objective methods from computational linguistics but rather rely on manual semantic analyses[83,84].

1.2.4 Advanced statistical methods

Beside the progression of computational linguistic methods, statistical methods applied in behavioural studies changed during the past years. Previous behavioural studies investigating the relationship of EFs and VF performance, mainly applied classical statistical approaches. Specifically, correlation analyses were used to determine the concrete relationship of the different EF subdomains and VF performance, assessed with the total sum of words. Therefore, group comparisons were calculated to determine disease-specific abnormalities in EF performance.

Taking into account the progress of methods applied in general e.g., in the field of radiology and neuroimaging, it is obvious that machine learning (ML) methods are gaining more and more interest. Due to the high number of medically labelled and digitalized patient data, such as anamnestic information, brain images or medication information, *supervised* ML algorithms can build complex models based on this data to predict e.g., disease specific symptoms[85]. In s*upervised ML* a model is trained based on a set of labelled observations. To ensure generalizability to independent data, the model is validated and tested with unseen data (out-of-sample testing). While the model is trained with the majority of the dataset, the rest of the dataset is held back to validate and test the model[86].

Depending on the algorithm used, it allows for the detection of multivariate interactions and non-linear relationships between the respective features and a specific target. There are two different options for how *supervised* ML methods can be applied depending on the present data structure: If the data is provided in different categories e.g., differentiating between healthy participants and patients with a specific disease, classification approaches are applied. The classifier learns the association between the features and the respective classification, creates the model and applies it to new data. Thus, the classification analysis leads to the percentage of correctly predicted elements. These classification approaches are for example used to classify and identify whether a patient suffers from ADHD, predict the specific subtype of ADHD or identifies the healthy subjects. In case of continuous data, that is not classified, regression

models are applied. Here the results of the model lead to the correlation of true and predicted values of the specific target.
To sum up, ML methods enable the modelling of non-linear and complex relationships while still generalizing to unseen data. However, ML methods require a high number of data points to gain reliable results.

1.3 Aim of the thesis

The main goal of this thesis was to investigate the potential of the VF task to predict EF performance. Therefore, the general construct of EFs was addressed (study 1) and the involvement of EFs in the VF task was investigated (study 2). Finally, a set of comprehensive features extracted from the VF task were identified to predict EF performance (study 3). In all studies, ML methods were applied to calculate predictions and allow for generalization of results.

Study 1

Camilleri, J.A., Eickhoff, S.B., Weis, S., Chen, J., Amunts, J., Sotiras, A. & Genon, S. A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan Executive Function System. *SciRep*. **11**, 16896 (2021).

The goal of study 1 was to investigate common structures of different EF tests and VF tasks to better understand the underlying cognitive processes involved in the VF task. Besides a high number of EF tests tapping in the different domains of EFs, different types of VF tasks were assessed. In detail, the D-KEFS tests battery consisted of nine different tests covering a broad spectrum of verbal and non-verbal EFs. The VF tests included semantic and lexical tasks as well as a switching component. This study built the basis for the following studies and investigated the general construct of EFs.

Study 2

Amunts, J., Camilleri, J.A., Eickhoff, S.B., Heim, S. & Weis, S. Executive functions predict verbal fluency scores in healthy participants. *Sci. Rep.* **10**, 11141 (2020).

The overall aim of study 2 was to better understand the detailed involvement of different EF domains in the VF task. Therefore, the relationship of specific domains of EFs and semantic VF performance was investigated using two different statistical methods. In a first step,

correlation analysis was performed to investigate linear relationships of different EF tests scores and the sum score of the semantic VF tasks. In a second step, a prediction analysis was computed to further elaborate more complex and non-linear relationships of the EF tests and the semantic VF task. Hormonal levels were taken into account to elaborate the influence of individual differences in VF performance. We hypothesised that VF performance is influenced by a combination of cognitive flexibility, working memory and inhibition test results.

Study 3

Amunts, J., Camilleri, J.A., Eickhoff, S.B., Patil, K.R., Heim, S., von Polier, G. & Weis, S. Comprehensive verbal fluency features predict executive function performance. *Sci. Rep*. **11**, 6929 (2021).

In a last step, study 3 investigated the predictive potential of the semantic VF task to draw conclusions on EF performance. Therefore, a comprehensive set of VF information was extracted from the VF task to predict the EF test scores. These VF features contained different aspects of the VF tasks i.e., sum scores, error types, speech breaks and semantic relatedness to cover the variety of cognitive demands that are involved in the VF task. Although isolated relationships of some of these aspects of VF performance and the different EF domains were found in previous studies, we were wondering whether the conglomerate of the various VF features could even predict EF test results applying ML methods.

1.4 Ethics vote

Study 1 used open-access data from the Enhanced Nathan Kline Institute – Rockland Sample (eNKI). The local ethics committee of the Heinrich-Heine University in Düsseldorf approved analysis of the data and all methods were carried out in accordance with relevant guidelines and regulations (study number: 4039).

Study 2 and 3 were performed in accordance with the positive vote by the ethics committee at the Heinrich-Heine University Düsseldorf (study number: 6055R; registration-ID: 2017064341).

2 A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan Executive Function System, Camilleri, J.A., Eickhoff, S.B., Weis, S., Chen, J., Amunts, J., Sotiras, A., Genon, S., Scientific Reports, 10: 11141, (2021)

scientific reports

Check for updates

OPEN

A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan Executive Function System

J. A. Camilleri[1,2]✉, S. B. Eickhoff[1,2], S. Weis[1,2], J. Chen[1,2,3], J. Amunts[1,2], A. Sotiras[4] & S. Genon[1,2]

While a replicability crisis has shaken psychological sciences, the replicability of multivariate approaches for psychometric data factorization has received little attention. In particular, Exploratory Factor Analysis (EFA) is frequently promoted as the gold standard in psychological sciences. However, the application of EFA to executive functioning, a core concept in psychology and cognitive neuroscience, has led to divergent conceptual models. This heterogeneity severely limits the generalizability and replicability of findings. To tackle this issue, in this study, we propose to capitalize on a machine learning approach, OPNMF (Orthonormal Projective Non-Negative Factorization), and leverage internal cross-validation to promote generalizability to an independent dataset. We examined its application on the scores of 334 adults at the Delis–Kaplan Executive Function System (D-KEFS), while comparing to standard EFA and Principal Component Analysis (PCA). We further evaluated the replicability of the derived factorization across specific gender and age subsamples. Overall, OPNMF and PCA both converge towards a two-factor model as the best data-fit model. The derived factorization suggests a division between *low-level* and *high-level* executive functioning measures, a model further supported in subsamples. In contrast, EFA, highlighted a five-factor model which reflects the segregation of the D-KEFS battery into its main tasks while still clustering higher-level tasks together. However, this model was poorly supported in the subsamples. Thus, the parsimonious two-factors model revealed by OPNMF encompasses the more complex factorization yielded by EFA while enjoying higher generalizability. Hence, OPNMF provides a conceptually meaningful, technically robust, and generalizable factorization for psychometric tools.

As of late, research in psychological and medical sciences has been subject to a replication crisis[1–4] that has infiltrated many disciplines interested in human behavior including differential psychology and cognitive neuroscience[2,5–7]. This crisis stems from the finding that a vast number of research results are difficult or impossible to replicate[8]. Several contributing factors have been pointed out and possible solutions have been proposed. Among the contributing factors, the limited sample size and the flexibility in the choice of analysis appear to play an important role[9–12]. Specific choices in the sample selection, measure of interest, and the criteria for significance, together with specific criteria for evaluating the relevance or validity of the analysis' outcomes are examples of factors that directly influence the final findings and conclusions of any study. This problem has been fully acknowledged and extensively discussed in the context of hypothesis-driven studies (i.e., testing a specific psychological effect), and potential solutions for the problem have been suggested. Pre-registration of confirmatory hypotheses has been recommended to limit a-posteriori choices driven by questionable practices such as p-hacking and data-fishing[13]. However, these practices are more difficult to implement in the case of exploratory studies of human behavior, where the analysis is data-driven rather than hypothesis-driven. This

[1]Institute of Neuroscience and Medicine (INM-7 Brain and Behaviour), Forschungszentrum Jülich, Jülich, Germany. [2]Institute of Systems Neuroscience, Heinrich-Heine University, Düsseldorf, Germany. [3]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou, China. [4]Mallinckrodt Institute of Radiology, Institute for Informatics, Washington University in Saint Louis, Saint Louis, USA. ✉email: j.camilleri@fz-juelich.de

applies to the search for latent structure in psychological data capitalizing on multivariate approaches. Actually, the replicability issue has been rarely raised in this domain, despite the influence of the choice of analysis on the findings has been often discussed[14,15].

A popular exploratory method widely used in psychological research is exploratory factor analysis (EFA), which has been introduced in the field by Spearman[16]. It aims to reduce a number of observed variables to fewer unobserved factors in order to identify a hidden structure in the data and to facilitate interpretability[14]. In a conceptual or theoretical perspective, these structures are used as constructs in sophisticated models describing different aspects of human behavior. The established models and structures are then considered as a ground theory on which following studies can build to further characterize human behavior. For example, studies have built on derived factorial models of executive functioning to establish relationships with other aspects of human behavior[17], to examine genetic influences[18], or to propose neural substrates[19] of this cognitive function. In that context, an exploratory factor analysis is generally used to identify latent structure in a set of behavioral variables, such as a test battery, and the derived structure then serves as a model which is usually a-priori imposed on a new dataset using a confirmatory factorial analysis[12]. Nevertheless, as noted by Treiblmaier and Filzmoser[14], many factor solutions can be derived from one correlation matrix and the final solution represents just one of many possible choices. Analyses methods, such as the EFA, involve a number of choices that require the researcher to make crucial decisions that have a substantial impact on the results and subsequent interpretation[20–23]. Such decisions include the number of factors to retain and the criteria used to select this, the type of rotation applied, and the interpretation of the resulting factor solution[24]. These are choices, that, in addition to the data collection aspects such as sample size and test battery, can have an influence on any type of study. Consequently, the lack of replicability of factorizations in the literature has been reported in a number of fields. For instance, one can point out the diverse and inconsistent factor solutions proposed for psychiatric scales[25]; personality scores[26], and executive functioning[27–31]. In this context, and considering the broader framework of the replication crisis in psychological research, it appears necessary to question the utility and generalizability (i.e., the external validity) of exploratory approaches to identify latent structure in psychological tools. Traditionally, Principal Component Analysis (PCA) has also been used for the investigation of the latent structure of behavioral data. To date, the literature is not in agreement as to which method is most appropriate in the context of behavioral data. Many authors argue against the use of PCA mainly because this is considered to be solely a data reduction method and not a true method of factor analysis in a psychological sciences perspective[32–35]. However, other authors disagree[36,37]. Generally, the main point of debate concerns the perspective in which the factorization is applied. As aforementioned, EFA specifically aims to identify hypothetical constructs (also referred to as factors, dimensions, latent variables, synthetic variables or internal attributes). In the behavioral sciences, these latent dimensions are assumed to be unobservable characteristics of people. Accordingly, the factors derived from an EFA are expected to have a theoretical validity. In contrast, PCA aims to provide a summary representation of the original variables into components, without having the specific aim to reflect theoretical constructs. Given their different aims, EFA and PCA have different ways of conceptualizing sources of variance in measured variables. EFA assumes that factors are not perfectly reflected by the measured variables, and thus distinguishes between variance in measures due to the common factors and variance due to unique factors. On the other hand, PCA does not make such a distinction and the resulting components contain a combination of common and unique variance[38]. Considering this distinction further implies that EFA factors are assumed to reflect latent constructs, and thus should not be expected to vary across subsamples. In contrast, from a data-science perspective, PCA and data reduction approaches in general, could be expected to provide different representations depending on the datasets by extracting a simplified representation of the data. Given these differences between the two approaches, the choice of one approach over the other can influence the result, perpetuating the problem of replicability in the identification of latent structures.

Executive functioning is one of the most studied psychological concepts in psychology and is continuously examined in cognitive neuroscience. Executive functioning refers to processes central to coordinated, goal-directed behavior and is thought to play a major role in a wide range of different psychiatric and neurological diseases[39]. However, despite its significance, the true nature of executive abilities remains rather elusive. One of the main reasons for this is that executive functioning is not a single process but rather a "macro-construct" encompassing various aspects of mental functioning[40]. Moreover, the lack of a clear formal definition of executive functioning is also due to the nature of the aspects that constitute it, the relationship among these and their contribution to the overall concept[41]. As a result, there is a constant interest in the study of the structure of executive functioning and its relationship with other traits and behaviors[17]. Throughout the years, several neuropsychological tests have been designed to capture and measure different executive abilities. However, the measurement of executive functioning poses several challenges[41–44] including the fact that executive functioning tests tend to be inherently impure[29]. Executive functioning operates on other cognitive processes, and thus any score derived from an executive functioning task will unavoidably include systematic variance that can be attributed to non-executive functioning processes associated with that specific task context[42,44]. This latter issue is referred to as the task impurity problem and is addressed by using factor analytical techniques. These map the shared variance between tests of executive functioning to a set of latent variables, providing a cleaner estimate of these higher-order cognitive abilities than the individual tests[42,45]. Consequently, numerous studies have investigated the latent structure of executive functioning using different factorization methods and executive functioning batteries. However, the different studies have resulted in diverse findings and conceptual models[27–31]. The long-term study of factors, or components, of executive functioning is thus particularly illustrative of the plurality of latent structures that can be derived from factorization methods in psychological research for a particular concept.

In the clinic, the most popular way of assessing executive functioning is by using test batteries that evaluate the diverse higher-order abilities through multiple tests[44]. One such test battery that has become increasingly common in clinical practice, as well as in research, is the Delis–Kaplan Executive Function System[46]. The D-KEFS

is one of the first normed set of tests developed specifically to assess executive functioning. It consists of nine tests comprising traditional and newly developed tests covering a wide spectrum of verbal and non-verbal executive functions, which are all designed to be stand-alone instruments that can be administered individually or together with other D-KEFS tests. Past studies have used different methods to attempt to evaluate the latent structure of this particular battery, identifying some evidence of diverse latent factors explaining performance on individual tests[17,45,47,48]. In summary, the D-KEFS represents a widely used psychological tool with applications in clinical settings, but for which different factorizations could be proposed in the healthy population.

Considering the heterogeneous factorization results in the literature of executive functioning and psychology in general, generalizability should be a crucial criterion of validity in order to reach a conceptual consensus in psychological sciences. However, as can be seen in the study of executive functioning, a plethora of models exists. In the context of a replicability crisis in psychological sciences, the heterogeneity of models is particularly problematic. The use of different models that examine different aspects of interindividual variability prevents comparison and integration across studies. However, practically evaluating generalizability is hard due to lack of data (and lack of funding support for replicability evaluation). This is particularly the case for factorization analyses, which require large sample sizes for each evaluation. Nevertheless, internal cross-validation can be used to give insight on how the model will generalize to unseen data that are not used for model derivation. As a common approach in the machine learning field, cross-validation consists of the partitioning of a dataset into subsets. The analysis is then performed on one subset (the training set) and validated on the other subset (the test set) across multiple runs with different training and test sets.

In recent years, the increased use of machine learning approaches has emphasized the use of internal cross-validation to increase robustness and to estimate generalizability to an independent dataset. This has led to the popularization of novel methods, which can also be used as factorization techniques, thus offering a novel perspective for behavioral sciences. While these novel approaches are commonly perceived as lacking interpretability and validity when compared to classical statistical approaches, some methods have been developed with the purpose of increasing these aspects by adding additional constraints. One such method, the OPNMF (or Orthonormal Projective Non-Negative Matrix Factorization), provides a relatively higher interpretability as compared to more traditional methods, such as the classic NMF. OPNMF was recently used to identify a robust and generalizable factor structure of Positive and Negative Syndrome Scale (PANSS) data from participants with schizophrenia[25]. The new factor-structure was moreover shown to more reliably relate to specific brain functions than the original PANSS subscales, demonstrating the usefulness of this OPNMF approach[49]. This technique could hence significantly contribute to the definition of robust factorization of psychological variables, in particular for widely used psychological tools, such as standard neuropsychological batteries, socio-affective questionnaires and clinical scales.

The motivation of this study was two-fold. Firstly, given the importance of generalizability in the identification of latent structures, one main goal of the present study was to compare the factorization obtained when using a machine learning approach (OPNMF) with a cross-validation scheme, with the factorization derived from more traditional approaches that tend to lack the generalizability aspect, in particular EFA, but also PCA. Furthermore, a second motivation of this study was to better understand the nature of EF and the tasks commonly used to investigate it. To this end, we capitalized on a large open access dataset of healthy adult scores of the D-KEFS provided by the Enhanced Nathan Kline Institute – Rockland Sample. This dataset is heterogenous in covering the whole adult life span, providing a good gender balance and including participants from the whole population (including different ethnicities), thus making it optimal for this study in which generalizability is central. EFA and PCA were here performed by using standard statistical techniques as implemented in open access statistical tools such as JASP[50]. Furthermore, the choice of the optimal number of factors or components for these traditional approaches was based on recent guidelines in the field, while the choice of the optimal number of components for OPNMF was based on standard criteria assessing not only the quality of the data representation, but also its generalizability. Finally, to further evaluate the quality of the different factorizations, we examine the stability or generalizability across age and gender subsamples.

Methods

Sample and measures. The current study used age-corrected scaled D-KEFS scores of 334 adults (18–85 years old; mean age = 46; 101 males) obtained from the Enhanced Nathan Kline Institute—Rockland Sample (eNKI)[51]. Written informed consent was obtained from all participants. The local ethics committee of the Heinrich-Heine University in Düsseldorf, Germany approved analysis of the data and all methods were carried out in accordance with relevant guidelines and regulations. The main variables of the analyses included 17 D-KEFS Total Achievement Scores (Table 1), which reflect global achievement scores on the 9 tests included in the D-KEFS battery and broadly reflect traditional measures of executive functioning[46]. Only participants that had scores for all 17 variables were included in the study resulting in the exclusion of 385 participants from the original eNKI dataset. Additional information regarding the education level and occupation of the participants can be found in the supplementary material. This study used five different (sub) groups: (1) the full dataset including 334 adults; (2) a subset of the data including only males (n = 101); (3) a subset of the data including only females (n = 233); (4) a subset of the data only including subjects aged over 50 (n = 144); and (5) a subset of the data only including subjects aged 50 or under (n = 220).

The D-KEFS battery offers a wide range of tests that tap into many of the established constructs of executive functioning. The D-KEFS battery includes the following tests: (a) **Trail Making Test**, which aims at assessing attention, resistance to distraction and cognitive flexibility; (b) **Verbal Fluency Test**, which assesses the ability of generating words fluently from overlearned concepts and thus reflects efficient organization of such concepts; (c) **Design Fluency Test**, which is a non-verbal version of the Verbal Fluency Test and assesses the ability of quickly

Test	Variable	Variable description	Measure
Trail making test	Number-Letter Switching	Requires examinees to switch back and forth between connecting numbers and letters in sequence	Completion time [s]
Verbal fluency	Letter Fluency	Requires examinees to say as many words as possible starting with a specific letter in 60 s	Sum of correct responses
	Category Fluency	Requires examinees to say as many words belonging to a specific semantic category in 60 s	Sum of correct responses
	Category Switching	Requires examinees to switch between two specific categories in 60 s	Sum of correct responses
Design fluency	Design Fluency—Filled dots	Measures the examinee's ability to draw as many different designs as possible in 60 s	Total number of correct designs
	Design Inhibition—Empty Dots only	Measures the examinee's ability to draw as many different designs as possible in 60 s while making sure that certain responses are inhibited	Total number of correct designs
	Design Switching	Measures the examinee's ability to draw as many different designs as possible in 60 s while requiring participants to engage in cognitive shifting	Total number of correct designs
Color word interference	CWI—Inhibition	Requires examinee to inhibit reading the words in order to name the dissonant ink colors in which the word is printed	Completion time [s]
	CWI—Switching	Requires examinee to switch back and forth between naming the dissonant ink color and reading the word	Completion time [s]
Sorting test	Confirmed Sorts	Participants are required to sort cards into two groups according to as many different categorization rules or concepts as possible	Total number of correct sorts
	Free Sorting Description	Participants are required to describe the concepts they used to generate each sort	Total number of correct descriptions
	Sort Recognition	Participants are required to identify the correct categorization rule or concept used to sort cards that have been sorted by the examiner	Total number of correct recognitions
Twenty questions test	Initial Abstraction Score	Examinee is shown pictures of common objects and the task is to ask the fewest number of yes/no questions possible to identify the object chosen by the examiner	Minimum number of objects eliminated by first question
	20 Questions—Total Achievement Score		Sum of weighted achievement scores across all items
Word context test	Word Context—Total Achievement Score	Examinee attempts to discover the meaning of a made-up word on the basis of its use in five clue sentences	Consecutively correct items
Tower test	Tower Test—Total Achievement Score	Examinee is required to move disks varying in size across three pegs to build tower in the fewest number of moves possible to match the target tower while following certain rules	Sum of achievement scores (summed up for all items)
Proverb test	Proverb Test—Total Achievement Score	Proverbs are read individually to the examinee who is required to interpret them orally without assistance or cues	Sum of achievement scores (summed up for all items)

Table 1. Description summary of all variables included in the study.

generating designs; (d) **Color-Word Interference Test**, which taps into inhibition and cognitive flexibility by assessing the ability to inhibit an overlearned verbal response in order to generate a conflicting response; (e) **Sorting Test**, aims at measuring multiple components of concept-formation and problem-solving abilities; (f) **Twenty Questions Test**, which assesses the ability to formulate abstract questions and to come up with problem-solving strategies; (g) **Word Context Test**, assesses skills such as deductive reasoning, information integration, hypothesis testing, and flexibility of thinking; (h) **Tower Test**, which assesses spatial planning and rule learning; and (i) **the Proverb Test**, which tests abstraction abilities. All variables included in the present study are presented in Table 1. All variables were examined for outliers and visually inspected for inappropriate distribution. Frequency distributions for each of the 17 EF variables used in the analyses can be found in the supplementary material.

Factorization of D-KEFS scores using OPNMF. NMF is a factorization method that enables the decomposition of a given matrix into two non-negative matrices: (1) a basis matrix with columns representing the resulting latent factors and (2) a factor-loading matrix representing the loading coefficients. The two resulting matrices together should approximate the original data matrix. NMF and its variants have been widely used in various recent biomedical studies including metagene discovery[52], classification of cancer subtypes[53,54], identification of structural brain networks[55], and identification of dimensions of schizophrenia symptoms[25]. Such applications of NMF and its variants have shown that such methods do not require the input data to be normally distributed. One such variant, the OPNMF, has in fact been shown to derive stable and generalizable factor solutions for data with skewed distributions[25,49]. The present study aims at discovering the latent structure of executive functioning by applying this promising method to D-KEFS performance scores. In order to achieve this in an interpretable fashion, the present study adopted a specific variant of NMF, the OPNMF, which adds additional constraints to the algorithm in an effort to promote sparsity and hence improved interpretability to the results[25,55,56].

The OPNMF algorithm was first applied to D-KEFS total achievement scores coming from the whole sample, with the number of factors ranging from 2 to 9. Additionally, the algorithm was applied to the subsets of the dataset that were split by gender and age. The optimal number of factors, and hence the most robust, stable, and generalizable factor model, was identified by using cross-validation in 10,000 split-half analyses[25]. Considering the different sizes of the sub-samples, the cross-validation scheme that was used (i.e., partitioning the dataset into subsets and then performing the analysis on the training set and validating it on the test set across multiple runs with different training and test sets), ensured the robustness of all analyses, including the ones using smaller subsets of the dataset, in a more direct way than classical power and its use in classical statistics. Specifically, the eNKI sample was split into two halves, and OPNMF was performed on each split sample to derive the basis matrix. Subsequently, each item was assigned to a specific factor based on its largest coefficient within the basis matrix. The adjusted Rand index[57], and variation of information[58] were then employed to assess the stability of item-to-factor assignments between the basis matrices derived from the two split samples. Although OPNMF generates almost clustering-like structure, it allows small contributions from multiple items to specific factors. Hence we further evaluated the stability of the whole entries by comparing the two basis matrices as assessed by the concordance index[59]. For the adjusted Rand-index and concordance index, a higher value indicates better stability across splits, while for the variation of information metric, better stability corresponds to lower values. Generalizability was assessed by quantifying out-of-sample reconstruction error by projecting the data of one split sample onto the basis matrix from the other split sample. A lower increase in out-of-sample error compared with within-sample reconstruction error indicates better generalizability[25]. All analyses were run using Matlab R2018a with customized codes, which are available upon request.

PCA and EFA. Data from each of the five different matrices was additionally subjected to exploratory factor analysis (EFA) and principal component analysis (PCA). In both analyses, loading matrices were rotated using promax oblique rotation as currently suggested in the field[15]. An oblique rotation (which allows correlation between the factors) was chosen because of an a priori expectation that higher order factors would reflect a coherent domain of executive functioning, as suggested by the goals of the D-KEFS[46]. Furthermore, previous studies showed that executive functioning tasks tend to be correlated[42,60–62], hence justifying the use of oblique rotation. In both EFA and PCA, the optimal number of factors/components was determined by using two different methods: the Scree test[63] and eigenvalue Monte Carlo simulation approach[64], (i.e., parallel analysis) The Scree Test has been traditionally used for the selection of number of factors and involves plotting the eigenvalues in descending order of their magnitude and determining where they level off to ultimately select the number of meaningful factors that capture a substantial amount of variance in the data[65]. On the other hand, parallel analysis simulates a set of random data with the same number of variables and participants as the real data from which eigenvalues are computed. The eigenvalues extracted from real data that exceed those extracted from random data then indicate the number of factors to retain[15]. This method formally tests the probability that a factor is due to chance and hence minimizes the over-identification of factors based on sampling error[66]. It is thus superior to the reliance upon eigenvalue scores generated by factor analytic processes alone. Parallel Analysis has also been shown to perform well when determining the threshold for significant components, variable loadings, and analytical statistics when decomposing a correlation matrix[67]. Finally, for the reader's information, we also reported here a typical goodness-of-fit measure in EFA, the Tucker-Lewis Index (TLI). TLI reflects the ratio of the model chi-square and a null-model chi-square. In the null-model, the measured variables are uncorrelated (thus there are no latent variables), consequently the null-model has usually a large chi-square (i.e., a poor fit). TLI values express the goodness-of-fit of the found model relative to the null-model and usually range between 0 and 1. As a rule of thumb, a value >0.95 indicates a good fit , a value >0.90 indicates an acceptable fit for and a value <0.90 indicates a poor fit[68].

Results

Optimal number of factors across different factorization approaches and subsamples. Based on results of the stability measures (Fig. 1), the OPNMF analysis on the full dataset indicated a two-factor model as the optimal solution. The adjusted Rand index, variation of information and concordance index between the basis matrices, all indicated the two-factor solution to be the most stable. The transfer reconstruction error indicated that the 2-factor solution was the most generalizable. Stability measures for the OPNMF analyses that were carried out on subsets of the data split by gender and age showed a similar pattern to the ones resulting from the full dataset, thus suggesting a two-factor model for each of the subsets of the data. Both the Scree plot and the Parallel analysis carried out for PCA also indicated that the optimal solution consisted of a 2-factor model for the full dataset analysis. This 2-factor model was consistent for most PCA analyses performed on the data subsets when looking at both selection indices with the exception of the male subset whose scree-plot indicated a 4-factor solution. Consistently, in the case of the EFA analyses, the Scree plot indicated a 2-factor model for the full dataset analysis (TLI = 0.732) as well as for all the analyses performed on the data subsets (male: TLI = 0.670; female: TLI = 0.761; older adults: TLI = 0.745; younger adults: TLI = 0.699, all suggesting a poor fit). However, the parallel analyses results yielded more heterogenous findings. EFA parallel analyses results carried out on the full dataset suggested a 5-factor solution (TLI = 0.931 suggesting an acceptable fit). When the full dataset was split by gender, the EFA analyses results suggested a 3-factor solution for the male subjects only dataset (TLI = 0.837 suggesting a poor fit) and a 5-factor solution for the female subjects only dataset (TLI = 0.906 suggesting an acceptable fit). When the full dataset was split by age, the EFA analyses results suggested a 4-factor solution for both older (TLI = 0.894 suggesting a marginally acceptable fit) and younger (TLI = 0.732 suggesting a poor fit) age groups. Given the previous literature showing that Parallel Analysis performs well (Franklin et al., 1995), as well as the TLI indices that have resulted from our analyses, the Parallel analysis was chosen to be the index of

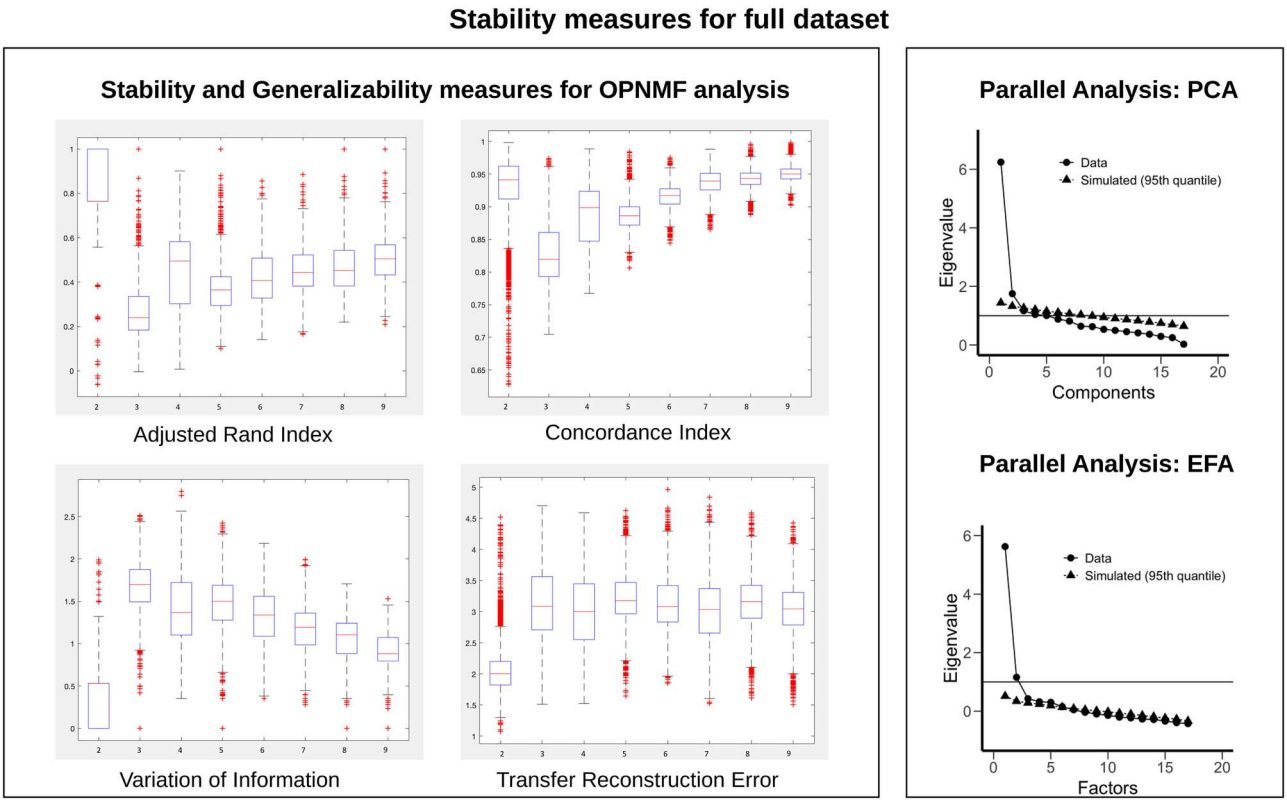

Figure 1. Stability measures for full dataset. Left panel shows plots for each of the stability measures used to identify the most robust factor solution for the OPNMF analysis. The right panel shows plots for the parallel analyses used to identify the most robust component/factor solutions in the PCA and EFA analyses.

choice. Consequently, the results reported below use the factor-model that was indicated by parallel analyses for both EFA and PCA. Figures showing the stability measures for each of the subsets of the data can be found in the supplementary material.

Factorization structure across different factorization approaches and subsamples. In the case of the OPNMF carried out on the full dataset, the resulting two factor solution consisted of one factor strongly loading on Color-Word Interference (CWI), Verbal Fluency and Design Fluency scores and moderately loading on switching components of the Design Fluency Test and the Trail Making Test. The second factor featured strong loadings on the Sorting Test, Proverbs Test, Word Context Test and the 20 Questions Test and a weaker loading on the Tower test (Fig. 2). This pattern was mostly consistent throughout the different subsamples of the data that were split by gender and age, with some minor exceptions. In the case of males only dataset, both switching components of the Verbal Fluency Test and the Trail Making Test showed weak loadings onto the first factor, while the switching component of the Design Fluency Test showed a stronger loading. In the case of females only dataset, the Word Context Test showed weak loadings onto the second factor together with the Tower Test. When the full dataset was split by age, the Tower Test, Proverb Test and Word Context Test all showed weak loadings onto the second factor in the dataset consisting of older adults, while the 20 Questions Test loaded weakly onto the second factor together with the Tower Test in young adults. Noticeably, all subsamples showed the same tests loading onto each of the two factors.

The PCA analyses resulted in component models that showed patterns that were strikingly similar to the OPNMF models for the full dataset as well as for each of the subsets. The component model resulting from the analysis of the full data set resulted in a two-factor solution that consisted of one factor strongly loading on CWI scores and Design Fluency scores and moderately loading on Verbal Fluency Scores and the Trail Making Test. The second factor featured strong loadings from the Sorting Test, moderate loadings from the Proverbs Test and Word Context Test and a weaker loading for the Tower test and the 20 Questions Test (Fig. 2). This pattern was repeated when the PCA analyses were carried out on subsets of female sand younger adults. When the PCA analysis was run on a subset that included only males, the factor solution consisted of one factor strongly loading on CWI scores and Design Fluency scores, moderately loading on Verbal Fluency Scores and the Trail Making Test and weakly loading on the Tower Test and the 20 Question Test. The second factor featured strong loadings from the Sorting Test, moderate loadings from the Proverbs Test, Word Context Test and weaker loading on the 20 Questions Test. The factor solution for the males only dataset consisted of one factor strongly loading on CWI scores and Design Fluency scores, moderately loading on Verbal Fluency Scores and the Trail Making Test and

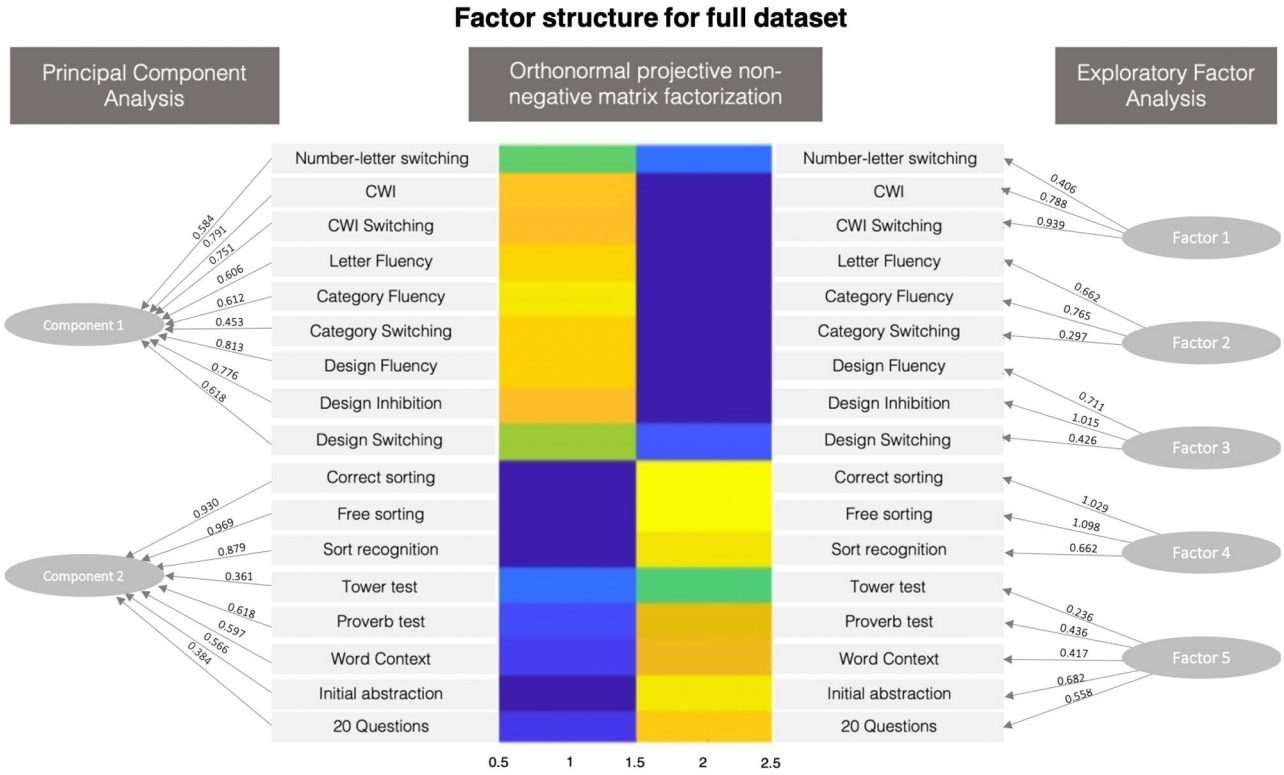

Figure 2. Factor structure and factor loadings resulting from the PCA, EFA and OPNMF analyses for the full data set. Figures show strongest loadings for each variable.

weakly loading on the Tower Test and the Word Context Test. The second factor featured strong loadings from the Sorting Test, moderate loadings from the Proverbs Test and the 20 Questions Test.

The EFA analyses resulted in a more heterogenous picture. The EFA analysis of the full dataset resulted in a five-factor solution consisting of one factor including scores from the Sorting Test; one factor that included scores from the CWI Test and the TMT test; one factor including scores from the Design Fluency Test; one factor including scores from the Proverbs Test, Word Context Test, 20 Questions Test and Tower Test; and another factor including scores from the Verbal Fluency Test. The EFA results for the males only dataset showed a three-factor solution with one factor including scores from the Sorting Test; one factor including scores from the Tower Test, Word Context Test and 20 Questions Test and the switching component of the Verbal Fluency Test; and one factor including the rest of the scores from the Verbal Fluency Test; the Trail Making Test, the Color-Word Interference Test, and the Design Fluency Test. In the females only dataset, the resulting factor structure consisted of a five-factor solution with one factor including scores from the Sorting Test; one factor including scores from the Verbal Fluency Test; one factor including two scores from the Design Fluency Test; one factor including the Trail Making Test, scores from the Color-Word Interference Test, the switching component of the Design Fluency Test, the Tower Test and the Word Context Test; and a final factor including scores from the Proverb Test and 20 Questions Test. When the full dataset was split by age, the EFA resulted in a four-factor solution in both subsets. In the case of the older adults dataset, the resulting factor structure consisted of one factor including scores from the Sorting Test; one factor including scores from the Verbal Fluency Test, the Color-Word Interference Test, the Trail Making Test and the Word Context Test; one factor including scores from the Design Fluency Test and the Tower Test; and a final factor including scores from the Proverb Test and 20 Questions Test. In the case of the younger adults dataset, results showed one factor including scores from the Verbal Fluency Test; one factor including scores from the Design Fluency Test; one factor including scores from the Color-Word Interference Test, and the Trail Making Test; and a factor grouping scores from the Sorting Test, Tower Test, Proverb Test, Word Context and 20 Questions Test. Result figures for each of the subsets can be found in the supplementary material. Importantly, all EFA and PCA analyses were replicated using another open access statistical software, Jamovi[69] (version 1.2, https://www.jamovi.org), and resulted in virtually identical results.

Discussion

Although the field of psychology has acknowledged and discussed the existence of a replicability crisis extensively, this issue has received less attention in the context of multivariate approaches for psychometric data factorization. This has resulted in heterogenous factorization results for several constructs in psychology, including executive functioning. Given the importance of replicability and generalizability in the identification of latent structures, the main goal of the present study was to compare the factorization obtained when using a machine learning approach (OPNMF) with a cross-validation scheme with the factorization derived from more traditional approaches, in particular EFA, but also PCA, in the D-KEFS. These latter approaches were performed

as typically implemented in standard statistical software and following current guidelines, which usually do not include generalizability evaluation. In addition to the evaluation of factorization approaches, this study provides further insight into the specific nature of the D-KEFS and hence also contributes more generally to the understanding of executive functioning. The following paragraphs start with a discussion of the results of the EFA analysis with regards to previous literature together with EFA theoretical background. We then discuss the convergent results obtained when using OPNMF and PCA from a methodological point of view and also with regards to previous literature on executive functioning and the related evaluation tools. Finally, we discuss the resulting two-factor solution in the context of a parsimonious and robust representation of executive functioning for various applications.

EFA analysis. Using traditional EFA analysis, our investigations of the factorization across subsamples first indicate that the optimal solution can vary across subsamples, hence suggesting that the generalizability of the factor solution derived by an EFA analysis can be relatively limited. Overall, in the whole dataset, a five-factor solution appeared to be the best model fit. This result suggests a segregation that reflects the structure of the D-KEFS battery with the Sorting, Design Fluency and Verbal Fluency Tests each being assigned to their own factors, while tasks that require a certain level of abstraction and problem-solving abilities were grouped together in one factor. Thus, overall, the factorial analysis was here strongly influenced by the specific structure of the test battery that was used. It is noteworthy that this finding is somewhat contradictory with the core assumption behind EFA that states that EFA reveals unobservable latent variables reflecting meaningful psychological constructs. A similar, albeit not identical structure is seen when performing an EFA on females only. In the case of males, results suggested a three-factor solution, while in both younger and older adults the EFA indicated a four-factor solution. The evaluation of the theoretical validity of the factorization derived here by the EFA in a psychological science perspective is complicated by the fact that the literature reports a multitude of different factor models, including various factor solutions, all using different methods of factorization, datasets and test batteries. In particular, similar exploratory studies that used EFA have also resulted in heterogenous factor solutions ranging from one[70] to six factors[71].

One model of executive functioning that has acquired a significant amount of empirical support is the three-factor model by Miyake et al.[42]. This influential study uses a confirmatory analysis approach as opposed to the exploratory approach established in the present study, and factorizes executive functioning into shifting, inhibition and updating. Shifting refers to the ability to switch between operations and perform new operations while being faced with interference[42]. Inhibition requires the ability to purposefully control automatic or dominant responses[42]. Finally, the updating factor represents tasks that require the monitoring and evaluating of new information and, if necessary, the updating of information in working memory for the successful completion of the task at hand[42]. Interestingly, the EFA findings of the present study do not overlap with the shifting, inhibition and updating factors suggested by Miyake et al.[42]. However, it is noteworthy that the three-factor model presented by Miyake and colleagues[42] was based on a limited set of tasks and did not include an exhaustive list of executive functions. Specifically, Miyake's study[42] and others[61,72], have focused mostly on tasks that require simpler cognitive abilities, and thus tend to not include tasks that tap into more complex abilities, such as problem-solving, abstraction and strategic thinking. On the other hand, the D-KEFS battery, which was used in the present study, offers a wide range of tests that tap into many of the established constructs of executive functioning, including more complex abilities, such as abstraction, reasoning, and problem solving[46,73]. Unsurprisingly, the specific set of tasks used will heavily impact the resulting factor model. The literature does include studies that have attempted to factorize D-KEFS measures using both confirmatory and exploratory approaches. Hence, Karr and colleagues[45] used confirmatory factor analysis, which led them to the conclusion that the D-KEFS taps into three EF factors, namely, inhibition, shifting and fluency. However, this study chose not to include tasks that tap into more complex abilities (i.e., Twenty Questions, Word Context, and Proverb Tests) in the input variables. On the other hand, Latzman and colleagues[17] used EFA to factorize D-KEFS measures and reported a three-factor model comprising Conceptual Flexibility, Monitoring and Inhibition, which was likened to the Miyake model by the authors.

A number of subsequent studies have supported the three factors of shifting, inhibition and updating presented by Miyake et al.[42] by reporting similar three factor solutions from a series of confirmatory factor analyses of diverse cognitive tasks[45,61,72,74]. Other similar confirmatory approaches have resulted in different factor solutions depending on the age group that was investigated[75–79]. To further understand the heterogeneity of findings reported in the literature and the divergence between the results of the EFA in the current study and previous conceptualization, it is important to note here that there is a fundamental difference between confirmatory and exploratory approaches in terms of their use to identify latent factors. Confirmatory approaches, such as Confirmatory Factor Analyses, use knowledge of the theory of the construct and previous empirical findings to test a hypothesis that has been postulated a priori. Therefore, the aim of this approach is to verify a specific factor structure of a set of observed variables. This approach will hence provide an evaluation that is in alignment with current research[45], however will be undeniably impacted by the initial research hypothesis used. On the other hand, exploratory approaches identify the underlying factor structure of a set of variables without the need of establishing an a priori hypothesis. The latter, thus, allows for the deeper understanding of a construct in an exploratory fashion. In other words, confirmatory approaches can be considered as "hypothesis-driven" approaches to some extent, while exploratory approaches can be considered as "data-driven" approaches. Differences in results when comparing confirmatory and exploratory factor analyses are therefore not surprising.

OPNMF and PCA. While suggesting a different factorization than EFA, PCA and OPNMF together converge toward a similar 2-component model. It is noteworthy that this convergence was observed despite the fact that the choice of optimal factor solution was based on different criteria within and between approaches

including the part of variance explained, data representation quality and stability evaluations. PCA and OPNMF factorization methods here resulted in one factor that designated loadings to Color-Word Interference scores, Verbal Fluency, Design Fluency Test and the Trail Making Test. The second factor featured strong loadings from the Sorting Test, Proverbs Test, Word Context Test and the 20 Questions Test and a weaker loading for the Tower test. These results seem to indicate a division between tasks that require monitoring and task-switching, and more complex tasks that require concept formation, abstraction, and problem-solving. Specifically, tasks that require a certain level of abstraction, strategic thinking and problem-solving abilities, such as the Sorting Test, Twenty Questions Test, Word Context Test, Tower Test, and the Proverb Test, were all grouped into one factor. On the other hand, tasks that require less complex abilities were grouped in another factor. The latter factor includes tests that tap into abilities such as monitoring, fluency, cognitive flexibility, and inhibition. Hence, in contrast to previous results, our results obtained from the OPNMF and PCA analyses suggest a stable and robust two factor model indicating a division between *Simple* and *Complex* (or *low*- vs *high-level*) executive functioning tasks. While previous factorization findings of executive functioning do not seem to support our findings indicating a split between *Simple* and *Complex* tasks, it has been previously shown that people suffering from executive functioning impairment, such as in the case of patients with mild cognitive impairment, tend to exhibit selective rather than global impairment with some studies showing a separation between impairment on simple versus more complex tasks[80–82]. The idea of simple versus complex is also reflected in neurobiological literature in which a separation of tasks between the dorsolateral prefrontal cortex and the ventrolateral prefrontal cortex[83] has been suggested. The former has been implicated in the context of more complex aspects of executive functioning although not all evidence supports this[83]. The notion of separation of tasks based on complexity is also in line with the proposed hierarchical organization of the frontal cortex[84]. When taking a deeper look at the individual measures that were included in the present study, it becomes apparent that there is a noteworthy difference between the different measurement approaches used and the subsequent processes that they could be eliciting. Specifically, while some of the variables are measures of accuracy (e.g., correct number of items), others rely more heavily on time pressure and processing speed (e.g., reaction time and completion time). This difference in measurement approaches seems to be reflected in the resulting dichotomy between *Simple* and *Complex* tasks. In fact, whereas the *Complex* tasks quite clearly emphasize accuracy, the *Simple* tasks appear to be more overtly driven by the element of time. The number-letter switching task, CWI and CWI switching are all direct measures of time while the other variables that have been grouped together with the *Simple* factor are measures of fluency which arguably also involves an aspect of time pressure since its measurement is related to time efficiency when recalling items. Additionally, although the factor labelled as *Complex* in the present study includes measures that tap into abilities such as reasoning, abstraction, problem-solving, and strategic thinking, this factor also includes measures coming from the Sorting Test. The D-KEFS Sorting Test and tests with a similar procedure, such as the Wisconsin Card Sorting Test[85], have been traditionally associated with the *Shifting* or *Conceptual Flexibility* factor[17,42,45,74]. This association appears to be appropriate since the Sorting Test and its variants require participants to shift from previous sorting rules to new rules to achieve a greater number of accurate sorts. However, the Sorting Test also taps into more abstract problem-solving strategies that go beyond simple shifting. This complexity of the Sorting Test is reflected by the results of the present study. Thus, the factorization derived from PCA and OPNMF appears parsimonious and meaningful from a psychological construct standpoint. This study hence demonstrated that the application of machine learning approaches to psychometric data can provide interpretable outcomes in a psychological science perspective. It should be noted here as well that OPNMF further promotes out-of-sample generalizability by evaluating reconstruction error in a left out set across multiple runs, which is a crucial aspect considering the replication issues in psychological sciences.

Despite the apparent divergence of factorization results between EFA on the one hand and OPNMF and PCA on the other hand, it should be noted that the results of our EFA analyses provide a higher factor model that reflects the segregation of tasks that was used in the D-KEFS battery while still assigning a single factor to tasks that require abstraction and problem-solving skills. Hence, the parsimonious two factor model can also be seen as encompassing the more complex factorization yielded by EFA. The results of the present study thus suggest that the OPNMF and PCA results provide a robust and stable two factor solution that separates tasks that require monitoring and task-switching from more complex tasks that require concept formation, abstraction, and problem-solving. Considering all the points discussed above, together with the fact that both methods converged towards one robust model, we suggest that our results may reflect a robust factor model that applies across a wide age range and across different factorization methods. Given the uncertainty and diverse findings of the factorial structure of executive functioning in the literature, this model offers a more scientifically parsimonious model from both technical and conceptual standpoints. From a technical standpoint, the approach established in the present study (i.e., that of reaching a consensus among different technical variations) is the most reasonable to our knowledge since it is commonly known that different approaches can result in different factor solutions. From a conceptual standpoint, the 2 factor solution presented in this study results in a scientifically parsimonious model since the differentiation between *Simple* and *Complex* is better at reflecting consensual real-world concepts than models with a higher number of factors. Considering these scientific qualities, the robust and parsimonious two-factor model that emerged from this study should be of higher practical utility for characterizing inter-individual variability in executive functioning performance at both the biological level (such as genetic and brain subtrates) and the environnmental level (external factors).

Summary and conclusion. In addition to demonstrating the advantages of a machine learning approach for the factorization of psychometric data in a replicability perspective, this study also provides a robust model of factorization of the D-KEFS. The derived factorization suggests a division between *low-level* and *high-level*

executive functioning measures, a model further supported in subsamples. In contrast, EFA, highlighted a five-factor model as the better fit to the overall cohort, but which was poorly supported in the subsamples. This five-factor factorization reflects the segregation of the D-KEFS battery into its main tasks while still clustering higher-level tasks together. Thus, the parsimonious two-factors model revealed by OPNMF underlies the more complex factorization yielded by EFA while enjoying higher generalizability. Hence the application of OPNMF to psychometric data in the present study provides conceptually meaningful, technically robust and generalizable factorization for psychometric tools.

Received: 29 March 2021; Accepted: 9 August 2021
Published online: 19 August 2021

References

1. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**(8), e124 (2005).
2. Masouleh, S. K., Eickhoff, S. B., Hoffstaedter, F., Genon, S. & Alzheimer's Disease Neuroimaging Initiative. Empirical examination of the replicability of associations between brain structure and psychological variables. *Elife* **8**, e43464 (2019).
3. Lindsay, D. S. Replication in psychological science. *Psychol. Sci.* **26**, 1827–1832 (2015).
4. Pashler, H. & Wagenmakers, E. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspect. Psychol. Sci.* **7**(6), 528–530 (2012).
5. Avinun, R., Israel, S., Knodt, A. R. & Hariri, A. R. Little evidence for associations between the big five personality traits and variability in brain gray or white matter. *Neuroimage* **220**, 117092 (2020).
6. Boekel, W. *et al.* A purely confirmatory replication study of structural brain-behavior correlations. *Cortex* **66**, 115–133 (2015).
7. Genon, S. *et al.* Searching for behavior relating to grey matter volume in a-priori defined right dorsal premotor regions: Lessons learned. *Neuroimage* **157**, 144–156 (2017).
8. Shrout, P. E. & Rodgers, J. L. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annu. Rev. Psychol.* **69**, 487–510 (2018).
9. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 1–7 (2020).
10. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**(11), 1359–1366 (2011).
11. Carp, J. On the plurality of (methodological) worlds: Estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6**, 149 (2012).
12. Martínez, K. *et al.* Reproducibility of brain-cognition relationships using three cortical surface-based protocols: An exhaustive analysis based on cortical thickness. *Hum. Brain Mapp.* **36**(8), 3227–3245 (2015).
13. Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* **7**(6), 632–638 (2012).
14. Treiblmaier, H. & Filzmoser, P. Exploratory factor analysis revisited: How robust methods support the detection of hidden multivariate data structures in IS research. *Inform. Manag.* **47**(4), 197–207 (2010).
15. Watkins, M. W. Exploratory factor analysis: A guide to best practice. *J. Black Psychol.* **44**(3), 219–246 (2018).
16. Spearman, C. "General intelligence," objectively determined and measured. *Am. J. Psychol.* **15**, 201–292 (1904).
17. Latzman, R. D. & Markon, K. E. The factor structure and age-related factorial invariance of the Delis–Kaplan Executive Function System (D-KEFS). *Assessment* **17**(2), 172–184 (2010).
18. Friedman, N. P. *et al.* Individual differences in executive functions are almost entirely genetic in origin. *J. Exp. Psychol. Gen.* **137**(2), 201 (2008).
19. Collette, F., Hogge, M., Salmon, E. & Van der Linden, M. Exploration of the neural substrates of executive functioning by functional neuroimaging. *Neuroscience* **139**(1), 209–221 (2006).
20. Armstrong, J. S. & Soelberg, P. On the interpretation of factor analysis. *Psychol. Bull.* **70**(5), 361 (1968).
21. Comrey, A. L. Common methodological problems in factor analytic studies. *J. Consult. Clin. Psychol.* **46**(4), 648 (1978).
22. MacCallum, R. A comparison of factor analysis programs in SPSS, BMDP, and SAS. *Psychometrika* **48**(2), 223–231 (1983).
23. Weiss, D. J., Rand McNally and Co & United States of America. Multivariate procedures. In *Handbook of Industrial and Organizational Psychology* (ed. Dunnette, M. D.) see ncj-52907 (1976).
24. Ford, J. K., MacCallum, R. C. & Tait, M. The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Pers. Psychol.* **39**(2), 291–314 (1986).
25. Chen, J. *et al.* Neurobiological divergence of the positive and negative schizophrenia subtypes identified on a new factor structure of psychopathology using non-negative factorization: An international machine learning study. *Biol. Psychiatr.* **87**(3), 282–293 (2020).
26. Blackburn, R., Renwick, S. J., Donnelly, J. P. & Logan, C. Big five or big two? superordinate factors in the NEO five factor inventory and the antisocial personality questionnaire. *Personal. Individ. Differ.* **37**(5), 957–970 (2004).
27. Amieva, H., Phillips, L. & Della Sala, S. Behavioral dysexecutive symptoms in normal aging. *Brain Cogn.* **53**(2), 129–132 (2003).
28. Bennett, P. C., Ong, B. & Ponsford, J. Assessment of executive dysfunction following traumatic brain injury: Comparison of the BADS with other clinical neuropsychological measures. *J. Int. Neuropsychol. Soc.: JINS* **11**(5), 606 (2005).
29. Burgess, P. W. Theory and methodology in executive function research. In *Methodology of Frontal and Executive Function* 87–121 (Routledge, 2004).
30. Chan, R. C. Dysexecutive symptoms among a non-clinical sample: A study with the use of the dysexecutive questionnaire. *Br. J. Psychol.* **92**(3), 551–565 (2001).
31. Robbins, T. W. *et al.* A study of performance on tests from the CANTAB battery sensitive to frontal lobe dysfunction in a large sample of normal volunteers: Implications for theories of executive functioning and cognitive aging. *J. Int. Neuropsychol. Soc.* **4**(5), 474–490 (1998).
32. Bentler, P. M. & Kano, Y. On the equivalence of factors and components. *Multivar. Behav. Res.* **25**(1), 67–74 (1990).
33. Floyd, F. J. & Widaman, K. F. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol. Assess.* **7**(3), 286 (1995).
34. Gorsuch, R. L. Common factor analysis versus component analysis: Some well and little known facts. *Multivar. Behav. Res.* **25**(1), 33–39 (1990).
35. Costello, A. B. & Osborne, J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract. Assess. Res. Eval.* **10**(1), 7 (2005).
36. Arrindell, W. A. & Van der Ende, J. An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Appl. Psychol. Meas.* **9**(2), 165–178 (1985).
37. Guadagnoli, E. & Velicer, W. F. Relation of sample size to the stability of component patterns. *Psychol. Bull.* **103**(2), 265 (1988).

38. Conway, J. M. & Huffcutt, A. I. A review and evaluation of exploratory factor analysis practices in organizational research. *Organ. Res. Methods* **6**(2), 147–168 (2003).
39. Zelazo, P. D. & Müller, U. Executive function in typical and atypical development. In *Handbook of Childhood Cognitive Development* 445–469 (2002).
40. Zelazo, P. D., Carter, A., Reznick, J. S. & Frye, D. Early development of executive function: A problem-solving framework. *Rev. Gen. Psychol.* **1**(2), 198–226 (1997).
41. Lezak, M. D. The problem of assessing executive functions. *Int. J. Psychol.* **17**(1–4), 281–297 (1982).
42. Miyake, A. *et al.* The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cogn. Psychol.* **41**(1), 49–100 (2000).
43. Jurado, M. B. & Rosselli, M. The elusive nature of executive functions: A review of our current understanding. *Neuropsychol. Rev.* **17**(3), 213–233 (2007).
44. Miyake, A. & Friedman, N. P. The nature and organization of individual differences in executive functions: Four general conclusions. *Curr. Dir. Psychol. Sci.* **21**(1), 8–14 (2012).
45. Karr, J. E. *et al.* The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychol. Bull.* **144**(11), 1147 (2018).
46. Delis, D. C., Kaplan, E. & Kramer, J. H. *Delis–Kaplan Executive Function System* (2001).
47. Floyd, R. G., Bergeron, R., Hamilton, G. & Parra, G. R. How do executive functions fit with the Cattell–Horn–Carroll model? Some evidence from a joint factor analysis of the Delis–Kaplan executive function system and the Woodcock–Johnson III tests of cognitive abilities. *Psychol. Sch.* **47**(7), 721–738 (2010).
48. McFarland, D. J. Factor-analytic evidence for the complexity of the Delis–Kaplan Executive Function System (D-KEFS). *Assessment* **27**(7), 1645–1656 (2020).
49. Chen, J. *et al.* Intrinsic connectivity patterns of task-defined brain networks allow individual prediction of cognitive symptom dimension of schizophrenia and are linked to molecular architecture. *Biol. Psychiatr.* **89**, 308–319 (2020).
50. Love, J. *et al.* Software to sharpen your stats. *APS Obs.* **28**(3), 27–29 (2015).
51. Nooner, K. B. *et al.* The NKI-rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Front. Neurosci.* **6**, 152 (2012).
52. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**(6), 600–606 (2016).
53. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**(11), 1108–1115 (2013).
54. Sadanandam, A. *et al.* A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.* **19**(5), 619–625 (2013).
55. Sotiras, A., Resnick, S. M. & Davatzikos, C. Finding imaging patterns of structural covariance via non-negative matrix factorization. *Neuroimage* **108**, 1–16 (2015).
56. Yang, Z. & Oja, E. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Trans. Neural Netw.* **21**(5), 734–749 (2010).
57. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985).
58. Meilă, M. Comparing clusterings—An information based distance. *J. Multivar. Anal.* **98**(5), 873–895 (2007).
59. Raguideau, S., Plancade, S., Pons, N., Leclerc, M. & Laroche, B. Inferring aggregated functional traits from metagenomic data using constrained non-negative matrix factorization: Application to fiber degradation in the human gut microbiota. *PLoS Comput. Biol.* **12**(12), e1005252 (2016).
60. Fisk, J. E. & Sharp, C. A. Age-related impairment in executive functioning: Updating, inhibition, shifting, and access. *J. Clin. Exp. Neuropsychol.* **26**(7), 874–890 (2004).
61. Lehto, J. E., Juujärvi, P., Kooistra, L. & Pulkkinen, L. Dimensions of executive functioning: Evidence from children. *Br. J. Dev. Psychol.* **21**(1), 59–80 (2003).
62. Hull, R., Martin, R. C., Beier, M. E., Lane, D. & Hamilton, A. C. Executive function in older adults: A structural equation modeling approach. *Neuropsychology* **22**(4), 508 (2008).
63. Cattell, R. B. The scree test for the number of factors. *Multivar. Behav. Res.* **1**(2), 245–276 (1966).
64. Horn, J. L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**(2), 179–185 (1965).
65. D'agostino, R. B. & Russell, H. K. Scree test. In *Encyclopedia of Biostatistics*, Vol. 7 (2005).
66. Wood, N. D., Akloubou Gnonhosou, D. C. & Bowling, J. W. Combining parallel and exploratory factor analysis in identifying relationship scales in secondary data. *Marriage Fam. Rev.* **51**(5), 385–395 (2015).
67. Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T. & Fralish, J. S. Parallel analysis: A method for determining significant principal components. *J. Veg. Sci.* **6**(1), 99–106 (1995).
68. McDonald, R. P. & Marsh, H. W. Choosing a multivariate model: Noncentrality and goodness of fit. *Psychol. Bull.* **107**(2), 247 (1990).
69. The jamovi Project. *jamovi* (Version 1.2) [Computer Software] (2021). Retrieved from https://www.jamovi.org.
70. Deckel, A. W. & Hesselbrock, V. Behavioral and cognitive measurements predict scores on the MAST: A 3-year prospective study. *Alcohol.: Clin. Exp. Res.* **20**(7), 1173–1178 (1996).
71. Testa, R., Bennett, P. & Ponsford, J. Factor analysis of nineteen executive function tests in a healthy adult population. *Arch. Clin. Neuropsychol.* **27**(2), 213–224 (2012).
72. Vaughan, L. & Giovanello, K. Executive function in daily life: Age-related influences of executive processes on instrumental activities of daily living. *Psychol. Aging* **25**(2), 343 (2010).
73. Baron, S. I. Delis–Kaplan executive function system. *Child Neuropsychol.* **10**(2), 147–152 (2004).
74. Karr, J. E., Hofer, S. M., Iverson, G. L. & Garcia-Barrera, M. A. Examining the latent structure of the Delis–Kaplan executive function system. *Arch. Clin. Neuropsychol.* **34**(3), 381–394 (2019).
75. Brydges, C. R., Reid, C. L., Fox, A. M. & Anderson, M. A unitary executive function predicts intelligence in children. *Intelligence* **40**(5), 458–469 (2012).
76. 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), 663–676 (2011).
77. Fournier-Vicente, S., Larigauderie, P. & Gaonac'h, D. More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. *Acta Physiol. (Oxf.)* **129**(1), 32–48 (2008).
78. de Frias, C. M., Dixon, R. A. & Strauss, E. Structure of four executive functioning tests in healthy older adults. *Neuropsychology* **20**(2), 206 (2006).
79. Glisky, E. L. *et al.* Differences between young and older adults in unity and diversity of executive functions. *Aging Neuropsychol. Cogn.* **8**, 1–26 (2020).
80. Traykov, L. *et al.* Executive functions deficit in mild cognitive impairment. *Cogn. Behav. Neurol.* **20**(4), 219–224 (2007).
81. Zhang, Y., Han, B., Verhaeghen, P. & Nilsson, L. Executive functioning in older adults with mild cognitive impairment: MCI has effects on planning, but not on inhibition. *Aging Neuropsychol. Cogn.* **14**(6), 557–570 (2007).
82. Brandt, J. *et al.* Selectivity of executive function deficits in mild cognitive impairment. *Neuropsychology* **23**(5), 607 (2009).
83. Elliott, R. Executive functions and their disorders: Imaging in clinical neuroscience. *Br. Med. Bull.* **65**(1), 49–59 (2003).

84. Badre, D. & Nee, D. E. Frontal cortex and the hierarchical control of behavior. *Trends Cogn. Sci.* **22**(2), 170–188 (2018).
85. Berg, E. A. A simple objective technique for measuring flexibility in thinking. *J. Gen. Psychol.* **39**(1), 15–22 (1948).

Acknowledgements
This study was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/2–1, EI 816/16-1 and EI 816/21-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain", the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement 785907 (HBP SGA2), 945539 (HBP SGA3), The Virtual Brain Cloud (EU H2020, no. 826421), and the National Institute on Aging (R01AG067103).

Author contributions
J.A.C., S.G. and S.B.E. planned and designed the study. J.C. and A.S. provided tools necessary for the analysis. J.A.C. processed the data, performed the analysis, drafted the manuscript and designed the figures. S.G. edited the manuscript and oversaw the study. J.A. helped with the designing of the figures and tables. J.A.C., S.G., S.B.E., J.A., and S.W. contributed to the interpretation of the results. All authors provided critical feedback and commented on the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-96342-3.

Correspondence and requests for materials should be addressed to J.A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3 Executive functions predict verbal fluency scores in healthy participants, Amunts, J., Camilleri, J.A., Eickhoff, S.B., Heim, S., Weis, S., Scientific Reports, 10: 11141, (2020)

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Executive functions predict verbal fluency scores in healthy participants

Julia Amunts[1,2] ✉, Julia A. Camilleri[1,2], Simon B. Eickhoff[1,2], Stefan Heim[3,4] & Susanne Weis[1,2]

While there is a clear link between impairments of executive functions (EFs), i.e. cognitive control mechanisms that facilitate goal-directed behavior, and speech problems, it is so far unclear exactly which of the complex subdomains of EFs most strongly contribute to speech performance, as measured by verbal fluency (VF) tasks. Furthermore, the impact of intra-individual variability is largely unknown. This study on healthy participants (n = 235) shows that the use of a relevance vector machine approach allows for the prediction of VF performance from EF scores. Based on a comprehensive set of EF scores, results identified cognitive flexibility and inhibition as well as processing speed as strongest predictors for VF performance, but also highlighted a modulatory influence of fluctuating hormone levels. These findings demonstrate that speech production performance is strongly linked to specific EF subdomains, but they also suggest that inter-individual differences should be taken into account.

Executive functions (EFs) refer to a set of cognitive processes that allow for goal-directed behavior through the regulation of various cognitive subprocesses. Since EFs permeate behavior, they also impact daily activities as well as social and personal development, including school or job success[1]. The importance and pervasiveness of EFs has led different fields of study to investigate these control mechanisms with the goal of differentiating the various subdomains of EFs. This, in turn, has resulted in a number of different conceptualizations based on different approaches, all attempting to subdivide EFs into different domains. While a consensus does not yet exist about how exactly to subdivide and name EFs, there is general agreement that there are three core EFs: (1) cognitive flexibility, (2) working memory and (3) inhibition[1] (but see Karr *et al.*[2,3]). Higher-order EFs, such as reasoning, planning and problem solving, are then built on the basis of these subdomains.

The various sub-domains of EFs have been shown to be impaired in a number of neurological and psychiatric diseases, such as attention-deficit/hyperactivity disorder (ADHD)[4], Parkinson's disease[5], depression[6] and schizophrenia[7]. Different diseases present their own typical EF deficits and clinical diagnosis attempts to assess the specific patterns of the disease. For example, in the case of Parkinson's disease, patients suffer from difficulties in dual-tasking which is reflected in the deficient combination of memorizing and manipulation of thoughts and tasks[8] but also in impaired speech characterized by semantic paraphasias and reduced word fluency due to a lack of EFs[5,9]. To assess these symptoms different test batteries have been developed. These batteries, which include tests tapping into the different EF sub-domains, are used for neuropsychological assessment in both clinical settings and lab-based environments. Commonly used batteries are the Delis-Kaplan Executive Function System (D-KEFS) and the *Vienna Test System*, both of which offer a wide range of tests probing each of the EF sub-domains and have been independently validated[10–12]. Commonly used tasks tapping into the different sub-domains of EFs comprise the *Wisconsin Card Sorting test* (WCST), *Tower of London* (ToL) and *Trail-Making* test (TMT) to assess cognitive flexibility[5,13,14], *n-back* tasks and the *Corsi block tapping* test to cover the sub-domain of working memory[15,16] and the *Stop-signal* task or the *Stroop* test (color-word interference) to probe the sub-domain of inhibition. All of these are commonly used in the clinical[5,17] as well as in the scientific context[14,18]. Importantly, due to the overlap of the different domains of EFs, these tests cannot be assumed to target one specific domain of EFs only.

[1]Institute of Neuroscience and Medicine (INM-7 Brain and Behaviour), Research Center Jülich, Jülich, Germany. [2]Institute of Systems Neuroscience, Heinrich-Heine University, Düsseldorf, Germany. [3]Institute of Neuroscience and Medicine (INM-1 Structural and functional organization of the brain), Research Center Jülich, Jülich, Germany. [4]Department of Psychiatry, Psychotherapy und Psychosomatics, Medical Faculty, RWTH Aachen University, Aachen, Germany. ✉e-mail: j.amunts@fz-juelich.de

Beside subdomain-specific EF tests, clinical and research test batteries also include a speech-based task, namely the verbal fluency (VF) task but the explicit involvement of different EF subdomains in the VF task is reported controversially especially when considering inter-individual differences[19]. The well-established and often-used VF test assesses the number of words generated in a given time (usually 60 seconds) and has been found to be a sensitive measurement for testing EFs in both non-clinical groups[19] as well as in neurological patients[20]. VF tests mainly comprise two types of tasks: The phonemic/lexical VF, requiring the generation of as many words as possible beginning with a specific letter (e.g. C, F); and semantic VF, in which the examinee is asked to produce words that belong to a specific semantic category (e.g. fruits, animals). Additionally, most VF tests include a switching task in which words from two different categories are produced in an alternating order[21].

The relationship between VF and the various subdomains of EFs has frequently been investigated in both healthy controls[19,22–24] and patients[17,25]. Concerning the relationship between VF and working memory, some studies showed that better working memory performance leads to less perseveration errors[26] and a higher total score of produced words in the VF task[22,27]. However, a clear link between working memory and VF performance has so far not been found[19,28]. Similarly, the relationship between VF and response inhibition is not clear yet. While some studies report lower scores in VF concomitating with a decline of inhibition performance[22] other studies failed to find a link between VF and inhibition[20,29]. Finally, regarding the relationship between VF and cognitive flexibility, studies report a positive correlation of switching between categories in VF tasks and cognitive flexibility performance[21]. However, there are also findings which indicate that there is no relationship between EFs and VF performance[30].

Altogether, results concerning the relationship of EFs and VF are ambiguous. This might, at least partly, be based on inter-individual variability of both EFs and VF. For example, a pronounced effect of age was identified by multiple studies showing a significant negative correlation between age and the different aspects of EFs as well as VF performance[31–34]. Furthermore, fluid intelligence has been found to be related to the performance in EFs tasks tapping into the subdomains of planning and reasoning and[35]. In contrast, inhibition was shown to be independent of intelligence in children with problems performing attention tests[36].

The complex involvement of EFs in VF performance has also been shown to be modulated by the influence of inter-individual variability like dynamically varying hormonal levels. Especially sex hormones like estradiol and progesterone have been shown to influence performance in EFs tasks[37,38]. It was shown that cognitive performance varies during the different phases of the menstrual cycle with high progesterone and estradiol levels leading to faster reaction times and better accuracy[37,39]. Moreover, cortisol, which is mostly associated with stress, appears to impact EFs, but literature addressing this topic is ambiguous. On the one hand, studies found a positive relationship e.g. between cortisol level and working memory[40] or cortisol level and performance in cognitive flexibility tasks[41] in men. On the other hand an inverse relationship was found in cognitive flexibility tasks in women[41] and in working memory performance[42]. Additionally to the influence on EFs varying hormonal levels could be also linked to VF performance[43]. To investigate the role of varying hormonal levels studies implement different procedures: While some inject specific hormones and assess the change of cognitive functions due to this injection[44,45] other studies analyze intra-individual differences measuring the naturally varying hormonal level at different points of time[37,40]. Test data of EFs and VF are commonly analyzed with classical statistical methods. For example, correlation analyses have been previously used to investigate the relationship of VF and specific subdomains of EFs[46]. Other studies have investigated group differences between patients and healthy controls to e.g. examine sex differences in VF strategies[19] or to explain the relationship of memory and VF in patients with Alzheimer's disease[28]. Furthermore, factor analysis has also been applied to investigate common cognitive structures of VF performance[24,30].

Considering previous literature investigating the relationship of EFs and VF in more detail it is obvious that each work contributes to a better understanding of this relationship but generalizing this knowledge is still difficult. Specifically, these limitations are e.g. due to the small subject size or reduced EF test batteries which does not represent overall EF performance. Generalizability is also restricted due to the applied methods. All the above-mentioned methods are applied to investigate within-sample effects to understand the theoretical hypothesis-driven neuropsychological relationship between VF performance and EFs. However, it is so far unclear to what extend VF task performance reflects the different subdomains of EFs. To address this question, more advanced statistical methods should lead to a more detailed insight into the complexity of VF performance. Machine learning models can be used to characterize complex behavior with the ultimate goal of identifying and predicting psychiatric diseases[47,48]. In contrast to classical statistical analyses, these prediction analyses use large sample sizes and a high number of variables as well as a cross-validation approach by training a model on part of the dataset and then validating it on unseen data. Applying machine learning methods on a wide variety of EF tests enables to capture the complex and non-linear relationship of EFs and VF performance.

To contribute to a deeper understanding of the so far inconclusive relationship between EFs and VF, the present study used a machine learning approach to investigate to what extent VF performance can be explained by subdomain-specific EF tests. We hypothesize that VF performance can be explained by a conglomeration of cognitive flexibility, working memory and inhibition test scores, which is further modulated by individual variations of fluctuating hormone levels.

Methods

Participants. The age of the 253 healthy participants was ranging from 20–55 years (mean age 35.3 ± 11.0, 99 males). Participants were monolingual German speakers and received different levels of education (finished middle school: 10, professional school/job training: 70, finished high school with a university-entrance diploma: 76, university degree: 97). Participants were recruited in North Rhine-Westphalia (Germany) via social networks and the Forschungszentrum Jülich mailing list. Testing sessions took place at the Forschungszentrum Jülich, with

Measure	Description	Main variables
Cognitive flexibility/Planning		
Trail-Making test	The task consists of 2 parts. In part A, numbers from 1–25 are displayed on the screen in a haphazard fashion. The task consists of clicking on the numbers in sequential order as quickly as possible. In part B numbers from 1–13 and letters from A-L are presented on the screen. participnts must click on the numbers and letters alternately and in ascending order.	Errors in part A/B, difference part B-A, quotient B/A
Raven's Standard Progressive Matrices	Eight items that form one pattern are shown to the participants. The task requires the participants to identify one missing item out of 6 choices to complete the pattern. The difficulty of recognizing each pattern increases during the course of this test.	Process time, correct items
Wisconsin Card Sorting test	Four stimulus cards illustrating different geometrical figures are presented. These cards differ in the number, color and form of the figures. The task is to match one additional card to one of the four cards using the correct rule (match for number, colour or form of figure) without knowing which rule is applied.Thus, participants are required to shift rules accordingly.	Number of perseveration/ non-perseveration errors
Tower of London	Three rods are presented on the screen: The left rod holds three balls, the middle rod two balls and the right rod one ball. The participants are asked to move the balls from the starting state to ta target position using a minumum number of moves.	Planning ability, number of correct respones
Cued task switching	A coloured figure is presented on the screen. Participants are required to respond to either the color or figure task. Figure task: Particpants press matching button (left or right) depending on the type of the figure (triangle or rectangle); colour task: Particpants press matching button (left or right) depending on the colour of the figure (blue or yellow).	Number of incongruent/ congruent errors
Working memory/Attention		
N-back non verbal	A sequence of 100 abstract successive figures are presented to the particpants. The task consists of indicating whether the current stimulus matches the figure shown two turns back (2-back paradigm).	Number of correct and false responses
Non-verbal learning test	Nonsensical, irregular, and geometric figures are presented on the screen. In the course of the test some figures are shown multiple times. For each figure the participants has to decide whether the current figure has already appeared or whether this figure is being shown for the first time.	Correct/false responses, sum of difference between correct minus false responses, process time
Corsi block tapping test	Nine irregularly arranged cubes are presented to the participants. A cursor touches a certain number of cubes in a specific order; The task is to repeat the given sequence correctly. The length of the sequence increases the more correct sequences the particpants complete.	Block span, correct/false items, error types (omission, sequence mistake)
WAF-G (divided attention)	The participants are required to focus on two geometric figures and one auditory stimulus. At a certain interval the stimuli change their intensitiy (figure gets lighter and/ or auditory stimulus gets higher). The participants have to respond when two stimuli become lighter/higher twice in a row.	Mean reaction time, number of false alarm, missed items
WAF-R (spatial attention)	Four triangles are presented in four spatial positions (similar to Posner paradigm). The participants are required to react if a triangle changes intensity (gets darker). In the neglect test a interfering/matching visual cue is given but this cue do not always indicate the correct answer.	Mean reaction time, number of false alarm, missed items
Inhibition		
Stop-signal task	The test consists of two parts: 1) The participants are asked to respond to the direction of an arrow stimulus. 2) The participants have to repeat task as in previous step but should withhold their motoric response whenever they hear an auditory signal.	Stop-signal reaction time, stop-signal delay, number of different error types
Simon task	The participants are asked to press the right button if they read the word "right" and the left button if they read the word "left". The words are either presented on the right or left part of the screen. The reaction time of the participants is usually longer whenever the stimlus is incongruent to its position (e.g. the word "left" is on the right side of the screen).	Interference reaction time, incompatible/compatible errors
Stroop test	Names of colors (e.g., "blue", "green", or "red") are displayed on the screen in a color which is not denoted by the name (i.e., the word "blue" is printed in red). The test consists of two conditions: 1) Naming - participants are asked to respond to the colour of the words; 2) Reading - participants are asked to respond to the meaning of the word with naming. A baseline measure is taken at the start of the test to assess reading and color naming (color and word refer to the same concept).	Baseline time of naming and reading, reading interference, naming interference, errors

Table 1. Overview of executive function test battery.

a duration of 150–180 minutes depending on the time needed for instructions and the speed with which the participants passed the tests. A remuneration fee of €50 was paid.

Data collection. Data was collected by four different examiners, all of whom conducted several pilot testings and were instructed by the study leader to ensure a common standard. The examiner gave standardized instructions before starting each test and help was provided by the examiner whenever the participant had any questions regarding the instructions or tests. The testing session included 13 EF tests and 3 semantic VF tasks. The EF test battery consisted of computerized versions of neuropsychological tests covering domains of inhibition, working memory and cognitive flexibility. Ten of these tests were taken from the *Vienna Testsystem* test battery and three were designed with *PsyToolkit*[49]. In this study, we assessed commonly used EF tests like the *Stroop* and *TMT*. We used a broad selection of EF tests to cover all subdomains of EFs and to detect most influencing tests and their variables. A complete list of the tests is shown in Table 1.

Results of the neuropsychological tests can be seen in Table 2.

The semantic VF tasks were based on the *Regensburger Wortflüssigkeitstest*[50]. This test is a standardized neuropsychological assessment that has been thoroughly tested for reliability, validity and objectivity[50]. Due to

	Variable	M ± SD	Min - Max
Age		35.33 ± 11.04	20–55
Education		4.05 ± 0.90	2–5
Cortisol		0.12 ± 0.08	0.001–0.42
Estradiol		3.61 ± 5.27	0.01–44.7
Progesterone		65.09 ± 93.62	6.25–940.97
Testosterone		79.96 ± 99.5	2.41–597.61
Trail Making Test	Difference part A-B [sec]	7.60 ± 6.25	−3.32–40.57
Raven's Progressive Matrices	Correct items	27.86 ± 3.3	14–32
Wisconsin Card Sorting Test	Perseveration errors	7.91 ± 3.45	4–24
Tower of London	Planning ability	7.51 ± 2.20	1–12
Cued Task-Switching	Switch costs (reaction time switch tasks - reaction time in non-switch tasks)	0.05 ± 0.08	−0.15–0.39
N-back nonverbal	Correct items	8.40 ± 2.94	1–14
Non-verbal learning Test	Sum of difference between correct minus false	19.54 ± 7.78	−4–35
Corsi Block Tapping Test	Block span	5.68 ± 1.10	3–9
WAF-G (divided attention)	False alarm (crossmodal)	3.10 ± 4.86	0–34
WAF-R (spatial attention)	Errors	3.61 ± 3.38	0–18
Stop-Signal Task	Stop signal reaction time (mean reaction time - mean stop signal delay) [sec]	0.21 ± 0.07	0.03–0.50
Simon Task	Reaction time difference (reaction time incongruent - reaction time congruent items) [sec]	0.03 ± 0.04	−0.14–0.16
Stroop Test	Reading interference [sec]	0.14 ± 0.08	−0.04–0.50
Stroop Test	Naming interference [sec]	0.13 ± 0.08	−0.02–0.46
Semantic Verbal Fluency sum1		36.77 ± 8.30	19–57
Semantic Verbal Fluency sum2		26.08 ± 6.60	11–45
Semantic Verbal Fluency sum3		21.98 ± 4.34	8–34
Semantic Verbal Fluency sum all	sum1 + sum2 + sum3	84.83 ± 15.45	50–125

Table 2. Neuropsychological data of participants.

language-specific differences in the frequency and usage of letters and categories[51] this German version of VF task was used. Two of the tasks were simple semantic VF tasks in which the participant had to name animals (t_1) and jobs (t_2). The third semantic VF task was a switching task in which the participant switched between fruits and sports (t_3) within the same task. Each of the three tasks was performed for 2 minutes. The VF tasks were presented with *Presentation* software (*Neurobehavioural Systems*) and the participant's responses were recorded automatically. Following the testing session, the recorded speech was transcribed and words were coded manually as being either correct answers or errors. The number of correct words were counted for each task (t_1, t_2, t_3) and the sum score of total number of correct words across all three VF tasks was used in all further analyses. To broadly represent VF performance, the sum of all VF tasks was selected to include different aspects of the task. This variety of VF performance is beneficial to build a machine learning model which is complex enough to reflect the complex patterns of VF performance.

In addition to the main test set of EFs and VF tasks, phenotypical data was collected through questionnaires to gather information regarding the physical and psychological well-being of the participants. These questionnaires included the Beck Depression Inventory (BDI-II) (Beck, Steer & Brown, 1996) which was used to collect information regarding depressive symptoms. Saliva samples were collected at the beginning and at the end of the test session. The two saliva samples of each subject were sent to an external lab which pooled both samples before carrying out analysis for cortisol, progesterone, estradiol and testosterone. Additionally, the testing session also comprised further speech tests (word-picture interference task, picture description, spontaneous speech), for which results will not be reported here, as they will be independently analyzed. This additional data will then be described in a subsequent paper. Moreover, we aim to publish a data paper which will describe all aspects of data collection, test selection and testing procedure in detail while also making this data publicly available.

Collection and analyses of the data presented here was approved by the ethics committee of the Heinrich-Heine University Düsseldorf. We confirm that all experiments were performed in accordance with relevant guidelines and regulations. Moreover, informed consent was obtained from all participants.

Data analysis. The original dataset of 253 participants was reduced to 235 due to missing data of some participants (94 males; 101 participants were aged between 20–31, 70 between 32–43, 64 between 44–55). From all EF tests 72 variables (Supplement 1) were extracted based on the features provided by the *Vienna Testsystem* and *PsyToolkit*[49]. VF performance was represented by the sum score of correct words across all VF tasks.

Two independent analyses were computed. In a first analysis, Spearman correlations were computed to analyze the relationship of each EF variable and VF sum scores. Here, a reduction of the 70 EF variables was used. Specifically, EF variables were selected based on the EF test manuals provided by the *Vienna Testsystem* in 10/13

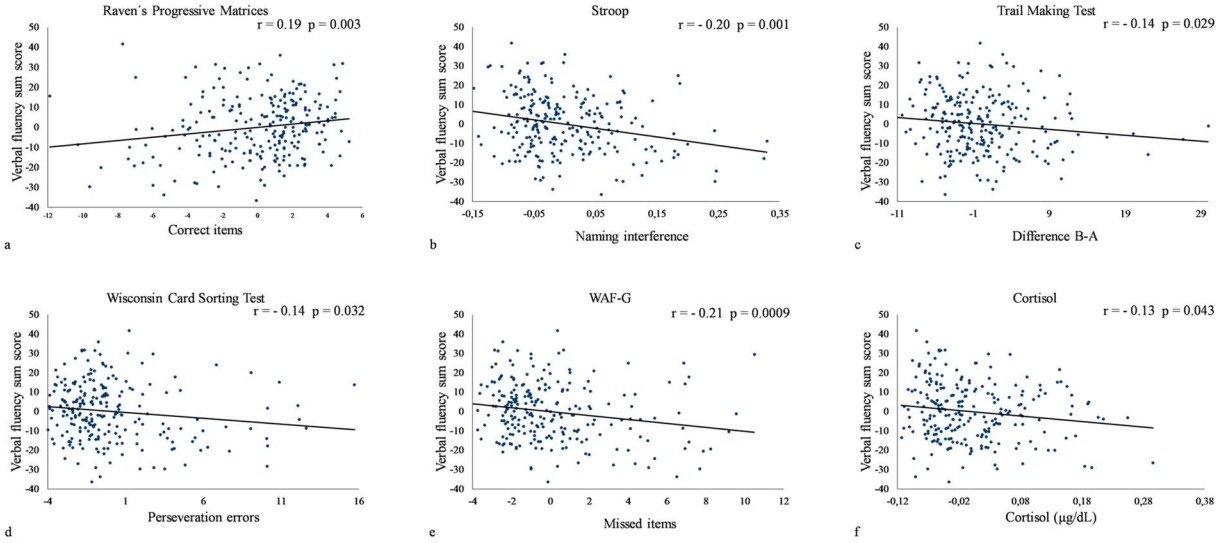

Figure 1. Plots of significant correlations of executive function tests and total verbal fluency sum score. The performance in verbal fluency task is represented by the total number of correct words produced across all three semantic VF tasks ($t_1 + t_2 + t_3$).The negative correlation in plot b-f are due to the divergent direction of the scores since these variables describe different types of errors, reaction or process times (the higher the worse the performance) while the performance in the verbal fluency is represented by the total amount of correct items (the higher the better).

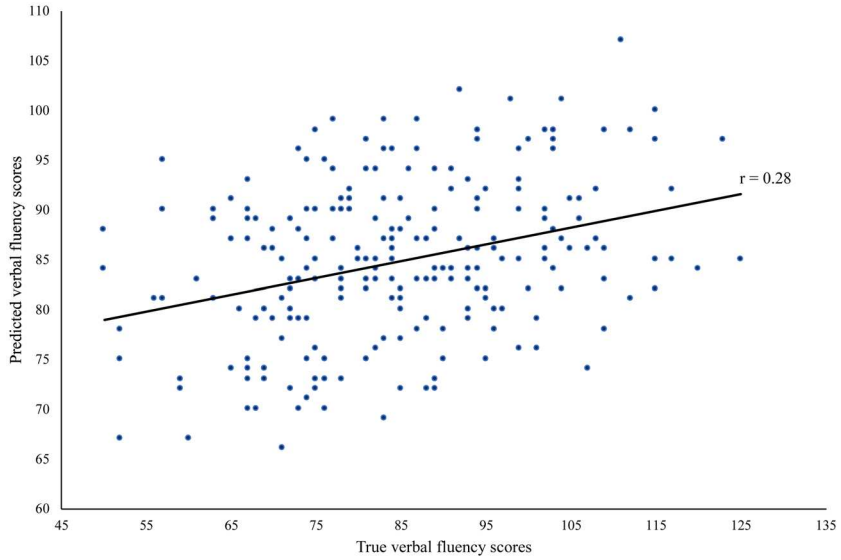

Figure 2. Correlation of true and predicted verbal fluency sum scores applying Relevance Vector Machine algorithm.

EF tests. In cases where multiple main variables were provided by the *Vienna Testsystem*, the main variable was selected based on previous literature investigating EF performance. In contrast to the *Vienna Testsystem*, tests run within *Psytoolkit* are not standardized and thus do not come with associated test manuals. Thus, the selection of main variables of tests designed with *Psytoolkit* (3/13) were selected based on previous literature.

Considering the influence of sex and age on the performance in EF and VF tasks[19,22,34] data were adjusted for these variables by linear regression and analyses were computed with the residuals.

In a second analysis, the possibility of predicting VF from EF scores was investigated by applying supervised learning via a sparse (relevance vector machine; RVM) and non-sparse (partial least squares; PLS) model using 72 EF variables (Supplement 1). Generally speaking, sparse models aim to reveal a sparse structure and detect correlations among redundant features[52]. Specifically, RVM is based on the Support Vector Machine (SVM) but is a Bayesian sparse technique which allows for the prediction of a specific target value from a set of different features. In contrast, PLS is similar to principal components regression and is based on covariance. Results given in the

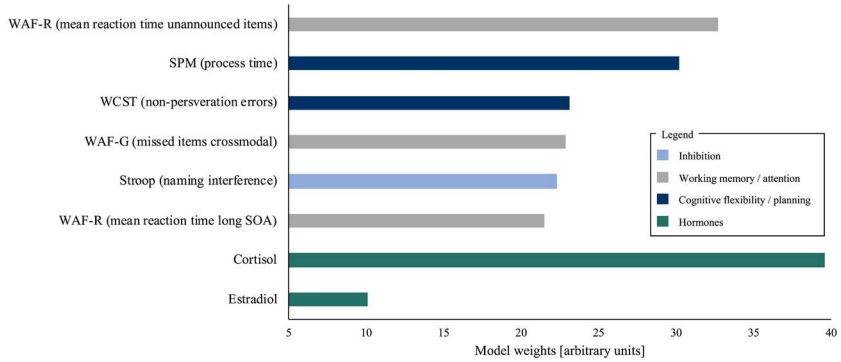

Figure 3. Features displaying strongest impact on prediction analysis.

main manuscript focus on the RVM approach, while results for the PLS analysis are given in the supplement. Sex and age were regressed out from VF score and from EF data in a cross-validation consistent way.

Before running the prediction analysis, data was transformed to z-scores. A 10-fold cross-validation was then performed for which the data set was randomly split into 10 sets, 9 of which were used for training while the 10th set was held back and used to perform the prediction in previously unseen data. Ten replications of the 10-fold cross-validation were performed and thus 100 prediction models were computed. Prediction performance was assessed by computing the correlation between real and predicted values.

Beside testing statistical significance of prediction performance, we also examined which specific EF features significantly impact prediction performance. To determine which EF features (EF test variables) contribute most strongly to the prediction, we employed an approximate permutation test procedure, in which associations between features (total set of EF variables) and labels (VF sum score of each participant) were randomized. That is, the VF performance score was randomly permuted while the feature matrix was kept unchanged. The RVM analysis was repeated for each permutation and accuracies for 100 permutations were used to construct an empirical null distribution for each feature, which was used to compute the statistical significance of the contribution of each feature as the proportion of permutated labels achieving a better prediction than then original labels.

Results

Correlations between executive function scores and verbal fluency performance. The correlation analyses identified multiple significant results which are shown in Fig. 1.

The highest negative correlation coefficient can be seen between the number of missed items in *WAF-G* and the VF performance ($r = -0.21$; $p = 0.0009$) indicating that a better performance in divided attention is associated with a higher VF score. Likewise, inhibition ability measured by the naming interference variable of the *Stroop* test ($r = -0.20$; $p = 0.001$) shows a negative correlation with the VF score. This result indicates that participants who successfully inhibited proponent behavior in the *Stroop* task perform better in the VF task. Additionally, abstract reasoning assessed with the *Raven's Progressive Matrices* test (SPM) reveals a positive correlation ($r = 0.19$; $p = 0.003$) to VF performance indicating a demand of cognitive flexibility and planning while generating words from a specific category. Similar results were found for the *TMT* ($r = -0.14$; $p = 0.029$) and the number of perseveration errors in the *WCST* ($r = -0.14$; $p = 0.032$) which particularly reflect the involvement of cognitive flexibility and working memory in the VF task. Additional to the EF battery we also found a significant negative correlation of the VF tasks and the Cortisol level of the subjects ($r = -0.13$; $p = 0.042$).

Prediction of verbal fluency performance from EF scores. The correlation of true and predicted values was $r = 0.28$ ($p < 0.0001$) (Fig. 2).

In order to quantify the contribution of the different EFs variables to VF performance, features with significant model weights in the approximate permutation test were identified. As can be seen in Fig. 3, 8 features belonging to 4 different EF tests and 2 hormones were identified.

The EF feature with the highest impact on the prediction analysis is "mean reaction times" of unannounced items of the spatial attention test *WAF-R*. The RVM analysis also revealed another "reaction time" feature of *WAF-R* which represents "reaction time" in items with a long stimulus onset asynchrony. The influence of attention on VF performance was also shown in a feature of *WAF-G* assessing the number of missed items in a divided attention test. These results show that participants reacting faster in attention tests also perform better in the VF task, identifying overall reaction speed and correctness as a central component in VF performance. Since *WAF-R* is not only assessing attention but also includes inhibitory requirements, these results highlight the role of attention and inhibition during VF performance. The explicit role of inhibition can be also detected in the variable "naming interference" of the *Stroop* test indicating that inhibition is an essential component to successfully produce words within or between two different categories. The analysis also revealed the predictive meaningfulness of cognitive flexibility and planning, by showing that "non-perseveration errors" in *WCST* and "process time" in *SPM* contribute essentially to the prediction analysis. The *WAF-R* was the only test presenting more than one variable represented in the most predictive features, both of which contain reaction time information. With regards to non-EF features, the RVM analysis also identified stress hormone cortisol and sex hormone estradiol as highly predictive variables (Fig. 3).

Corresponding results of the PLS analysis revealed a correlation of true and predicted values of $r = 0.35$ ($p < 0.0001$). However, in contrast to the results of the RVM analysis, approximate permutation test did not reveal any significant *p*-values identifying specific EF features. Detailed results of the PLS analysis are given in the Supplementary Material (Supplement 2).

Discussion

The aim of the study was to elucidate to what extent VF performance can be explained by different subdomains of EFs and which types of EF variables contribute most strongly to the prediction of VF performance. In a first step, we correlated the different EF scores with the number of correctly produced words across the three semantic VF tasks. This analysis revealed significant correlations between *SPM*, *Stroop*, *TMT*, *WCST*, *WAF-G*, *WAF-R* and the VF task performance. These EF tests tap into two EF domains, namely cognitive flexibility and inhibition. We further investigated the relationship of EF scores and VF by prediction analyses to gain insight into the contribution of the different EF test variables. We showed that EF data predict VF performance and that beside cognitive flexibility and inhibition, reaction time and attention play important roles in predicting VF performance. Additionally, hormonal influences were identified as meaningful parameters to predict VF performance, highlighting the influence of inter-individual differences in VF performance. We first discuss the results in the direct context of the different EF subdomains cognitive flexibility, inhibition and working memory. Secondly, the involvement of EFs in the VF task is discussed in a more general context addressing the role of attention as well as the meaningfulness of reaction times. Finally, the influence of varying hormonal levels illustrates the impact of inter-individual differences on VF performance.

Multiple tests within the domain of cognitive flexibility were shown to be related to VF performance. The highest correlation was found for the *SPM* test followed by *TMT* and *WCST*. While the correlation analysis revealed a relationship of these tests with VF performance, the prediction analysis confirms the importance of the features describing errors in *WCST*. Additionally, the prediction analysis highlights the component of processing speed during *SPM* which was not identified by correlation analysis. In congruence with these results, previous studies have linked VF with cognitive flexibility[21,53,54]. Paula *et al.*[21] investigated this relationship in healthy adults, using simple and switching semantic VF tasks and three different EF tests, including the *TMT*. They found that this particular measure of cognitive flexibility correlated well with both simple and switching VF tasks. The influence of cognitive flexibility on VF was also examined in a study by Troyer *et al.*[54] who discussed the importance of cognitive flexibility assuming that two different abilities are needed for VF: (1) verbal memory for the creation of clusters and production of words belonging to a specific subcategory; (2) strategic search and cognitive flexibility which enables shifting between clusters[54].

It should be noted that the present results concerning the *SPM* might have to be treated with caution since this test also encompasses aspects of fluid intelligence[55–57]. Due to the relationship between EFs and fluid intelligence[58,59], it may not surprise that fluid intelligence also impacts VF performance as has been shown in studies on schizophrenia[60] and bipolar disorder patients[61] as well as healthy controls[60].

Altogether, considering that three out of five cognitive flexibility tests contribute to VF performance, the present results point to a crucial influence of cognitive flexibility on VF performance, especially to cluster words and switch between categories.

In addition to the domain of cognitive flexibility, inhibition tests were also identified to play a role in VF performance. Specifically, both the correlation analyses and prediction analysis revealed that the naming interference of the *Stroop* test is related to VF. Previous studies report ambiguous results when investigating the relationship between inhibition and speech production. For example, a positive correlation between inhibition, assessed with a stop-signal task, and the reaction time in picture naming[46] was found but could not be validated in VF tasks[23]. Discussing these ambiguous results, the authors suggest that while stop-signal tasks measure the participant's ability stopping a planned response (*response inhibition*), VF tasks tend to involve the ability of suppressing the activation of competitive target responses (*selective inhibition)*. In the present study selective inhibition was assessed by the *Stroop* test. In the naming subtask of the *Stroop* test the participant is asked to name the color in which the word is printed. Incongruent items, in which the color of the word does not match the written word evoke a longer reaction time, indicating that prepotent responses (i.e. the meaning of the written word) have to be suppressed. This is very similar to the kind of inhibitions that participants are challenged with in the VF task when needing to suppress words which have already been produced. In accordance with previous literature[24], the present suggests that selective inhibition, specifically as reflected in the *naming interference* of the *Stroop* test, is a key parameter to drive VF performance. An alternative interpretation of the *naming interference* of the *Stroop* test and the total number of words produced in the VF task relates to the association between verbal processing speed and dominance of word reading. Individuals with a high verbal processing speed can be assumed to also show a stronger dominance of word reading then those with slower verbal processing. This stronger dominance of word reading, in turn, can be expected to go along with a stronger interference effect in the *STROOP* task, thus explaining the correlation between naming inference in the *STROOP* task and VF performance.

In addition to cognitive flexibility, working memory and inhibition we also investigated the role of attention in VF performance. While divided attention (*WAF-G*) was linked to VF performance in the correlation analysis, multiple variables of the spatial attention test (*WAF-R*) and divided attention test (*WAF-G*) contributed to VF performance in the prediction analysis.

Previous studies also described attention as a crucial cognitive function to perform VF[54]. In particular, it could be shown that divided attention particularly impacts the switching component of the VF tasks. At first glance, the influence of the spatial attention test on the speech task might be surprising since there are no spatial requirements in the VF task. Nevertheless, we assume that beside the component of attention *per se* the involvement of inhibition which is also part of this task might also have an effect on these results. With respect to the relationship

of VF and divided attention, results are consistent with previous literature highlighting the influence of divided attention especially in the VF switching task[54]. To sum up, attention might be a crucial aspect for performing VF task. In particular, we hypothesize that attention is a fundamental and permanent cognitive requirement during updating the current status of already produced words as well as being efficient in producing words within or between two.

Surprisingly, the present study did not find a relationship between working memory and VF performance in both the correlation and prediction analysis. However, previous studies investigating the involvement of working memory in VF tasks also report ambiguous results: For example, one study assessing a digit-span and spatial-span test did not find a significant relationship between working memory and VF[28]. However, other studies have reported results that indicate updating of information and working memory performance have a high impact on VF scores[23,62,63]. Additionally, another study found the relationship between memory performance and VF to be specific to women[19]. The missing link between working memory and VF in the current study might be the type of variable which was selected to represent VF performance. Here, the sum of correctly produced words was used as the main measure of VF performance. While measuring VF performance, this variable does not contain information about the types of errors that occur during the VF task. Although word repetitions (*perseveration errors*) did not count as correct produced words this error type is not analyzed separately. However, perseveration errors are described as a sensitive indicator of working memory performance. Additionally, the relationship between working memory and VF performance measured with the total sum of words has so far been mainly investigated in patients or older participants[22,27] but rarely in healthy controls. Thus, we assume that the number of correct produced words might be less meaningful to reflect working memory in healthy controls than error-specific parameters like perseveration errors.

In general, prediction results reveal specific EF tests and variables which are closely linked to VF performance. Besides differences in the strength of the relationship between certain EFs and VF, the EF test constructs themselves might also have partially influenced analysis. In particular, the reliability of some EF tests is discussed controversially[64,65]. Thus, some EF tests might not represent actual EF performance well and such a poor reliability of EF test might be reflected in the relationship of EF to VF as studied here.

Comparing correlation and prediction analyses, crucial differences in the results were observed. At first glance, the EF tests identified in the prediction analyses are similar to those of the correlation analyses but include additional variables. Specifically, the prediction analyses reveal a number of additional variables that measure how fast participants completed the tests and how many errors they made. While there is limited literature about the influence of processing speed in cognitive tasks on VF performance, some studies have addressed processing speed in general in the context of cognitive functions[66,67]. Another study found relationships between processing speed, working memory, inhibition and VF scores[24]. Additionally, poorer processing time has been associated with poorer cognitive performance in older adults[66]. The association between processing speed and EFs has been also observed in patients with depression[67]. In line with these findings, another study investigated the role of processing speed in schizophrenia patients and suggest that especially in working memory tasks assessing speed might be helpful to detect patterns of schizophrenia[68]. In respect to VF performance, processing speed was identified as being closely related to speech production[31] and is reported as a predictor for VF in ageing[69]. Based on previous literature and our present findings, we assume that processing speed is a general aspect involved in both EFs tests and VF tasks. Particularly, it can be assumed that due to the time limit of 2 minutes in the VF tasks participants are zealous to name as many words as possible. This general behavior might also be relevant in EF tests. Thus, we suggest that processing speed and reaction times indicate that people acting fast in cognitive tasks also perform more successfully in VF tasks than participants thinking more in detail about their answer. Additionally, we assume that the complex influence of processing speed on VF performance might be beyond what can be described as a linear relationship. This might explain why the impact of speed is detectable in the prediction computation but was not found the correlation analysis.

In addition to the relationship between EFs and VF this study also assessed the influence of hormonal fluctuations to investigate the influence of inter-individual differences. The results showed that the stress hormone cortisol and the sex hormone estradiol have a high impact on VF performance. In line with other studies[44,45] our analyses indicates that there is a negative correlation between cortisol level and the performance in cognitive functions. However, previous studies have also linked an increase of cortisol to better cognitive performance[40,41]. To our knowledge, rather little is known about the influence of estradiol on VF performance. However, studies investigating the influence of estradiol on EFs show that higher estradiol levels particularly leads to better performance in shifting and cognitive flexibility tasks[37,39]. Moreover, a link between hormonal contraceptives and VF performance[43] has been shown. These results demonstrate that women taking hormonal contraception and consequently having significantly lower estradiol and progesterone levels, perform worse in the VF task than the control group[43]. The high impact of cortisol and estradiol in the prediction analysis suggest that fluctuating hormones are essential parameters for predicting VF performance and that intra-individual differences in hormone levels need to be considered when examining the relationship of cognitive functions and speech production tasks. Thus, it shows that although EF test variables are closely related to VF performance VF is a complex construct which is also driven by hormones and attention.

Our prediction analyses yielded important insights into the relationships between EFs, VF and inter-individual differences. However, some open questions remain concerning both inter-individuality and speech related topics. Firstly, due to the fact that inter-individuality influences both EFs and VF performance, further studies would benefit from gathering additional inter-individual parameters. For example, a test for assessing intelligence might be useful to control for the influence of intelligence on each test, especially on the *SPM*, making it possible to better differentiate the impact of cognitive flexibility on VF performance. Secondly, intra-individual differences could be further investigated by gathering saliva samples at two different time points. In this study saliva samples of each participant were pooled. A comparison of the hormones level before and after testing might help to

provide insights into individual strategies dealing with stress. Beside inter-individual influences of hormonal levels on EFs[41], studies also report intra-individual variety e.g. due to different phases of the menstrual cycle[38]. Therefore, an analysis of hormonal levels within each participant taken at different time points could reveal an additional dimension representing intra-individual differences.

Considering speech-specific issues, a vocabulary test could contribute to better understand inter-individual differences. Previous studies showed that the vocabulary size has a positive impact on VF performance[54,70]. Moreover, additional parameters reflecting VF performance could help to gain deeper insights of searching strategies during VF tasks. In particular, semantic analyses provide details of clustering and switching[24] and could indicate the participant's strategies which could then be linked to EF performance.

A more general consideration is related to the predictive methods as used in this study. An independent data set assessing the same variables that were used in the study does not yet exist. Thus, it was not possible to validate our results in a totally independent dataset. Instead, we applied 10-fold cross-validation by repeatedly training the model on parts of the data while keeping a subset out as a validation sample. However, we are aware of the need to validate our results in an independent dataset to better generalize our results and suggest a replication of this study on an independent sample which could prevent study-specific biases. However, due to the broad and specific collection of the EF test battery finding a similar data set could be difficult. An additional independent dataset with similar EF tests could be used to test split-half reliability investigating the construct of EF tests. Due to the high number of participants which is needed to apply machine learning methods it was not possible to split our data in two groups and running the prediction analysis on the split data. The ambiguous results of the RVM and PLS analysis also need to be considered. While both approaches revealed a significant correlation between true and predicted values, the PLS approach did not identify any significant features (Supplement 2). This might be due to the fact that PLS is a non-sparse machine learning method, which will include all features in the prediction model. In contrast, it is the nature of sparse models like RVM to build the prediction model based on most relevant features only.

Due to the high number of participants and the large battery of EF tests this study provides a detailed view on the involvement of EFs in VF tasks and examines the influence of fluctuating hormones. It investigated to what extent EF tests can represent semantic VF performance and shows that cognitive flexibility and inhibition are the main domains involved in performance on the VF task. Additionally, attention seems to be a central component of the VF task. The most striking observation to emerge from the data analysis was the new and more detailed view of the EF tests and variables that are best at predicting VF performance. While correlation analyses provided first insights into the relationship of EFs and VF, the prediction analyses revealed the importance of speed parameters. In particular, our results suggest that beside the influence of specific EFs, more general components such as attention and speed are crucial aspects of successful VF performance. These results also highlight the advantage of the prediction analysis since it revealed concrete variables of EF tests which also represents cognitive abilities not directly linked to specific EF subdomains or representing standard variables.

A better understanding of the cognitive demands that are required for the successful performance of VF tasks can potentially lead to a more wide-spread use of VF tests in the clinical context, thus EF tests that tend to be time-consuming and inaccurate. Additionally, VF tests tend to better reflect real-life conditions than lab-based EF batteries. A detailed knowledge of meaningful test variables could later on lead to insights into which subdomains of EFs could be replaced by VF tasks and which subdomains of EFs still have to be assessed by additional EF tests. This link between EF and VF represents a first step towards a speech-based EF-test. Furthermore, it indicates that in investigating the relationship of EF and VF the complex construct of VF performance should be considered in research and clinical context.

Furthermore, taking the influence of varying hormonal levels into account our study suggests that beside inter-individual differences intra-individual fluctuations could play an important role in evaluating VF performance in clinical context.

Received: 8 January 2020; Accepted: 2 May 2020;
Published online: 07 July 2020

References

1. Diamond, A. Executive functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
2. Karr, J. E. *et al*. The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychol. Bull.* **144**, 1147–1185 (2018).
3. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
4. Nigg, J. T., Blaskey, L. G., Huang-pollock, C. L. & Rappley, M. D. Neuropsychological Executive Functions and DSM-IV ADHD Subtypes. *J. Am. Acad. Child Adolesc. Psychiatry* **41**, 59–66 (2002).
5. Kudlicka, A., Clare, L. & Hindle, J. V. Executive functions in Parkinson's disease: Systematic review and meta-analysis. *Mov. Disord.* **26**, 2305–2315 (2011).
6. Tavares, J. V. T. *et al*. Distinct profiles of neurocognitive function in unmedicated unipolar depression and bipolar II depression. *Biol. Psychiatry* **62**, 917–924 (2007).
7. Barch, D. M. The cognitive neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.* **1**, 321–353 (2005).
8. Wu, T. & Hallett, M. Neural correlates of dual task performance in patients with Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **79**, 760–766 (2008).
9. Altmann, L. J. P. & Troche, M. S. High-Level Language Production in Parkinson's Disease: A Review. *Park. Dis*. **2011** (2011).
10. Bowden, S. C. *et al*. The Reliability and Internal Validity of the Wisconsin Card Sorting Test. *Neuropsychol. Rehabil.* **8**, 243–254 (1998).
11. Erdodi, L. A. *et al*. The Stroop test as a measure of performance validity in adults clinically referred for neuropsychological assessment. *Psychol. Assess.* **30**, 755–766 (2018).
12. Sánchez-Cubillo, I. *et al*. Construct validity of the Trail Making Test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc. JINS* **15**, 438–450 (2009).

13. Kortte, K. B., Horner, M. D. & Windham, W. K. The trail making test, part B: cognitive flexibility or ability to maintain set? *Appl. Neuropsychol.* **9**, 106–109 (2002).
14. Miyake, A. *et al*. The unity and diversity of executive functions and their contributions to complex 'Frontal Lobe' tasks: a latent variable analysis. *Cognit. Psychol.* **41**, 49–100 (2000).
15. Guariglia, C. C. Spatial working memory in Alzheimer's disease: A study using the Corsi block-tapping test. *Dement. Neuropsychol.* **1**, 392–395 (2007).
16. Jacola, L. M. *et al*. Clinical utility of the N-back task in functional neuroimaging studies of working memory. *J. Clin. Exp. Neuropsychol.* **36**, 875–886 (2014).
17. van den Berg, E., Jiskoot, L. C., Grosveld, M. J. H., van Swieten, J. C. & Papma, J. M. Qualitative Assessment of Verbal Fluency Performance in Frontotemporal Dementia. *Dement. Geriatr. Cogn. Disord.* **44**, 35–44 (2017).
18. Alvarez, J. A. & Emory, E. Executive Function and the Frontal Lobes: A Meta-Analytic Review. *Neuropsychol. Rev.* **16**, 17–42 (2006).
19. Weiss, E. M. *et al*. Sex differences in clustering and switching in verbal fluency tasks. *J. Int. Neuropsychol. Soc.* **12** (2006).
20. Henry, J. D. & Crawford, J. R. Verbal fluency deficits in Parkinson's disease: a meta-analysis. *J. Int. Neuropsychol. Soc. JINS* **10**, 608–622 (2004).
21. Paula, J. J., de Paiva, G. C. de C. & Costa, D. de S. Use of a modified version of the switching verbal fluency test for the assessment of cognitive flexibility. *Dement. Neuropsychol.* **9**, 258–264 (2015).
22. Fisk, J. E. & Sharp, C. A. Age-Related Impairment in Executive Functioning: Updating, Inhibition, Shifting, and Access. *J. Clin. Exp. Neuropsychol.* **26**, 874–890 (2004).
23. Shao, Z., Janse, E., Visser, K. & Meyer, A. S. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* **5** (2014).
24. Unsworth, N., Spillers, G. J. & Brewer, G. A. Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *Q. J. Exp. Psychol.* **64**, 447–466 (2011).
25. Nikolai, T. *et al*. Semantic verbal fluency impairment is detectable in patients with subjective cognitive decline. *Appl. Neuropsychol. Adult* **25**, 448–457 (2018).
26. Azuma, T. Working Memory and Perseveration in Verbal Fluency. *Neuropsychology* **18**, 69–77 (2004).
27. Hedden, T. & Yoon, C. Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology* **20**, 511–528 (2006).
28. Benjamin, M. J., Cifelli, A., Garrard, P., Caine, D. & Jones, F. W. The role of working memory and verbal fluency in autobiographical memory in early Alzheimer's disease and matched controls. *Neuropsychologia* **78**, 115–121 (2015).
29. Fournier-Vicente, S., Larigauderie, P. & Gaonac'h, D. More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. *Acta Psychol. (Amst.)* **129**, 32–48 (2008).
30. Whiteside, D. M. *et al*. Verbal Fluency: Language or Executive Function Measure. *Appl. Neuropsychol. Adult* **23**, 29–34 (2016).
31. Elgamal, S. A., Roy, E. A. & Sharratt, M. T. Age and Verbal Fluency: The Mediating Effect of Speed of Processing. *Can. Geriatr. J. CGJ* **14**, 66–72 (2011).
32. Little, D. M. & Hartley, A. A. Further evidence that negative priming in the Stroop color-word task is equivalent in older and younger adults. *Psychol. Aging* **15**, 9–17 (2000).
33. Palmer, E. C., David, A. S. & Fleming, S. M. Effects of age on metacognitive efficiency. *Conscious. Cogn.* **28**, 151–160 (2014).
34. Souchay, C. & Isingrini, M. Age related differences in metacognitive control: Role of executive functioning. *Brain Cogn.* **56**, 89–99 (2004).
35. Duncan, J., Emslie, H., Williams, P., Johnson, R. & Freer, C. Intelligence and the Frontal Lobe: The Organization of Goal-Directed Behavior. *Cognit. Psychol.* **30**, 257–303 (1996).
36. Friedman, N. P. *et al*. Greater Attention Problems During Childhood Predict Poorer Executive Functioning in Late Adolescence. *Psychol. Sci.* **18**, 893–900 (2007).
37. Hidalgo-Lopez, E. & Pletzer, B. Interactive Effects of Dopamine Baseline Levels and Cycle Phase on Executive Functions: The Role of Progesterone. *Front. Neurosci.* **11** (2017).
38. Sundström Poromaa, I. & Gingnell, M. Menstrual cycle influence on cognitive function and emotion processing—from a reproductive perspective. *Front. Neurosci.* **8** (2014).
39. Berent-Spillson, A. *et al*. Distinct cognitive effects of estrogen and progesterone in menopausal women. *Psychoneuroendocrinology* **59**, 25–36 (2015).
40. Stauble, M. R., Thompson, L. A. & Morgan, G. Increases in cortisol are positively associated with gains in encoding and maintenance working memory performance in young men. *Stress* **16**, 402–410 (2013).
41. McCormick, C. M., Lewis, E., Somley, B. & Kahan, T. A. Individual differences in cortisol levels and performance on a test of executive function in men and women. *Physiol. Behav.* **91**, 87–94 (2007).
42. Oei, N. Y. L., Everaerd, W. T. A. M., Elzinga, B. M., Well, S. V. & Bermond, B. Psychosocial stress impairs working memory at high loads: An association with cortisol levels and memory retrieval. *Stress* **9**, 133–141 (2006).
43. Grikšiene, R. & Ruksenas, O. Effects of hormonal contraceptives on mental rotation and verbal fluency. *Psychoneuroendocrinology* **36**, 1239–1248 (2011).
44. McAllister-Williams, R. H. & Rugg, M. D. Effects of repeated cortisol administration on brain potential correlates of episodic memory retrieval. *Psychopharmacology (Berl.)* **160**, 74–83 (2002).
45. Newcomer, J. W. *et al*. Decreased memory performance in healthy humans induced by stress-level cortisol treatment. *Arch. Gen. Psychiatry* **56**, 527–533 (1999).
46. Shao, Z., Roelofs, A. & Meyer, A. S. Sources of individual differences in the speed of naming objects and actions: the contribution of executive control. *Q. J. Exp. Psychol. 2006* **65**, 1927–1944 (2012).
47. Bedi, G. *et al*. Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophr.* **1**, 15030 (2015).
48. Rezaii, N., Walker, E. & Wolff, P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *Npj Schizophr.* **5**, 9 (2019).
49. Stoet, G. PsyToolkit: A software package for programming psychological experiments using Linux. *Behav. Res. Methods* **42**, 1096–1104 (2010).
50. Aschenbrenner, S., Tucha, O. & Lange, K. W. *Regensburger Wortflüssigkeits-Test: RWT*. (Hogrefe, Verlag für Psychologie, Göttingen).
51. Ruff, R. M., Light, R., Parker, S. B. & Levin, H. S. Benton Controlled Oral Word Association Test: reliability and updated norms. *Arch. Clin. Neuropsychol. Off. J. Natl. Acad. Neuropsychol.* **11**, 329–338 (1996).
52. Deng, Y., Dai, Q. & Zhang, Z. An Overview of Computational Sparse Models and Their Applications in Artificial Intelligence. In *Artificial Intelligence, Evolutionary Computing and Metaheuristics: In the Footsteps of Alan Turing* (ed. Yang, X.-S.) 345–369, https://doi.org/10.1007/978-3-642-29694-9_14 (Springer Berlin Heidelberg, 2013).
53. Koren, R., Kofman, O. & Berger, A. Analysis of word clustering in verbal fluency of school-aged children. *Arch. Clin. Neuropsychol.* **20**, 1087–1104 (2005).
54. Troyer, A. K., Moscovitch, M. & Winocur, G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology* **11**, 138–146 (1997).
55. Duncan, J., Burgess, P. & Emslie, H. Fluid intelligence after frontal lobe lesions. *Neuropsychologia* **33**, 261–268 (1995).
56. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **6**, 316 (2003).

57. Hayashi, M., Kato, M., Igarashi, K. & Kashima, H. Superior fluid intelligence in children with Asperger's disorder. *Brain Cogn.* **66**, 306–310 (2008).
58. Aken, L., van Kessels, R. P. C., Wingbermühle, E., Veld, W. Mvander & Egger, J. I. M. Fluid intelligence and executive functioning more alike than different? *Acta Neuropsychiatr.* **28**, 31–37 (2016).
59. Friedman, N. P. *et al*. Not all executive functions are related to intelligence. *Psychol. Sci.* **17**, 172–179 (2006).
60. Roca, M. *et al*. The relationship between executive functions and fluid intelligence in schizophrenia. *Front. Behav. Neurosci.* **8**, 46 (2014).
61. Goitia, B. *et al*. The relationship between executive functions and fluid intelligence in euthymic Bipolar Disorder patients. *Psychiatry Res.* **257**, 346–351 (2017).
62. Daneman, M. Working memory as a predictor of verbal fluency. *J. Psycholinguist. Res.* **20**, 445–464 (1991).
63. Rosen, V. M. & Engle, R. W. The role of working memory capacity in retrieval. *J. Exp. Psychol. Gen.* **126**, 211–227 (1997).
64. Enkavi, A. Z. *et al*. Reply to Friedman and Banich: Right measures for the research question. *Proc. Natl. Acad. Sci.* **116**, 24398 (2019).
65. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186 (2018).
66. Donoghue, O. A. *et al*. Association Between Timed Up-and-Go and Memory, Executive Function, and Processing Speed. *J. Am. Geriatr. Soc.* **60**, 1681–1686 (2012).
67. Sheline, Y. I. *et al*. Cognitive Function in Late Life Depression: Relationships to Depression Severity, Cerebrovascular Risk Factors and Processing Speed. *Biol. Psychiatry* **60**, 58–65 (2006).
68. Trapp, W. *et al*. Speed and capacity of working memory and executive function in schizophrenia compared to unipolar depression. *Schizophr. Res. Cogn.* **10**, 1–6 (2017).
69. Rodríguez-Aranda, C. Reduced writing and reading speed and age-related changes in verbal fluency tasks. *Clin. Neuropsychol.* **17**, 203–215 (2003).
70. Sauzéon, H. *et al*. Verbal Knowledge as a Compensation Determinant of Adult Age Differences in Verbal Fluency Tasks over Time. *J. Adult Dev.* **18**, 144–154 (2011).

Acknowledgements

This research was supported by The Deutsche Forschungsgemeinschaft (DFG, EI 816/11-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1) 785907 (HBP SGA2). We are very grateful to Lisa Mochalski, Natalie Schlothauer and Hannah Hensen for help with testing participants. We also thank Kaustubh R. Patil for his methodological support and for providing the algorithms of the prediction analysis.

Author contributions

The study was designed by all authors of the manuscript. In particular, the executive functions test selection was mainly supported by Julia Camilleri and Stefan Heim's expertise mainly contributed to the discussing of speech-related topics. Data collection was done by Julia Amunts, supported by research assistants and master students, mentioned in the acknowledgements. Data analysis was mainly driven by Susanne Weis, Julia Camilleri, Simon Eickhoff and Julia Amunts, especially involving Susanne Weis' and Simon Eickhoff's knowledge of advanced data science. The manuscript was mainly written by Julia Amunts who was supervised throughout the writing process by Susanne Weis and Julia Camilleri who provided improvements for structuring the manuscript as well as improving the wording. Moreover, Simon Eickhoff and Stefan Heim provided detailed feedback on the overall manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-65525-9.

Correspondence and requests for materials should be addressed to J.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

4 Comprehensive verbal fluency features predict executive function performance, Amunts, J., Camilleri, J.A., Eickhoff, S.B., Patil, K.R., Heim, S., von Polier, G., Weis, S., Scientific Reports, 11: 6929, (2021)

scientific reports

Check for updates

OPEN

Comprehensive verbal fluency features predict executive function performance

Julia Amunts[1,2]✉, Julia A. Camilleri[1,2], Simon B. Eickhoff[1,2], Kaustubh R. Patil[1,2], Stefan Heim[3,4], Georg G. von Polier[1,5,6] & Susanne Weis[1,2]

Semantic verbal fluency (sVF) tasks are commonly used in clinical diagnostic batteries as well as in a research context. When performing sVF tasks to assess executive functions (EFs) the sum of correctly produced words is the main measure. Although previous research indicates potentially better insights into EF performance by the use of finer grained sVF information, this has not yet been objectively evaluated. To investigate the potential of employing a finer grained sVF feature set to predict EF performance, healthy monolingual German speaking participants (n = 230) were tested with a comprehensive EF test battery and sVF tasks, from which features including sum scores, error types, speech breaks and semantic relatedness were extracted. A machine learning method was applied to predict EF scores from sVF features in previously unseen subjects. To investigate the predictive power of the advanced sVF feature set, we compared it to the commonly used sum score analysis. Results revealed that 8 / 14 EF tests were predicted significantly using the comprehensive sVF feature set, which outperformed sum scores particularly in predicting cognitive flexibility and inhibitory processes. These findings highlight the predictive potential of a comprehensive evaluation of sVF tasks which might be used as diagnostic screening of EFs.

Executive functions (EFs) comprise cognitive processes that enable goal directed behaviour[1]. Previous literature investigated the general cognitive processes that fall under the umbrella term of EFs and encompass both lower-level cognitive processes and higher-level processes. The former include working memory, inhibition and cognitive flexibility which represent the building blocks for higher-level processes such as planning, reasoning and problem solving[2].

While the number and definition of different EF subprocesses remains controversial[3], there is strong evidence that EFs are impaired in a large number of neurological[4,5] and psychiatric[6,7] diseases. Therefore, the measurement of EFs forms a crucial part of the clinical neuropsychological diagnostical routine in order to detect and specify impairments such as frontal lobe damages[8]. Multiple test batteries such as the *Delis-Kaplan Executive Function System* (D-KEFS)[9] and the *Vienna Test System*[10] provide numerous EF tests to capture a wide range of the different aspects of EFs. However, many EF tests are mainly based on pen-and-paper versions which tend to be time consuming while also lacking accuracy. Moreover, there are discrepancies between unnatural test instructions and naturalistic tasks in everyday life which leads to a lack of ecological validity of commonly used EF tests[11].

There is consensus, that EFs play a crucial role in speech production processes[12,13]. Cognitive flexibility is required to activate general lexical concepts while later working memory capacities are needed for remembering already produced words. Here, the episodic buffer and phonological loop, which are also related to the working memory system, serve as central components[12]. Since EFs are also involved in speech production, verbal fluency (VF) tests are integrated in several clinical diagnostic batteries to assess EFs. E.g. in *B-CATS*—an assessment tool for schizophrenia; *NIH stroke scale* – assessment for quantifying stroke severity; BCSB—screening for mild

[1]Institute of Neuroscience and Medicine (INM-7 Brain and Behaviour), Forschungszentrum Jülich, Wilhelm-Johnen-Str, 52428 Jülich, Germany. [2]Institute of Systems Neuroscience, Heinrich-Heine University, Moorenstr. 5, 40225 Düsseldorf, Germany. [3]Institute of Neuroscience and Medicine (INM-1 Structural and functional organization of the brain), Forschungszentrum Jülich, Wilhelm-Johnen-Str, 52428 Jülich, Germany. [4]Department of Psychiatry, Psychotherapy und Psychosomatics, Medical Faculty, RWTH Aachen University, Pauwelsstraße 30, 52074 Aachen, Germany. [5]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe-Universität Frankfurt am Main, Deutschordenstraße 50, 60528 Frankfurt am Main, Germany. [6]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Medical Faculty, RWTH Aachen University, Neuenhofer Weg 21, 52074 Aachen, Germany. ✉email: j.amunts@fz-juelich.de

Alzheimer´s disease; and D-KEFS—broadly applicable assessment used for assessing diseases such as epilepsy or Parkinson´s disease.

Two different types of VF tests are commonly used. On the one hand, VF is assessed with a lexical task. In this task participants are asked to produce as many words as possible with a specific initial letter within a specific timeframe (usually 1–2 min). Due to the fact that all requested words start with the same phoneme, the lexical VF task is also commonly referred to as *phonological* VF task. On the other hand, the semantic VF (sVF) task requires the production of words belonging to a specific category (e.g. animals or fruits), regardless of the initial letter of the word. The lexical VF task is driven by phonological and lexical cues, whereas the sVF task requires attributes of a specific semantic category.

Within each type of the VF task, it is also possible to modulate the demand on EFs by applying a switching component. Here, participants are asked to switch between two different categories in alternating order within the same task (e.g. fruits-jobs). VF performance is generally evaluated by calculating the total number of correctly produced items. However, in the neurological literature, it has been shown that specific patterns of VF impairment greatly depend on the damaged brain regions[14,15]. Thus, studies suggest the need for a more differentiated analysis of VF performance[16].

In general, there is consensus on the involvement of EFs in the VF task in healthy controls[17] as well as their impairment in patients[15,18]. In detail, it is assumed that semantic knowledge and memory as well as cognitive flexibility are required to build semantic associations in sVF tasks whereas the lexical VF tasks require the suppression of grouping words with shared associations[19]. Additionally, in both types of VF tasks, inhibition is presumably needed to suppress competitive responses and to avoid perseveration errors[20,21], while attention, updating and working memory processes are simultaneously involved to keep the processing speed high, to remember already produced items and to produce as many items as possible[12].

Although previous findings undergirded the involvement of executive control processes in the VF task[22], the diagnostic validity of VF tasks to assess EF performance remains controversial[23,24]. In particular, it has been found to be affected by multiple factors such as the underlying language component in the VF task, underlying cognitive processes such as intelligence, and fluctuating hormonal levels[25,26]. Moreover, the literature is not in agreement with regards to the specific relationship between VF and EF. Various studies report a positive correlation between working memory, inhibition, cognitive flexibility performance and the total score of produced words[22,27]. In contrast, other studies failed to identify a clear relationship between VF performance and EFs in one or more EF domains[24,28]. Notably, in previous studies, classical statistic methods were used to e.g. investigate group comparisons of EF performance in patients and healthy controls. Applying correlational analyses, studies investigated linear relationships of VF sum scores and different EF domains[29].

However, within the last years VF tasks per se have gained more interest as a predictive tool for clinical decision making, e.g. in schizophrenia[30] or mild cognitive impairments[31] since they offer an alternative to the highly time-consuming testing procedure of EFs[11]. The growing interest in the predictive value of VF tasks might be a result of the increasing use of machine learning algorithms investigating speech production to predict disease specific properties[32–34]. The main appeal of the machine learning approach is its ability to train a predictive model by identifying patterns in high dimensional data which can be subsequently used to make predictions in unseen data. Additionally, interpreting models can provide information with regards to which specific features contribute most to accurate predictions. Based on a data-driven learning, predictive modelling enables researchers to capture (non)-linear relationships, generalize associations and to potentially subsequently transfer these to a clinical context.

Although the VF task is commonly evaluated based on the total sum of correct produced words[35–37], other variables can also be employed to gain deeper insights into cognitive performance.

Recent studies have demonstrated the potential of advanced parameters taken from the VF task, i.e. error types[38], latencies[39] and semantic distances[18] to complement the common analysis of the total sum of words. These additional variables, assessed within the sVF task, were shown to reflect the complex involvement of executive processes in disorders such as dementia as well as in better differentiation between patients and healthy controls[16,40].

To interpret VF performance in more detail, studies have also investigated error types that occurred in the course of the VF task such as those based on the breaking of sVF-specific rules (e.g. naming words from a different category, creating neologisms) and category errors. Perseveration and category errors are particularly reported in the switching VF task when participants fail to switch to the second category, name words from a different category or repeat the same category twice[37]. Thus, perseveration and category errors can provide qualitative information when measuring VF performance, in addition to the commonly used total sum of words.

Additionally, information of the VF task can also be assessed on a semantic level, analysing semantic relatedness of produced words. This concept was first investigated by Troyer et al.[41] who manually organized produced words in the sVF task into conceptually related clusters and switches. Specifically, semantically related words were clustered based on specific subcategories[41]. For example, animals were clustered based on their living environment, human use and zoological categories. According to these clusters, which are usually defined as a minimum of a two-word-sequence within the same subcategory, switches were calculated as the total number of shifts between these clusters[41]. Here, two types of switches were defined: While *cluster switches* describe a transition between multiword and adjacent clusters, *hard switches* represent transitions between a cluster and non-clustered words[42]. Later research showed that the ability to create new subcategories and generate new cues is more important for performing the sVF task than creating large cluster sizes[43]. Moreover, authors highlighted the importance of working memory capacity for self-generating category cues in healthy participants[43] and suggested the sVF task as a diagnostic tool in cognitive impairment[44,45]. Nevertheless, this assessment of semantic information from the sVF task was traditionally done manually and thus was highly time-consuming and partially subjective due to the manual determination and assignment of subcategories[46,47]. However, this problem can be

addressed with the help of computational linguistics providing automated computational approaches (e.g. Latent semantic analysis[48], Word2Vec[49]). Nowadays large text corpora and fine grained information of semantic relatedness are available (e.g. WordNet[50], DISCO[51]). In general, different conceptual structures are implemented in these models. On the one hand, some systems provide the hierarchical structure of a lexical semantic net[52] based on semantic concepts (e.g. fishes, birds, mammals)[50]. In contrast to this hierarchical and ontological approach, vector-based systems rely on the co-occurrence of words within a big text corpus. Here, words are represented as a point in a multi-dimensional space creating word embeddings[53]. Applying these computerized and automated systems, studies were able to identify dementia risk in healthy participants based on semantic relatedness[16] and to distinguish between patients with forms of disorganization and healthy controls[54].

Alongside the semantic information, the sVF task also provides prosodic information such as speech latencies (speech pauses between each word). Latencies convey information about the approximate time needed to access lexical items[13,55]. Although there is little literature on the relationship between speech latencies in the VF task and EF performance, some findings indicate that it might be meaningful[39,43]. Specifically, studies suggest that a higher incidence of unfilled pauses are more likely to occur in situations in which participants are confronted with a higher planning load[56]. Other studies also report a relationship between prosodic information and EF demands showing a decreased production of words within the progress of the VF task[39]. Since a decrease of the number of produced words in the VF task also indicates an increase of speech latencies[39] these findings suggest that speech latencies could provide additional information on VF performance with respect to the involvement of EFs.

In summary, previous studies indicate the potential of additional quantitative measures for evaluating sVF performance to gain better insight into cognitive processes. However, diagnostic batteries used in the clinical context as well as in the scientific environment still heavily rely on the sole use of the sum of correct words as the main indicator of EF performance. Consequently, the aim of the present study was to investigate the predictive power of a comprehensive set of sVF measures and compare it to the commonly used sum score analysis. As a first step into deeper insights of the predictive power of the VF task, we focus on the semantic VF task which allowed us to exploit the vast information within the semantic relatedness features. In this exploratory study, machine learning methods were applied to predict performances of well validated but highly time-consuming EF tests from a broad set of objective and mainly computerized VF measurements in unseen participants. We expected the extended sVF feature set to outperform the basic analysis of sum scores in predicting EF test results.

Methods

Participants. In this study, 230 healthy participants with an age range of 20–55 years (mean age 35.2 ± 11.1; 92 males) were tested. Before the actual testing session, participants were asked for previously detected diagnoses. Only participants without neurological or psychiatric diagnoses were included in this study. Moreover, participants were monolingual German speakers, i.e. their native language was German and they did not learn an additional language before going to school. Participants received different levels of education (finished middle school: 8, professional school/job training: 63, finished high school with a university-entrance diploma: 69, university degree: 90). The recruitment took place in North Rhine-Westphalia (Germany) via social networks and the Forschungszentrum Jülich mailing list. Participants were tested at the Forschungszentrum Jülich, and the testing session included an EF test battery together with VF tasks, with a duration of 150–180 min depending on the individual time needed for instructions and the speed with which the participants passed the tests. A remuneration fee of €50 was paid. All experiments were performed in accordance with relevant guidelines and regulations. Moreover, informed consent was obtained from all participants. Collection and analyses of the data presented here was approved by the ethics committee at Heinrich-Heine University Düsseldorf.

Executive function assessment. The EF test battery consisted of 14 computerized versions of commonly used neuropsychological tests covering domains of cognitive flexibility, working memory and inhibition. While 11 of these tests were taken from the *Vienna Testsystem*[10], three were designed with *PsyToolkit*[57]. The *Vienna Testsystem*[10] is a standardized computerized test battery providing numerous EF tests and test manuals. Every EF test provided multiple variables which were extracted automatically by the respective test system. While some of these variables represent main variables, others solely include processing time information which are not directly linked to the EF performance. EF tests which were designed within *PsyToolkit*[57] do not come with associated test manuals and the selection of variables of these tests was thus based on previous literature[58–60].

Cognitive flexibility was assessed using five tests, namely, the *Trail Making Test*[61] (TMT), *Raven´s Standard Progressive Matrices*[62] (SPM), *Wisconsin Card Sorting Test*[63] (WCST), *Tower of London* (TOL)[64] and *Cued-Task Switching*[65] (SWITCH).

Working memory performance was examined using three tests: *N-back non-verbal Test*[66] (NBN), *Non-verbal Learning Test*[67] (NVLT) and *Corsi Block Tapping Test*[68] (CORSI).

Inhibition was tested using *Stop-Signal Task*[69] (STOP), *Simon Task*[70] (SIMON) and *Stroop Test*[71] (STROOP).

Additionally, we also assessed divided and spatial attention (WAF-G[72], WAF-R[72]) as well as vigilance (*Mackworth Clocktest*[60] (CLOCK)). In total, 68 variables were extracted from EF tests. The full set of EF test variables is provided in the supplementary material (Table S1).

Semantic verbal fluency tasks. The sVF tasks were based on the *Regensburger Wortflüssigkeitstest*[37] (RWT) which is equivalent to the English *Controlled Oral Word Association Test*[73] (COWAT). The German standardized neuropsychological version of the VF task was used due to language-specific differences in the frequency and usage of letters and categories[36]. Two of the tasks were simple sVF tasks in which the participant had to name animals (t_1) and jobs (t_2). The third sVF task (t_3) was a switching task in which the participant switched

VF features	Description
Correct words t1 + t2 + t3	Sum of all correct produced words in task1, task2, task3
Correct words	Sum of correct produced words in each task
Switch coefficient	Relationship of correct items in simple and switching tasks; switching coefficient = sum3/((sum1 + sum2)/2))
Repetition error	Repetition errors in task 1, task 2
Category error	Category errors in task 3
Latency mean	Mean of speech breaks in each task
Latencies 1st quarter	Mean of speech breaks in seconds 0–30 (i1) for each task
Latencies 2nd quarter	Mean of speech breaks in seconds 31–60 (i2) for each task
Latencies 3rd quarter	Mean of speech breaks in seconds 61–90 (i3) for each task
Latencies 4th quarter	Mean of speech breaks in seconds 91–120 (i4) for each task
Latency difference	Progress of speech breaks (i4-i1) in each task
Sequential mean	Semantic mean of all sequential word pairs in each task; computed with GermaNet (hierachical)
Cumulative mean	Semantic mean of all possible word pairs (cumulative) in each task; computed with GermaNet (hierarchical)
Sequential mean cat1 t3	Semantic mean of all sequential word pairs (sequential) in catergory 1 (sports) of switching task; computed with GermaNet (hierarchical)
Sequential mean cat2 t3	Semantic mean of all sequential word pairs (sequential) in catergory 2 (fruits) of switching task; computed with GermaNet (hierarchical)
Sequential mean DIS	Semantic mean of all sequential word pairs in each task; computed with DISCO (Word2Vec)
Cumulative mean DIS	Cumulative mean of all possible word pairs in each task; computed with DISCO (Word2Vec)
Sequential mean cat1 t3 DIS	Semantic mean of all sequential word pairs in category 1 (sports) of switching task; computed with DISCO (Word2Vec)
Sequential mean cat2 t3 DIS	Semantic mean of all sequential word pairs in category 2 (fruits) of switching task; computed with DISCO (Word2Vec)

Table 1. Overview of Verbal fluency features.

between fruits and sports within the same task. Each of the three tasks was performed for 2 min. The sVF tasks were presented with *Presentation* software[74] and the participant´s responses were recorded automatically.

Following the testing session, the recorded speech was transcribed and words were coded manually as being either *correct answers* or *errors*. Furthermore, errors were differentiated into perseveration and category errors. Sum scores of each sVF tasks separately, sum score of correct produced words across all sVF and errors (perseveration, category errors) were included in the prediction analysis. In general, the sum scores solely include correct produced items in all three sVF tasks. A list of extracted sVF features is shown in Table 1.

Speech latencies were automatically detected and manually corrected using *PRAAT*[75], and the mean of the speech *latencies* within each task was calculated. Moreover, the task was divided into four 30-seconds intervals (i_1, i_2, i_3, i_4) and the mean of the speech latencies within each interval was determined. Additionally, these means of intervals were then used to determine an increase or decrease of speech latencies within each task (i_4-i_1). Latency means of each task and of each interval as well as latency differences were defined as sVF features for prediction analysis.

Semantic distances were computed using two different approaches to ensure that the results of prediction analysis are not dependent on a specific semantic system. One of the semantic systems was a hierarchical structured lexical-semantic net of *GermaNet*[52] and *GermaNet-Pathfinder*[76]. Specifically, this lexical network is partitioned into various sets of semantic concepts (*synsets*) that are intertwined by semantic relations and create nodes. These synsets are related conceptually in different ways including, hypernymy, part-whole relations, entailment and causation[52], leading to hierarchical-structured subcategories. *GermaNet-Pathfinder*[76] provides different measurements[77] for the determination of how closely two nouns are related to each other. In this study, we selected a path-based measure which describes the relatedness between concepts. In detail, the path-based system takes the distance between two synset nodes and the longest possible shortest path between any two nodes in GermaNet into account.

$$sim(s_1, s_2) = \frac{MAXSHORTESTPATH - length(s_1, s_2)}{MAXSHORTESTPATH}$$

length(s_1,s_2) = shortest path between synset s1 and synset s2.

MAXSHORTESTPATH = maximum of all shortest paths within GermaNet.

Applying this formula, semantic relatedness is represented by values between 0 and 1. While closely related words lead to values approximating 1 (German Shepard x Labrador → sim = 0.94), more distanced word pairs lead to smaller values (e.g. German Shepard x dolphin → sim = 0.77).

The other semantic system that was used to determine semantic similarity between words was *DISCO*[78] applying a Word2Vec[49] approach. This system is based on co-occurrences in large text corpora. Specifically, this corpus contains 1.5 billion tokens including German Wikipedia entries, newspaper articles, parliamentary debates, movie subtitles and more. Each unique word is represented by a word vector and is part of the vector space.

Within this vector space, word vectors are located based on shared common contexts building word embeddings. As in *GermaNet*[52], a high semantic similarity is represented by numbers approximating 1.

Each sVF task of the participants was analysed automatically using *GermaNet Pathfinder*[76] and *DISCO API*[78]. For our feature-set which was later used for the prediction analysis, two different types of semantic relations were extracted: (1) Sequential distance was computed across each consecutive word pair in order of the produced words. (2) Cumulative distance was computed over the entire task regardless of the order in which they appear within the task. As an output, the relatedness between each word-pair was extracted and the mean of all semantic relations within one task was calculated. In the case that *GermaNet* contained more than one synset for one word, the synset with closest relatedness to the paired word was selected. Moreover, missing lexical entries in *GermaNet* or *DISCO* led to a deletion of the corresponding word pair. All semantic information, including means of sequential and cumulative distances of both systems (*GermaNet* and *DISCO*) were added as features to prediction analysis.

Altogether, 43 features were extracted from the sVF tasks containing information of sum of correct words, error types, speech latencies and semantic distances calculated with two different systems. A complete overview of VF feature scores is provided in the supplementary material (Table S2).

Machine learning analysis. In this study, we applied a machine approach using a cross-validation procedure. Here, just parts of the data are used to train the model while the other part is used to validate the model; i.e. EF scores were predicted in unseen participants which allows for generalization of results to a certain degree.

EF performance was predicted from sVF variables (*features*) applying supervised learning via random forests[79,80] (RF). The sVF features were used to predict each of the 68 EF scores (*targets*) in separate and independent analyses. Generally speaking, RF creates a "forest" of decision trees as weak learners by randomly sampling the features before learning each decision tree. The trees are used as an ensemble and the prediction of individual trees is averaged to get the final prediction[81]. In the present study 100 trees were used to compute prediction analysis.

Previous work indicates that performance in the VF task is negatively related to age[82,83]. Moreover, sex was found to be associated with differential solving strategies in the VF task[84]. Likewise, a higher level of education was associated with better performance in VF tasks[82,85]. Therefore, data was transformed to z-scores and sex, age and education were regressed out from the sVF features within cross-validation. A tenfold cross-validation procedure was performed for which the data set was randomly split into 10 sets, 9 of which were used for training while the 10th set was held back and used to assess the prediction performance in previously unseen data. Ten repetitions of the tenfold cross-validation were performed and thus 100 prediction models for each EF target were computed. Prediction performance was assessed by computing the mean correlation (*Pearson*) between real and predicted values within cross-validation folds and subsequently across all repetitions. EF targets which were predicted from sVF features at a significance level of $p < 0.01$ were considered *highly predictable* EF targets.

To compare the predictive power of the comprehensive and the classical feature set, the prediction analysis was computed for classical sVF features, solely containing information from sum scores of sVF tasks.

The sVF features which contributed most strongly to the prediction analyses of each *highly predictable* EF target were identified. Feature importance was defined by the permutation of out-of-bag predictor observations as implemented in *Matlab*[86]. The top five sVF features with the highest feature performance were identified to further investigate the (non) linear relationship of these sVF features with the respective EF performance. Here, rank correlations (*Spearman*) of sVF features and EF test scores were calculated. Due to the high number of extracted EF variables, only one highly significantly predicted EF test variable of each EF test is presented to exemplarily demonstrate the complex relationship of sVF features and EF performance. The selection of this representative EF variable was based on the test EF manuals and previous literature describing specific main variables of each EF test.

Results

Prediction of EF variables from verbal fluency data. To investigate which EF targets were predictable from sVF features, we computed two independent prediction analyses. In the first analyses the full set of sVF features, including sum scores, errors, latencies and semantic relatedness was used (Fig. 1). The second analysis was performed with variables containing only information regarding the number of correctly produced items in each sVF task (Fig. 2). Both figures show the EF targets that were significantly predicted from sVF features at a significance level of $p < 0.01$. Detailed results of all prediction analyses are given in the supplementary material (Table S3).

In sum, 20 EF targets, pertaining to 8 different EF tests and tapping into all subdomains of EFs, could be predicted significantly from the full feature set. With respect to cognitive flexibility, TMT, SPM and WCST were predicted from sVF data. The highest correlation between true and predicted values was identified in processing times of part A ($r = 0.41$; $p = 3.2e^{-10}$) and B ($r = 0.33$; $p = 2.6e^{-7}$) of TMT. While these results are primarily related to overall processing speed, an explicit relationship between sVF performance and cognitive flexibility was found in the "*difference between part B-A*" of the TMT ($r = 0.17$ $p = 0.007$) as well as in the test results of SPM and WCST. Here, the number of correct items in the SPM ($r = 0.20$; $p = 0.001$) and different error types in the WCST revealed the complexity of cognitive requirements and planning ability in conducting the sVF task. With regards to tests assessing working memory capacity, two of three EF tests, namely NVLT ($r = 0.24$; $p = 0.0002$) and NBN ($r = 0.16$; $p = 0.009$) were predicted significantly. Beside EF targets referring to cognitive flexibility and working memory, the analysis also identified inhibition targets. Particularly, *naming interference* ($r = 0.24$; $p = 0.0002$) and processing time in STROOP ($r = 0.23$; $p = 0.0003$) were predicted.

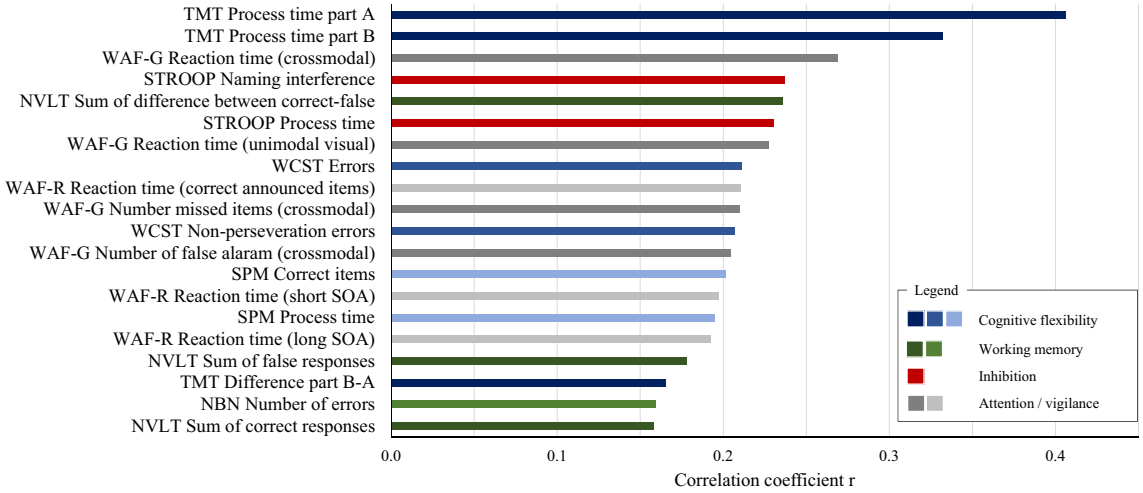

Figure 1. Correlation coefficients of true and predicted executive function variables computed with full feature set. Executive function variables were predicted based on 43 verbal fluency features. Results shown in this table illustrate executive function variables which could be predicted at $p < 0.01$ from verbal fluency data; Colour groups indicate EF domains and colour gradients denote different EF tests within this EF domain; *NBN* N-back non-verbal; *NVLT* Non-verbal learning test; *SOA* Stimulus onset asynchrony; *SPM* Raven's standard progressive matrices; *STROOP* Stroop test; *TMT* Trail making test; *WCST* Wisconsin Card Sorting Test; *WAF-G* Divided attention; *WAF-R* Spatial attention.

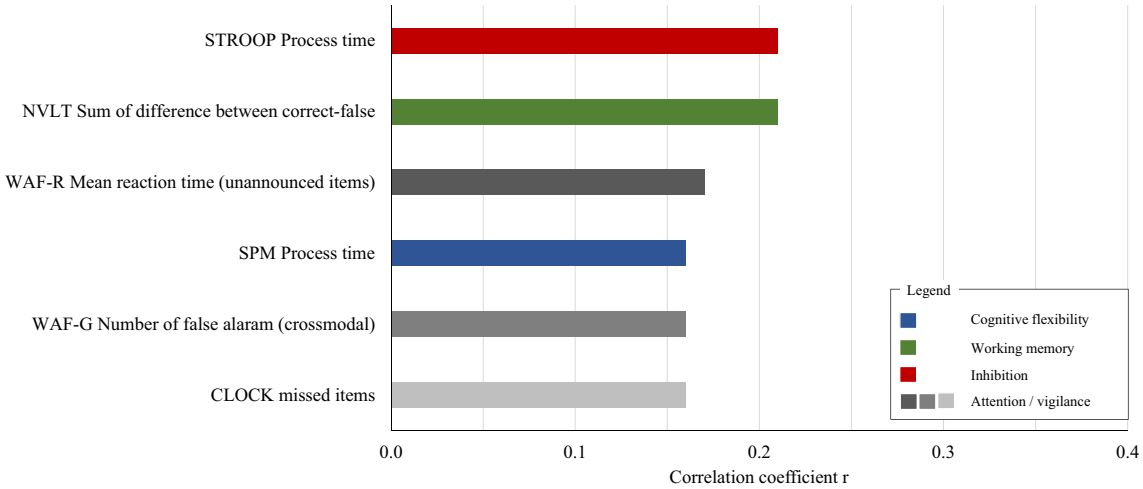

Figure 2. Correlation coefficients of true and predicted executive function variables computed with classical feature set. Executive function variables were predicted based on the sum scores of all 3 semantic verbal fluency tests as well as the total sum score across these 3 tests, which led to a total number of four verbal fluency features. Results shown in this table illustrate executive function variables which could be predicted with $p < 0.01$ from verbal fluency data; Colour groups indicate EF domains and gradients denote different EF tests within this EF domain; *MACK* Mackworth Clock Test; *NVLT* non-verbal learning test; *SPM* Raven's Standard Progressive Matrices; *STROOP* Stroop Test; *WAF-G* Divided attention; *WAF-R* Spatial attention.

Across all subdomains of EFs, variables displaying general processing speed and reaction times performance were detected. The role of attention and general processing speed is also represented in test results referring to divided and spatial attention. Here, seven targets of attention tests were predicted significantly. In general, tests from all EF subdomains were predicted significantly and no dominance of one specific subdomain was apparent.

The focus of this study was the predictive power of an advanced VF feature set. To compare the predictive power of the advanced features with the commonly used VF information, i.e. the sum of correctly produced words, an additional prediction analysis was computed using solely sum scores. Here, the sum scores of each sVF tasks as well as a total sum score across all three tests were included. In this analysis only six EF targets were predicted significantly (Fig. 2). Prediction performance was lower than in the analysis with full feature set and particularly targets of processing speed and reactions times were detected. In contrast to the first analysis, vigilance was predicted with missed items in CLOCK ($r = 0.16$; $p = 0.009$).

SPM—correct items				TMT—difference part B-A				WCST—non-perseveration errors			
Top 5 sVF features		r	p	Top 5 sVF features		r	p	Top 5 sVF features		r	p
1	Repetition error t_1	− 0.05	0.44	1	Latencies 4th quarter t_1	− 0.01	0.88	1	Repetition error t_3	− 0.10	0.12
2	Latencies 2nd quarter t_1	− 0.09	0.18	2	**Repetition error t_3**	**− 0.12**	**0.08**	2	**Latencies 1st quarter t_1**	**0.17**	**0.01***
3	Category error t_3	− 0.01	0.87	3	Latency difference t_1	− 0.02	0.80	3	Category error t_3	0.01	0.85
4	**Correct words t_1**	**− 0.12**	**0.08**	4	Category error t_3	− 0.07	0.27	4	**Correct words t_2**	**− 0.16**	**0.02***
5	**Cum. mean t_3**	**0.14**	**0.05***	5	Repetition error t_1	− 0.01	0.80	5	**Total sum score $t_1 + t_2 + t_3$**	**− 0.16**	**0.02***

Table 2. Spearman correlations of five most important semantic verbal fluency (sVF) features with significantly predictable cognitive flexibility targets. 1–5 = Top five sVF features with regards to predictor performance based on feature importance; correlations with $p < 0.1$ are marked in bold; significant correlations ($p < 0.05$) are marked with *. *SPM* Raven's Standard Progressive Matrices; *TMT* Trail-Making Test; *WCST* Wisconsin Card Sorting Test. t_1 = VF test (animals); t_2 = VF test (jobs); t_3 = Switching VF test (sports/fruits); *Cum* cumulative.

NBN—errors				NVLT—difference correct minus errors			
Top 5 sVF features		r	p	Top 5 sVF features		r	P
1	Repetition error t_3	− 0.01	0.91	1	**Latencies 4th quarter t_1**	**− 0.12**	**0.08**
2	**Sequ. mean t_1**	**− 0.19**	**0.00***	2	Category error t_3	− 0.03	0.63
3	Category error t_3	− 0.01	0.92	3	**Latency difference t_1**	**− 0.12**	**0.08**
4	Repetition error t_1	0.11	0.11	4	Correct words t_1	0.07	0.28
5	**Cum. mean DIS t_3**	**− 0.12**	**0.07**	5	Latency mean t_2	0.03	0.69

Table 3. Spearman correlations of five most important semantic verbal fluency (sVF) features with significantly predictable working memory targets. 1–5 = Top five VF features with regards to predictor performance based on feature importance; correlations with $p < 0.1$ are marked in bold; significant correlations ($p < 0.05$) are marked with *. *NBN* N-back non-verbal; *NVLT* non-verbal learning test. t_1 = VF test (animals); t_2 = VF test (jobs); t_3 = Switching VF test (sports/fruits); *Cum* cumulative; *Sequ* sequential; *DIS* semantic system *DISCO*.

STROOP—naming interference			
Top 5 sVF features		r	p
1	**Cum. mean t_2**	**− 0.18**	**0.01***
2	Latency difference t_1	0.03	0.62
3	Latencies 4th quarter t_1	0.05	0.48
4	**Sequ. mean DIS cat$_1$ t_3**	**− 0.15**	**0.03***
5	**Total sum score $t_1 + t_2 + t_3$**	**− 0.22**	**0.00***

Table 4. Spearman correlations of five most important semantic verbal fluency (sVF) features with significantly predictable inhibition target. 1–5 = Top five VF features with regards to predictor performance based on feature importance; correlations with $p < 0.1$ are marked in bold; significant correlations ($p < 0.05$) are marked with *. t_1 = VF test (animals); t_2 = VF test (jobs); t_3 = Switching VF test (sports/fruits); *Cum* cumulative; *Sequ* sequential; *DIS* semantic system *DISCO*.

Impact of sVF features on prediction analysis. The impact of single sVF features on EF performance was quantified based on the feature importance scores of the prediction analysis. Due to the high number of significantly predicted EF targets, only one EF target for each of the significantly predicted EF tests is discussed in detail here. We focus on the main variables for the respective EF tests based on previous literature and the EF test manuals. For each of these, the five most important sVF features were extracted and correlations with the respective EF target were calculated (Tables 2, 3, 4, 5) to enable a comparison of present results with commonly used univariate analyses.

Across all EF domains, the most important sVF features for the prediction results included information about number of correctly produced words, error types, latencies and semantic distances. Out of these most predictive sVF features, some showed a significant correlation with the EF target ($p < 0.05$), while others displayed a trend level significance ($p < 0.1$) or no significant correlation at all. In the following, we assessed the top five sVF features that are related to the different EF subdomains of cognitive flexibility, working memory, inhibition as well as to attention. Due to the high number of EF scores that were predicted significantly from sVF features, one

WAF-G reaction time crossmodal				WAF-R—reaction time correctly announced			
Top 5 sVF features		r	P	Top 5 sVF features		r	p
1	Repetition error t_3	− 0.09	0.17	1	Repetition error t_2	− 0.01	0.93
2	Repetition error t_2	− 0.03	0.70	2	Repetition error t_1	− 0.09	0.17
3	**Latencies 1st quarter t_3**	**0.13**	**0.06**	3	Sequ. mean DIS t_1	0.04	0.61
4	Repetition error t_1	− 0.05	0.45	4	Cum. mean DIS t_1	0.07	0.29
5	Cum. Mean t_1	− 0.03	0.59	5	Repetition error t_3	0.01	0.88

Table 5. Spearman correlations of five most important semantic verbal fluency (sVF) features with significantly predictable attention targets. 1–5 = Top five VF features with regards to predictor performance based on feature importance; correlations with $p < 0.1$ are marked in bold; significant correlations ($p < 0.05$) are marked with *. *WAF-G* divided attention test; *WAF-R* spatial attention test. t_1 = VF test (animals); t_2 = VF test (jobs); t_3 = Switching VF test (sports/fruits); *Cum* cumulative; *Sequ* sequential; *DIS* semantic system *DISCO*.

EF variable of each significantly predicted test is presented here. A complete overview of the correlation matrix of all sVF features and significantly predicted EF scores is given in the supplementary material (Tables S4–S6).

With regards to cognitive flexibility (Table 2) 7/15 sVF features were related to errors participants produced within the sVF task. Repetition errors in simple and switching sVF tasks as well as category errors in the switching task were found to be important sVF features for predicting EF targets. Particularly, repetition and category errors were determined as highly relevant in predicting *TMT* performance. However, no significant (linear) correlation between errors and cognitive flexibility performance was found. In contrast, a linear relationship of sVF information and EF performance was shown for the number of correctly produced words. Here, significant correlations of correctly produced words and cognitive flexibility targets were primarily found in the *WCST*. Similar but not significant results were also found in the *SPM*. In all three significantly predicted EF tests (*SPM, TMT, WCST*) latencies within the sVF $task_1$ (animals) were identified as important sVF features but did not reveal correlations with EF targets except for latency patterns assessed in i_1. Here, longer speech breaks were shown to positively correlate with errors in WCST. With regards to semantic relatedness the cumulative mean within the sVF switching task (t_3), calculated with the hierarchical structured approach of *GermaNet*, was identified as a meaningful feature predicting *SPM* performance. Specifically, participants naming closely related words across both switching categories (sports and fruits) achieved better *SPM* targets.

Within the EF domain of working memory, the *NBN* and *NVLT* were identified as highly predictable EF tests (Table 3). Here, the sum of correctly produced words was selected as an important sVF feature less often than for cognitive flexibility tests and no significant correlation with EF target was found. Non-linear relationships of sVF features and working memory performance was additionally found for sVF features *errors* which were mainly important for predicting *NBN* performance. Among the five most important sVF features predicting NBN performance, the sequential as well as cumulative mean of the semantic relatedness were found to be highly relevant. Similar to results in cognitive flexibility tests (Table 1), a smaller search space ($r = -0.12$ $p = 0.07$) and closely related words ($r = -0.19$ $p = 0.005$) led to better results in *NBN*. While semantic relatedness was particularly important for predicting errors in *NBN*, latencies were relevant for *NVLT* performance. Here, results indicated a relationship between smaller speech breaks in end of the sVF task and higher NVLT target ($r = -0.12$ $p = 0.08$).

With respect to inhibition, naming interference in the *Stroop* test was predicted significantly. While error types were not selected as most important sVF features, the total sum score across all three sVF tests was determined as meaningful and revealed a significant correlation with *Stroop* performance ($r = -0.22$ $p < 0.001$) (Table 4). Important features for predicting *naming interference* performance were semantic relatedness and latencies. In particular, the searching space in t_2 represented by the cumulative mean was identified as highly important. These results indicate a better inhibition performance if participants searched for less distanced words ($r = -0.18$ $p = 0.01$). Similar results were also found in sVF features of sequential relatedness. Searching for closely related words in the first category within the switching sVF task (cat_1 t_3) was related to better inhibitory performance ($r = -0.15$ $p = 0.03$). Beside sematic relatedness and total sVF sum score, the analysis also points toward the relevance of latency patterns within the first sVF task (animals) for predicting inhibitory processes.

Finally, we investigated sVF features in the prediction of attentional performance (Table 5). Here, the results demonstrate a predictive importance of repetition errors in simple as well as in switching sVF tasks. The results revealed no significant correlation between number of errors and attention performance. Latencies within the first quarter of the switching sVF task (t_3) were selected as relevant for attention performance, indicating that a higher processing speed in the beginning of the sVF task resulted in faster reaction times in the divided attention test. Similar to previously reported results in other EF subdomains, semantic relatedness features in simple sVF task (animals) were selected as meaningful variables for attention performance.

To sum up, across all subdomains of EFs, a variety of different types of sVF features, including sum scores, error types, sematic relatedness and latencies showed high relevance for the prediction of EF performance. Out of these, about one third showed significant or trend level correlation with EF targets, while the remaining VF features that were identified as important for prediction accuracy, did not show any linear relationship with the respective EF target.

Discussion

Main findings. This study aimed to investigate whether EF performance can be predicted from sVF tasks using Machine Learning methods. In a first step, we applied a RF approach to determine which EF tests could successfully be predicted from a wide range of VF information. Results of this machine learning analysis identified EF tests tapping into all subdomains of EFs. In total, 20 of 44 EF scores were predicted significantly when using the full set of sVF features which included errors, latencies and semantic distances.

Moreover, prediction results of the full sVF features set was compared to a classical feature set including only sum scores of sVF tasks, as commonly used in clinical settings. The comparison of these two approaches revealed a larger number of significantly predicted EF scores as well as higher prediction accuracy of the advanced feature set. Particularly for cognitive flexibility performance, the comprehensive feature set achieved a higher prediction accuracy as compared to the commonly used sum score evaluation. Thus, the present results clearly demonstrate the advantage of using more comprehensive sVF features over the sole use of sum scores, which to date still tend to be the most common measure used to asses sVF tasks. In a second step, we further investigated the concrete involvement of different types of sVF features to gain insights into the impact of specific VF aspects on EF performance. Results showed that all types of sVF features, i.e. sum scores, errors, latencies and semantic relatedness contributed to the prediction of EF. With regards to the different EF subdomains no dominance of specific VF types was detected. Moreover, the correlation analyses revealed that good sVF predictors do not necessarily correlate with the respective EF score.

The following section starts with a discussion of the influence of different sVF features on prediction results. Here, predictable EF tests within each subdomain are presented and the contributions of sVF features are interpreted. Additionally, the role of general processing speed is addressed. Secondly, advantages of an elaborated VF feature set are delineated. In the end, limitations of this study are considered.

Sum scores. Summarizing scores of correctly produced items is the most commonly used way of evaluating VF tasks in the clinical and scientific context to date. The present study included separate sum scores for each sVF test as well as a total one across all sVF tasks. Results revealed the importance of sum score features for the prediction of cognitive flexibility, working memory and inhibition performance. In contrast, sum scores were not identified as important for predicting attention scores. Particularly sum scores resulting from t_1 (animals) as well as total sum scores revealed high feature importance. Furthermore, a positive linear relationship of relevant sVF sum scores and EF performance in the domains of inhibition and cognitive flexibility was found.

The findings from the present study can be directly linked to previous studies. In particular, Paula et al.[87] reported a positive correlation between cognitive flexibility performance, assessed with the TMT, and the sum of correct produced words in the switching task. With regards to working memory, another study found an association between the sum of correct produced items and working memory performance[43]. With respect to inhibition, our findings are also in line with multiple studies that demonstrated the positive linear relationship of inhibition performance and the total sum of words, assessed within the VF task, both in older[22] and young[88] adults.

Overall, based on previous literature and the results of the current study, sum scores were shown to contribute to the prediction results. In accordance with previous findings, this contribution appears to be based on a positive linear relationship of sum scores with EF performance.

Error types. When predicting EF test scores from sVF features, repetition and category errors were identified to mainly contribute to the prediction of cognitive flexibility, working memory and attention test result. Conversely, errors were not identified as important features for the prediction of inhibition scores. While both repetition and category errors were shown to be equally important for the prediction of cognitive flexibility and working memory, only repetition errors contributed to predicting attention performance. Importantly, in contrast to sum scores, most error features did not show a linear relationship with the respective EF test performance. The prediction results of TMT were the only ones to reveal a correlation trend, indicating that fewer repetition errors in sVF tasks are associated with better cognitive flexibility performance.

These findings partially contradict previous findings investigating the linear relationship between errors in the VF task and EF performance. Particularly, previous studies suggested that executive inhibitory dysfunction and reduced working memory performance lead to a higher number of perseveration errors in healthy participants[20,89]. Similar findings have also been reported in patients with brain damages[90] and schizophrenia[38]. In contrast, some studies did not find an increase in the number of perseveration errors in Parkinson´s patients compared to healthy controls[91].

Although in the present study repetition and category errors were shown to be important for the successful prediction of EF performance in all EF domains except for inhibitory processes, results revealed that a low number of produced errors does not necessarily result in better EF performance. Due to the importance of errors in prediction results and the non-linear relationship with EF performance, we assume that some participants adopt strategies where a higher number of errors is accepted in order to achieve a better score in the sVF task. Thus, successful EF performance does not necessarily go along with fewer errors.

Latencies. With regards to latency patterns our results revealed the importance of speech breaks for the prediction of all domains of EF as well as for attention scores. Latency patterns contributed differently to the prediction of different EF scores. Latency patterns during the first interval of the sVF task (i_1) were revealed as a meaningful feature for inhibitory processes, cognitive flexibility and attention performance. However, additional latency patterns, such as the mean of all latencies within each task and the progress of latencies (namely *latency differences*) also contributed to the prediction results. Interestingly, our results indicate an ambiguous relationship between latency patterns and EF test results. On the one hand, correlation analyses revealed some

significant correlations between latency patterns and EF scores with, for example, longer speech breaks in i_1 were related to a higher amount of errors in the WCST assessing cognitive flexibility performance. On the other hand, most of the latency features did not show a linear relationship with EF performance.

To our knowledge, the relationship between speech breaks and EF performance in the context of VF has rarely been reported in previous literature, with existing studies tending to rather focus on unfilled pauses in free speech[56]. However, previous findings support our results with respect to the importance of speech breaks within the first interval of the VF task in that previous studies found a relationship between longer latencies in the beginning of the VF task and cognitive flexibility performance[39]. Moreover, other studies suggest that a decrease of speech latencies over the course of the VF task is related to the cluster patterns of the participants. While participants are assumed to produce clusters with high-frequent words in the beginning of the task, less frequent words are produced during the progression of the task leading to more switches and increased searching times[92].

In general, previous studies support the positive relationship between the duration of speech breaks and higher cognitive demands[56]. However, our results revealed mostly non-linear relationships between latencies and significantly predicted EF scores. This might suggest that shorter speech breaks per se do not go along with better EF performance. Rather, we assume that the heterogeneity of searching strategies, including processes such as clustering and switching, lead to ambiguous latency patterns.

Semantic relatedness. Investigating the role of semantic relatedness between produced words within the sVF task, two different semantic analysis systems were applied. On the one hand, a hierarchical approach was used (*GermaNet*)[52]. On the other hand, an approach based on word embeddings was applied (DISCO)[51]. The main goal of including both approaches was to assess as much diverse semantic information as possible. Our results revealed that semantic relatedness measures from both semantic systems contribute essentially to the prediction of all EF domains as well as to attention performance. Although not all semantic features revealed a linear relationship with EF performance, results indicate that searching for closely related words might be related to stronger EF test results.

These findings are partially in line with previous studies which apply earlier approaches of cluster and switching quantification to investigate the importance of switches in the sVF task[87]. Authors have found a positive relationship between fewer switches and better cognitive flexibility performance in healthy participants[87]. In contrast, other studies reported a decreased number of switches in depressive patients with reduced cognitive flexibility[93]. Although the present study did not differentiate between the two types of switches[42], the semantic systems applied in this study[51,52] provided additional semantic distances which are similarly interpretable. In detail, these semantic measurements also quantify semantic distances of sequential and cumulative word pairs. Thus, a higher semantic mean in the present study can be equated to a higher cluster size and less hard switches. However, the present study did not aim to investigate such a fine-grained semantic approach as Troyer´s[41] approach but rather strived to investigate the general importance of semantic distances within the sVF task.

In general, we assume that the production of semantically distanced words puts higher demands on cognitive processes. However, for the sVF task, participants are asked to simply produce as many words as possible, with no demands on the number of different subcategories these words come from. Thus, producing closely related words and building high cluster sizes might represent the most efficient strategy of successful EF performers.

Superiority of advanced sVF feature set. While the full feature set of sum scores, errors, latencies and semantic relatedness was applied for the main analysis, we also predicted EF scores using sum score features only. Using the sophisticated feature set, test variables from all EF domains as well as attention performance and 8/14 EF tests were successfully predicted. While many of the predictable EF scores contained general information of processing speed and reaction times, results also comprised EF scores which are considered as characteristic variables for specific EF tests. For example, TMT is represented by the *difference between part A-B*[94], Stroop by *naming interference*[71], SPM by *correct items*[62] and WCST by *non-perseveration errors*[58], all of which were found to be predictable EF scores.

In contrast, analysis with a classical sVF feature solely containing information of the sum scores, predicted only 6/14 EF tests most of which were related to general processing speed rather than to specific EF functions. Only one EF score of NVLT contained characteristic information of working memory performance. EF scores representing cognitive flexibility and inhibitory performance did not include information which are directly linked to EF performance but rather related to general speed.

To our knowledge, so far, no other study has attempted combining different types of sVF measurements to predict EF scores. However, previous research has demonstrated the advantages of advanced approaches evaluating additional information over the sole use of the total number of correctly produced words. For example, it was shown that the switching sVF task, which was also used in the present study (t_3), contained more information of cognitive flexibility than simple VF tasks[87]. Our findings are also in line with another study investigating the digitalized evaluation of semantic relatedness with WordNet[50]. In particular, semantic relatedness was found to be highly associated with EFs and serve as an indicator for mild cognitive impairments which are difficult to detect with sum scores[95].

The comparison of prediction analysis with and without an extended set of sVF features mainly indicated that sum scores alone capture mostly working memory performance and attention scores. On the other hand, an advanced sVF feature set including sum scores, errors, latencies and sematic relatedness allows for the prediction of cognitive flexibility, working memory and inhibition performances as well as attention scores.

Investigating the relationship between the most important sVF features and EF performance in more detail, multiple non-linear relationships were detected. These findings highlight the advantages of machine learning approaches which are able to detect complex, non-linear relationships in addition to straightforward linear ones.

Also, these approaches can take into account multivariate interactions between different VF features to reveal patterns which could not have been identified based on each single feature alone.

In general, our findings indicate that the use of a comprehensive set of VF features might have the potential to replace time-consuming and artificial EF tests. Due to the use of abstract symbols like numbers and letters, commonly used neuropsychological tests are criticized for their lacking ecological validity[11]. In contrast, producing words which are related to a specific category better represents daily needs and requirements of participants. Moreover, the lack of ecological validity might have influenced the correlations of the abstract EF test scores and the more natural sVF features. However, it remains open whether comprehensive sVF features may be even more helpful in clinical practice than commonly used EF test batteries.

Role of processing speed. In both analyses, variables which are not directly linked to EF performance but rather represent overall processing speed or reaction times, were predicted significantly. Similar findings were reported in our previous study predicting VF sum scores from EF tests variables[96]. The relationship of processing speed and sVF performance is also reported in other studies[83,97]. These authors suggest that processing time reflects general cognitive abilities such as intelligence to some extent[98] but may also be related to age[99] or personality traits such as extraversion[100]. Additionally, the presence of a time indication within some EF tests might facilitate processing speed similarly as in sVF tasks.

Limitations. Our results yielded insights into the involvement of EFs in the sVF task and highlighted the informative value of the sVF task to predict EF performance using a comprehensive feature set. Moreover, our results revealed complex and mostly non-linear relationships of VF features and EF performance. Hence, a detailed examination of individual differences in searching strategies might improve our understanding of which sVF patterns are related to higher EF performance in certain domains. As with all analyses of individual differences such research is dependent on large data sets comprising detailed information on EF and VF performance.

An additional consideration relates to the generalizability of our results. Ideally, our findings should be validated in a fully independent data set. To date, such a data set of sufficient size is not yet available. Hence, we applied a cross-validation approach within our sample. Here, the model was trained on some parts of the data while other parts of the data were held back. The model was then validated in the previously held back participants. This within-dataset validation represents the best alternative when a fully independently acquired dataset is not yet available.

Summary and outlook. Our study revealed insights into the advantages of an elaborated analysis of sVF tasks which successfully predicts EF performance. In comparison to the commonly used approach of evaluating sum scores of correctly produced words, we detected a lucid advantage of an extended feature analysis. In particular with regards to cognitive flexibility and inhibition our study demonstrated that an evaluation of sVF sum scores does not capture actual EF performance but rather assesses overall processing speed. Thus, we suggest the utilization of a comprehensive analysis of VF performance including features of error types, latencies and semantic distances. The present study applied primarily automated and digitalized methods ensuring a time-efficient and objective evaluation of VF performance. Further studies ought to develop a fully automated software tool integrating and further developing our feature set. Here, it would be highly interesting to also include features from the lexical VF task. A computerized toolbox allowing for an extensive assessment of VF could serve as a screening tool for EFs in a clinical diagnostic process as well as in a research context. Such a tool could include an audio system that records the speech of the patient and converts it into text. Subsequently, an automated software could be used to automatically determine a comprehensive set of VF features including sum scores, errors, latencies and semantic distances from the transcribed data. This can in turn result in a digitalized and quantified evaluation of the patient´s EFs compared to healthy controls based on VF performance, which can be then used by the clinician as part of the diagnostic process. Consequently, this toolbox could allow for higher ecological validity while also saving time in clinical routine.

However, we do not suggest that VF assessments will be able to fully substitute an initial extensive assessment of EFs with commonly used EF test batteries. We rather propose an extended and fully digitalized VF analysis as part of progress diagnostics in the form of a screening to assess EF performance in e.g. Parkinson´s disease or ADHD. Additionally, this screening-tool could be used in patients with predispositions of schizophrenia before manifestation of clinical symptoms. Here, an advanced sVF analysis could provide insights into subtle changes of EF performance. In the future, this work might contribute to an automated digitalized speech analysis supporting clinicians in diagnostic processes.

Altogether, the present study demonstrated the predictive superiority of an extended VF feature evaluation. Additionally, the results provided a first step towards an automated analysis of VF serving as a predictor for EFs.

Received: 17 November 2020; Accepted: 9 March 2021
Published online: 25 March 2021

References

1. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
2. Diamond, A. Executive functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
3. Karr, J. E. *et al.* The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychol. Bull.* **144**, 1147–1185 (2018).

4. Kudlicka, A., Clare, L. & Hindle, J. V. Executive functions in Parkinson's disease: Systematic review and meta-analysis. *Mov. Disord.* **26**, 2305–2315 (2011).
5. Umarova, R. M. *et al.* Cognitive reserve impacts on disability and cognitive deficits in acute stroke. *J. Neurol.* **266**, 2495–2504 (2019).
6. Tavares, J. V. T. *et al.* Distinct profiles of neurocognitive function in unmedicated unipolar depression and bipolar II depression. *Biol. Psychiatry* **62**, 917–924 (2007).
7. Nigg, J. T., Blaskey, L. G., Huang-pollock, C. L. & Rappley, M. D. Neuropsychological executive functions and DSM-IV ADHD subtypes. *J. Am. Acad. Child Adolesc. Psychiatry* **41**, 59–66 (2002).
8. Stuss, D. T. & Alexander, M. P. Executive functions and the frontal lobes: a conceptual view. *Psychol. Res.* **63**, 289–298 (2000).
9. Fine, E. M. & Delis, D. C. Delis-Kaplan executive functioning system. In *Encyclopedia of Clinical Neuropsychology* (eds Kreutzer, J. S. *et al.*) 796–801 (Springer, New York, 2011).
10. *Wiener Testsystem*. (SCHUHFRIED GmbH, 2016).
11. Chan, R., Shum, D., Toulopoulou, T. & Chen, E. Assessment of executive functions: Review of instruments and identification of critical issues. *Arch. Clin. Neuropsychol.* **23**, 201–216 (2008).
12. Baddeley, A. Working memory and language: An overview. *J. Commun. Disord.* **36**, 189–208 (2003).
13. Levelt, W. J. Accessing words in speech production: Stages, processes and representations. *Cognition* **42**, 1–22 (1992).
14. Ho, A. K. *et al.* Verbal fluency in Huntington's disease: A longitudinal analysis of phonemic and semantic clustering and switching. *Neuropsychologia* **40**, 1277–1284 (2002).
15. Canning, S. J. D., Leach, L., Stuss, D., Ngo, L. & Black, S. E. Diagnostic utility of abbreviated fluency measures in Alzheimer disease and vascular dementia. *Neurology* **62**, 556–562 (2004).
16. Pakhomov, S. V. S. & Hemmy, L. S. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex* **55**, 97–106 (2014).
17. Weiss, E. M. *et al.* Sex differences in clustering and switching in verbal fluency tasks. *J. Int. Neuropsychol. Soc.* **12**, 502 (2006).
18. Nikolai, T. *et al.* Semantic verbal fluency impairment is detectable in patients with subjective cognitive decline. *Appl. Neuropsychol. Adult* **25**, 448–457 (2018).
19. Gonçalves, H. A. *et al.* Clustering and switching in unconstrained, phonemic and semantic verbal fluency: The role of age and school type. *J. Cogn. Psychol.* **29**, 670–690 (2017).
20. Azuma, T. Working memory and perseveration in verbal fluency. *Neuropsychology* **18**, 69–77 (2004).
21. Rosen, V. M. & Engle, R. W. The role of working memory capacity in retrieval. *J. Exp. Psychol. Gen.* **126**, 211–227 (1997).
22. Fisk, J. E. & Sharp, C. A. Age-related impairment in executive functioning: Updating, inhibition, shifting, and access. *J. Clin. Exp. Neuropsychol.* **26**, 874–890 (2004).
23. Shao, Z., Janse, E., Visser, K. & Meyer, A. S. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* **5**, 772 (2014).
24. Whiteside, D. M. *et al.* Verbal fluency: Language or executive function measure?. *Appl. Neuropsychol. Adult* **23**, 29–34 (2016).
25. Robinson, G., Shallice, T., Bozzali, M. & Cipolotti, L. The differing roles of the frontal cortex in fluency tests. *Brain* **135**, 2202–2214 (2012).
26. Hidalgo-Lopez, E. & Pletzer, B. Interactive effects of dopamine baseline levels and cycle phase on executive functions: The role of progesterone. *Front. Neurosci.* **11**, 403 (2017).
27. Hedden, T. & Yoon, C. Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology* **20**, 511–528 (2006).
28. Fournier-Vicente, S., Larigauderie, P. & Gaonac'h, D. More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. *Acta Psychol. (Amst.)* **129**, 32–48 (2008).
29. Benjamin, M. J., Cifelli, A., Garrard, P., Caine, D. & Jones, F. W. The role of working memory and verbal fluency in autobiographical memory in early Alzheimer's disease and matched controls. *Neuropsychologia* **78**, 115–121 (2015).
30. Patra, A., Bose, A. & Marinis, T. Performance difference in verbal fluency in bilingual and monolingual speakers. *Biling. Lang. Cogn.* https://doi.org/10.1017/S1366728918001098 (2019).
31. Clark, D. G. *et al.* Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **2**, 113–122 (2016).
32. Zhu, Z., Novikova, J. & Rudzicz, F. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *ArXiv180806570 Cs* (2019).
33. Cummins, N., Sethu, V., Epps, J. & Krajewski, J. Relevance vector machine for depression prediction. In *INTERSPEECH* (2015).
34. Bedi, G. *et al.* Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophr.* **1**, 1–7 (2015).
35. van den Berg, E., Jiskoot, L. C., Grosveld, M. J. H., van Swieten, J. C. & Papma, J. M. Qualitative assessment of verbal fluency performance in frontotemporal dementia. *Dement. Geriatr. Cogn. Disord.* **44**, 35–44 (2017).
36. Ruff, R. M., Light, R., Parker, S. B. & Levin, H. S. Benton Controlled Oral Word Association Test: reliability and updated norms. *Arch. Clin. Neuropsychol.* **11**, 329–338 (1996).
37. Aschenbrenner, S., Tucha, O. & Lange, K. W. *Regensburger Wortflüssigkeits-Test: RWT*. (Hogrefe, Verlag für Psychologie, Göttingen).
38. Galaverna, F., Bueno, A. M., Morra, C. A., Roca, M. & Torralva, T. Analysis of errors in verbal fluency tasks in patients with chronic schizophrenia. *Eur. J. Psychiatry* **30**, 305–320 (2016).
39. Wolters, M. K., Kim, N., Kim, J.-H., MacPherson, S. E. & Park, J. C. Prosodic and linguistic analysis of semantic fluency data: A window into speech production and cognition. *Interspeech* https://doi.org/10.21437/Interspeech.2016-420 (2016).
40. Pakhomov, S. V. S., Eberly, L. & Knopman, D. Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia* **89**, 42–56 (2016).
41. Troyer, A. K., Moscovitch, M. & Winocur, G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology* **11**, 138–146 (1997).
42. Abwender, D. A., Swan, J., Bowerman, J. & Connolly, S. Qualitative analysis of verbal fluency output: Review and comparison of several scoring methods. *Assessment* https://doi.org/10.1177/107319110100800308 (2001).
43. Unsworth, N., Spillers, G. J. & Brewer, G. A. Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *Q. J. Exp. Psychol.* **64**, 447–466 (2011).
44. Zhao, Q., Guo, Q. & Hong, Z. Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment. *Neurosci. Bull.* **29**, 75–82 (2013).
45. Price, S. E. *et al.* Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology* **26**, 490–497 (2012).
46. Rich, J. B., Troyer, A. K., Bylsma, F. W. & Brandt, J. Longitudinal analysis of phonemic clustering and switching during word-list generation in Huntington's disease. *Neuropsychology* **13**, 525–531 (1999).
47. Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P. & Stuss, D. Clustering and switching on verbal fluency: The effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia* **36**, 499–504 (1998).
48. *Handbook of latent semantic analysis*. xii, 532 (Lawrence Erlbaum Associates Publishers, 2007).
49. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *ArXiv13104546 Cs Stat* (2013).
50. *WordNet: An Electronic Lexical Database*. (The MIT Press, 1998). https://doi.org/10.7551/mitpress/7287.001.0001.

51. Kolb, P. DISCO: A Multilingual Database of Distributionally Similar Words. 8.
52. Henrich, V. & Hinrichs, E. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* 420–426 (Association for Computational Linguistics, 2011).
53. Jurafsky, D. & Martin, J. H. *Speech and Language Processing (2nd Edition)*. (Prentice-Hall, Inc., 2009).
54. Pauselli, L. *et al.* Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry Res.* **263**, 74–79 (2018).
55. Clark, H. H. Managing problems in speaking. *Speech Commun.* **15**, 243–250 (1994).
56. Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F. & Brennan, S. E. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Lang. Speech* **44**, 123–147 (2001).
57. Stoet, G. PsyToolkit: A software package for programming psychological experiments using Linux. *Behav. Res. Methods* **42**, 1096–1104 (2010).
58. Bowden, S. C. *et al.* The reliability and internal validity of the wisconsin card sorting test. *Neuropsychol. Rehabil.* **8**, 243–254 (1998).
59. Cohen, A.-L., Bayer, U. C., Jaudas, A. & Gollwitzer, P. M. Self-regulatory strategy and executive control: Implementation intentions modulate task switching and Simon task performance. *Psychol. Res.* **72**, 12 (2006).
60. Mackworth, N. H. The breakdown of vigilance during prolonged visual search. *Q. J. Exp. Psychol.* **1**, 6–21 (1948).
61. Reitan, R. M. Validity of the trail making test as an indicator of organic brain damage. *Percept. Mot. Skills* **8**, 271–276 (1958).
62. Raven, J. C., Raven, J. & Court, J. H. *SPM Manual (Deutsche Bearbeitung und Normierung von St. Bulheller und H. Häcker)*. (Swets & Zeitlinger B.V.).
63. Grant, D. A. & Berg, E. A. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *J. Exp. Psychol.* **38**, 404–411 (1948).
64. Kaller, C. P., Unterrainer, J. M. & Stahl, C. Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychol. Assess.* **24**, 46–53 (2012).
65. Meiran, N. Reconfiguration of processing mode to task performance. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 1423–1442 (1996).
66. Schellig, D., Schuri, U. & Arendasy, M. *NBN- NBACK-nonverbal*. (SCHUHFRIED GmbH, 2009).
67. Sturm, W. & Willmes, K. *NVLT Non-Verbal Learning Test*. (SCHUHFRIED GmbH, 2016).
68. Schellig, D. & Hättig, H. A. Die Bestimmung der visuellen Merkspanne mit dem Block-Board. *Z. Für Neuropsychol.* **4**, 104–112 (1993).
69. Kaiser, S., Aschenbrenner, S., Pfüller, U., Roesch-Ely, D. & Weisbrod, M. *Response Inhibition*. (SCHUHFRIED GmbH, 2016).
70. Simon, J. R. & Wolf, J. D. Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics* **6**, 99–105. https://doi.org/10.1080/00140136308930679 (1963).
71. Schuhfried, G. *Interferenz nach Stroop*. (SCHUHFRIED GmbH, 2016).
72. Sturm, W. *Wahrnehmungs- und Aufmerksamkeitsfunktionen: Geteilte Aufmerksamkeiten*. (SCHUHFRIED GmbH, 2016).
73. Benton, A. L., Hamsher, S. K. & Sivan, A. B. *Multilingual aplasia examination* (AJA Associates, Iowa city, 1983).
74. *Presentation software*. (Neurobehavioral Systems, 2018).
75. Boersma, P. Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2002).
76. Cramer, I. M. & Finthammer, M. Tools for exploring GermaNet in the context of cl-teaching. *KONVENS* https://doi.org/10.1515/9783110211818.3.195 (2008).
77. Barsukova, A. *et al. Tutorial: Semantic Relatedness API for GermaNet*. (University of Tübingen; Department of General and Computational Linguistics, 2018).
78. Kolb, P. *Experiments on the Difference Between Semantic Similarity and Relatedness*. 8.
79. Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* **1**, 278–282 (1995).
80. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
81. Denil, M., Matheson, D. & De Freitas, N. Narrowing the Gap: Random Forests in Theory and in Practice. In *Proceedings of the 31st International Conference on International Conference on Machine Learning: Volume 32* I-665-I–673 (JMLR.org, 2014).
82. Tombaugh, T. N., Kozak, J. & Rees, L. Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Arch. Clin. Neuropsychol.* **14**, 167–177 (1999).
83. Elgamal, S. A., Roy, E. A. & Sharratt, M. T. Age and verbal fluency: The mediating effect of speed of processing. *Can. Geriatr. J.* **14**, 66–72 (2011).
84. Lanting, S., Haugrud, N. & Crossley, M. The effect of age and sex on clustering and switching during speeded verbal fluency tasks. *J. Int. Neuropsychol. Soc.* **15**, 196–204 (2009).
85. Obeso, I., Casabona, E., Bringas, M. L., Alvarez, L. & Jahanshahi, M. Semantic and phonemic verbal fluency in Parkinson's disease: Influence of clinical and demographic variables. *Behav. Neurol.* **25**, 111–118 (2012).
86. Mathworks. *Predictor Importance Estimates by Permutation of Out-of-Bag Predictor Observations for Random Forest of Regression Trees—MATLAB*. https://www.mathworks.com/help/stats/regressionbaggedensemble.oobpermutedpredictorimportance.html.
87. Paula, J. J., Paiva, G. C. & Costa, D. D. Use of a modified version of the switching verbal fluency test for the assessment of cognitive flexibility. *Dement. Neuropsychol.* **9**, 258–264 (2015).
88. Ardila, A., Galeano, L. M. & Rosselli, M. Toward a model of neuropsychological activity. *Neuropsychol. Rev.* **8**, 171–190 (1998).
89. Carr, M., Saint-Onge, K., Blanchette-Carrière, C., Paquette, T. & Nielsen, T. Elevated perseveration errors on a verbal fluency task in frequent nightmare recallers: A replication. *J. Sleep Res.* **27**, e12644 (2018).
90. Fischer-Baum, S., Miozzo, M., Laiacona, M. & Capitani, E. Perseveration during verbal fluency in traumatic brain injury reflects impairments in working memory. *Neuropsychology* **30**, 791–799 (2016).
91. Suhr, J. A. & Jones, R. D. Letter and semantic fluency in Alzheimer's, Huntington's, and Parkinson's dementias. *Arch. Clin. Neuropsychol.* **13**, 447–454 (1998).
92. Raboutet, C. *et al.* Performance on a semantic verbal fluency task across time: Dissociation between clustering, switching, and categorical exploitation processes. *J. Clin. Exp. Neuropsychol.* **32**, 268–280 (2010).
93. Fossati, P., Bastard Guillaume, L., Ergis, A.-M. & Allilaire, J.-F. Qualitative analysis of verbal fluency in depression. *Psychiatry Res.* **117**, 17–24 (2003).
94. Kortte, K. B., Horner, M. D. & Windham, W. K. The trail making test, part B: cognitive flexibility or ability to maintain set?. *Appl. Neuropsychol.* **9**, 106–109 (2002).
95. Pakhomov, S. V. S., Hemmy, L. S. & Lim, K. O. Automated semantic indices related to cognitive function and rate of cognitive decline. *Neuropsychologia* **50**, 2165–2175 (2012).
96. Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S. & Weis, S. Executive functions predict verbal fluency scores in healthy participants. *Sci. Rep.* **10**, 11141 (2020).
97. Brébion, G. *et al.* Verbal fluency in male and female schizophrenia patients: Different patterns of association with processing speed, working memory span, and clinical symptoms. *Neuropsychology* **32**, 65–76 (2018).
98. Kail, R. & Salthouse, T. A. Processing speed as a mental capacity. *Acta Psychol. (Amst.)* **86**, 199–225 (1994).
99. Sliwinski, M. & Buschke, H. Cross-sectional and longitudinal relationships among age, cognition, and processing speed. *Psychol. Aging* **14**, 18–33 (1999).
100. Pearman, A. Basic cognition in adulthood: Combined effects of sex and personality. *Personal. Individ. Differ.* **47**, 357–362 (2009).

Acknowledgements
This research was supported by The Deutsche Forschungsgemeinschaft (DFG, EI 816/11-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1) 785907 (HBP SGA2). We are very grateful to Lisa Mochalski, Natalie Schlothauer and Hannah Hensen for their help with testing participants. We also thank Tobias Kadelka for his help with the automated digitalization of the VF evaluation.

Author contributions
The study was designed by all authors of the manuscript. In particular, the executive functions test selection was mainly supported by J.A.C. while S.H expertise mainly contributed to the discussion of speech-related topics. Data collection, data preprocessing and generation of VF features was done by J.A. Data analysis was mainly driven by S.W, K.R.P., J.A.C., J.A. and S.B.E.; Here, K.R.P. knowledge of different machine learning approaches and their implementation in *Matlab* was essential. Moreover, S.W. and S.B.E. contributed their knowledge of advanced data science. G.P. contributed contents relating to the psychiatric context as well as the influences of processing speed and intelligence on cognitive performance in psychiatric patients. The manuscript was mainly written by J.A., supervised throughout the writing process by S.W. and J.A.C. who provided improvements for structuring the manuscript as well as improving the wording. Moreover, S.B.E., K.R.P., S.H. and G.P. provided detailed feedback on the overall manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85981-1.

Correspondence and requests for materials should be addressed to J.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5 Discussion

The three studies presented in this thesis showed the potential of the VF task predicting fine-grained EF performance and provided evidence for the use of the VF task serving as a differentiated and time-efficient screening for EFs. Particularly, EF test scores tapping into all subdomains i.e., cognitive flexibility, working memory, and inhibition were predicted successfully from a comprehensive set of VF features. To enable that, common structures of different EF tests and VF tasks were identified to gain first insight into the involvement of EFs in the VF task. Results showed the involvement of rather simple aspects of the different EF domains and VF performance. Moreover, results of study 2 demonstrated that relationships of EFs and VF performance are just partially linear and not fully depictable applying correlation analysis. Instead, ML methods are needed to capture the complex relationships of differentiated EFs and VF task performance.

5.1 The role of the verbal fluency task in executive function test batteries

Although the VF task has been designed to test verbal functionating[87], there is consensus that the VF task is as well a valid instrument for testing EFs. Thus, many neuropsychological test batteries such as the D-KEFS battery include VF tasks. However, literature reports ambiguous results of the concrete involvement of different EF subdomains in the VF tasks. It is therefore not totally clear whether the VF test results are able to represent the full variety of EF performance. To discuss the present results of the three studies in a broader context, links to different diseases are provided representing different EF impairments.

5.1.1 The relationship of verbal fluency and executive functions

Simple aspects of EFs in the VF task

In all studies the involvement of EFs in the VF task was shown. Particularly, the strong inclusion of simple aspects of EFs and the role of attention were highlighted. The studies differed, however, in their concrete aims and settings. For example, studies 2 and 3 did not include abstraction tasks. In the first study, the D-KEFS battery was investigated including different EF tests as well as lexical, semantic, and switching VF tasks. Results of study 1 identified two factors, differentiating 9 different EF tests of the D-KEFS test battery. Results

of the ML approach as well as of the traditional dimension reduction demonstrated a subdivision of the EF tests in a simple and a more complex factor. The first factor included all three VF tasks as well as *Colour Word Interference* (CWI), design fluency and the TMT. Thus, the tests of the first factor mainly represented the subdomains of cognitive flexibility, attention, and inhibition. The second, more complex factor included the WCST, *Proverbs Test*, *Word Context Test*, *Twenty Questions Test* and *Tower of London* (TOL). The main cognitive requirements involved in the second factor were consequently abstract thinking and problem-solving abilities.

The detected affiliation of more simple EF abilities and VF performance is also supported by the results of study 2 and study 3. Beside the concrete relationship of specific EF tests and VF performance, an additional insight was obtained highlighting the overall role of processing speed and attention that impacted especially simple EF abilities. Here, study 3 revealed a high number of EF processing and reaction time labels predicted by VF performance, particularly in the TMT assessing cognitive flexibility and attention. Interestingly, the TMT, also named *Number-Letter Switching*, was identified in the same factor of simple EFs than in the VF tasks in study 1.

The close link of reaction times, attention and EFs was already investigated in ADHD patients[88]. The study has assessed both the lexical and semantic VF task in children and compared results with EF and attention performance. It has found significantly lower scores in the ADHD group in the lexical but not in the semantic VF tasks compared to healthy controls[88]. Although inattentiveness plays an essential role in ADHD adults[89], it has been seen as difficult to define a concrete neuropsychological profile in this patient group[90].

Besides the influence of attention in processing speed and EFs, studies also investigated the influence of state-dependent effects. Particularly, the effects of mood[91] and reward-related manipulations of motivation were identified to essentially influence EF test performance[92].

Due to the strong link between processing/reaction times and simple EF tasks across all our three studies, we assume that this aspect of EF, as well as of the VF task, plays a central role in the mutual connection of EFs and VF performance. One reason for this close relationship might be because both are influenced by the general concentration and motivation of the participant. In detail, attentive, motivated, and ambitious participants might try to solve EF and VF tasks as quickly as possible. Simple EF tasks with less cognitive complexity might better allow for fast reaction times than more complex tasks. Thus, the attention and motivation of the patient should be taken into account when evaluating EF and VF tasks.

Cognitive flexibility

While there was a conclusive link between inhibitory EF tests and VF performance, the findings concerning cognitive flexibility and VF performance were ambiguous to some degree across the three studies. Study 2 demonstrated that the VF sum score could be predicted from cognitive flexibility tests. Additionally, study 3 predicted cognitive flexibility performance from VF information. In particular, SPM, TMT and WCST scores were predicted successfully, or where mainly involved to predict the VF sum score. Interestingly, study 1 assigned most cognitive flexibility tests to a different factor than the VF task. The *Sorting tests* (equivalent to the WCST) as well as the *Tower tests* (equivalent to the TOL) were assigned to the complex EF factor and not to the simple factor that included the VF tests. At the same time, the *Number-Letter Switching* task (equivalent to the TMT) was associated to the simple EF factor.

The strong link between cognitive flexibility and VF performance is well known[11]. It represents just one reason why the VF task in the past was implemented in several EF test batteries, e.g. to assess cognitive flexibility in patients with anorexia[93] or obsessive-compulsive disorder[94]. However, the conceptual discussion of the involved domains of EFs in the respective EF test led to an impeded interpretation of the concrete test results. For example, previous studies partially support our findings with regards to the TOL. Some studies suggest that the TOL should not be used as a planning task due to the strong involvement of *updating* processes that are primarily linked to working memory performance[95,96]. Besides the complexity of EF test constructs, study results seem to vary depending on the specific semantic VF task. For example, a study assessing the TMT as a cognitive flexibility measure to investigate the relationship to different semantic VF tasks in healthy adults found diverse results depending on the specific VF task[59]. In the present work, correlations of the TMT score and the semantic VF task *animals,* as well as the semantic VF switching task, were significant. In contrast, the correlation of the TMT and the semantic VF task *fruits* was not significant[59]. Therefore, our results in this respect are in line with the literature.

The present results provide evidence that VF performance better captures simple cognitive flexibility abilities. The level of complexity should be considered when selecting the right EF test assessing cognitive flexibility as well as when choosing the specific category assessed in the VF task. While studies testing patients with mild EF impairments or healthy controls, high-frequent categories, such as *animals*, might not lead to a successful indication of present EF deficits. Additionally, we suggest considering the involvement of multiple EF domains within the same task and critically reflect the explicit task used to potentially discover further EF deficits, such as working memory performance.

Inhibition

In the present three studies, the demand of inhibitory processes in the VF task was shown across all three studies. Particularly, the cognitive demands in the *Stroop* test appear similarly to inhibitory demands in the VF task. The concrete involvement of inhibitory processes in the VF tasks was compelling in all studies. Study 1 demonstrated that the VF tasks load on the same factor as the inhibitory tests, i.e., assessed with the *design inhibition* task and the CWI. Similarly, study 2 results showed that test scores from the *Stroop* test (which is a synonym for the CWI test and assesses inhibitory performance) played an essential role to predict the VF sum score. Additionally, study 3 demonstrated the predictive power of a comprehensive set of VF features to predict *Stroop* performance.

The results of these three studies partially differ from those in previous literature dealing with the link between inhibitory processes and VF performance. Investigating healthy participants, Escobar et al.[97] showed that although bilingual children performed better in VF tasks, inhibitory control did not differ from monolingual children. In contrast, another study investigating inhibitory control in bilingual adults found significant positive correlation between inhibition and VF performance in bilingual but not in monolingual speakers[83]. The divergent performance of inhibitory and VF tasks was also discussed in the context of inhibitory processes. Researchers highlighted the different demands of inhibitory processes depending on the specific EF task[65]. In detail, differences of inhibitory control between the stop-signal task, testing *response inhibition*, and the *Stroop*, assessing *selective inhibition*, were mentioned. The latter type of inhibition is meant to also be involved in inhibitory processes in the VF task. Thus, different results might occur depending on the task used in the respective study.

The present studies highlight the predictive power of the VF task with respect to inhibitory performances in healthy participants and indicate the potential of the predictive power of the VF task in patients suffering from inhibitory deficits, e.g., in ADHD or schizophrenia. We suggest that EF test batteries should include additional inhibition tests beside the *Stroop* to gain broader insights into the patient´s EF performance.

Working memory

Working memory has been described as a multi-component model including a phonological loop that is responsible to maintain and manipulate information over a short time period[98]. This ability is known to play an important role within the VF task to remember and suppress already produced words[68].

With respect to the involvement of working memory performance in the VF task, the three present studies are not directly comparable with each other with respect to the relationship of working memory and VF performance due to missing explicit working memory tests (e.g., a span task) in study 1. Working memory tests used in study 2 did not reveal a strong impact on the prediction of the VF sum score. However, study 3 has been designed in such a way that it is differentiating VF performance in more detail. As a result, working memory performance was predicted successfully from VF features.

The present results support some findings from previous literature. A recent study applying ML methods in patients with Multiple Sclerosis demonstrated the involvement of working memory abilities in the VF task[99]. This study similarly used digit span tasks and the *Corsi* test to assess working memory performance as well as detailed information of the VF task (e.g., errors, clusters, switches). Interestingly, this study identified a closer link between working memory and the semantic VF task compared to the lexical VF task[99]. In line with these results, *Troyer* found a closer relationship of working memory and EFs within the semantic VF task compared to the lexical VF task. Notably, he also analysed fine-grained information of the VF tasks such as number of cluster and switches [100]. Discussing the lack of a relationship between VF and working memory, the general involvement of working memory in span tasks should be considered. Here, previous literature described missing differences in working memory performances in young healthy participants when a controlled situation is provided that do not include distraction[101].

In general, previous literature as well as study 3 highlighted the importance of working memory in VF performance. However, the involvement of working memory in the assessed EF span might not completely reflect the involvement of working memory in the VF task. Rather, more distracting and interfering EF tasks, such as the *Stroop* test, might better include similar working memory performance as required in the VF task.

5.1.2 The importance of individual differences

EF test batteries in clinical context usually include standard values with the aim to better differentiate the patient's performance and define a cut-off value distinguishing between healthy people and patients with neuropsychological deficits. For example, the age of the patient is considered to allow for an age-appropriated evaluation of the actual EF and VF task performance. However, additional dependent variables such as sex, education or stress level are

suggested to be taken into account to control further individual differences and measure actual VF performance[62,78]. Particularly the VF task is well known to evoke considerable differences in sexes, favouring women[102] while men usually outperform women in visual-spatial tasks[102]. Furthermore, the influence of stress is reported in many studies and is partially suggested as positively influence cognitive performance in men and women[75,103].

Sexual hormones

Sex-related differences go beyond the mere definition of "males" and "females" but also include hormonal levels. The importance of fine-grained differences is reflected in the scientific discussion of the terminology of *sex* and *gender*, moving from a binary classification towards the attention of biological variation traditionally associated with *sex*[104]. While the term *sex* is supposed to be used classifying individuals according to their reproductive organs, *gender* refers to the individual self-representation of a person. Due to the influence of hormones on physiology as well as on behaviour, hormonal levels play an important role when evaluating human behaviour and cognitive performances.

Study 2 included sex-related hormones as features in the prediction analysis to predict VF performance. Results demonstrated the predictive power of estradiol, the major female sex hormone regulating the female reproductive system[105] as one of the strongest variables impacting the prediction analysis, while *progesterone* and *testosterone* were determined to be far less relevant.

The importance of sex-related hormonal levels was also highlighted in various studies investigating sex differences in men and women[76], or assessing cognitive performance during different menstrual cycle phases[106]. Here, our results of the positive effect of *estradiol* on cognitive performance support findings of previous research. For example, a study investigated the effects of *estrogen* in men and postmenopausal women who received *estrogen* replacement therapy (ERT) assessing fluency tasks, working memory, attention and mood[107]. The study reported that women receiving ERT performed better in the semantic VF task, attention and working memory tests than men. Men performed better in attention and working memory tests than women without ERT. Interestingly, women without ERT outperformed men in the fluency task. Besides the comparison of cognitive performances, this study also investigated the influence of *estradiol* on the mood. They found fewer depressive symptoms in women with ERT compared to the female control group not receiving ERT. Additionally, the authors reported lower anger scores in men with higher *estradiol* levels[107]. In contrast to the findings supporting the positive effect of e*stradiol* on cognitive performance, other studies could not

confirm this effect. For example, *Leeners* et al.[108] investigated the effect of *estradiol* on working memory, attention and complex cognitive functions in women receiving fertility treatment. Significant changes of cognitive performance could not be identified between groups and also were not identified in intra-individual analyses[108].

The present results highlight the need for a differentiated assessment and interpretation beyond binary sex classification. The consideration of fluctuating hormonal levels is highly relevant when assessing cognitive performance in transgenders of transexual individuals but also affects e.g., women's performance in the different phase of the menstrual cycle. Hence, even if a hormonal analysis is not possible due to time or cost restrictions, at least an extended assessment of anamnestic information including gender and hormonal specific questions should be part of a comprehensive testing to avoid misleading interpretation of VF results.

Stress-related hormones

To investigate the influence of stress on VF performance, the stress-related hormone cortisol was assessed. Study 2 demonstrated that *cortisol* was the most important feature and even influenced VF performance to a larger extent than actual EF test scores.

The effects of *cortisol* on cognitive performance have been controversially discussed. On the one hand, positive effects of higher *cortisol* levels were associated with faster reaction times in men with higher *cortisol* levels compared to men without cortisol application[103]. Additionally, a study[75] in which young men completed a stress test known to produce *cortisol* (and did not receive external *cortisol* application) found a positive effect of *cortisol* on working memory performance. However, various studies have found contradictory results and reported a negative effect of *cortisol* on working memory performance at high working memory loads[75,109]. The influence of c*ortisol* on other EF subdomains depends on the respective EF subdomains tested and stress levels investigated. In detail, *cortisol* was shown to increase accuracy in updating flexibility tasks in healthy participants[110]. However, patients suffering from acute stress were associated with decreased performance in switching tasks[111].

Based on the heterogeneity of applied methods and findings investigating the influence of *cortisol* on cognitive performance, we suggest that the effects of *cortisol* depend on different criteria. Firstly, the stress load *per se* might serve as a reasonable factor influencing the increase or decrease of cognitive performance. Specifically, a moderate workload causing a moderate increase of *cortisol* might result in faster reaction times while a high demand of cognitive processes might lead to a decrease in performance. However, a continuous increased stress

level, e.g., caused by the general situation of the participant, might lead to worse performance. Secondly, the type of administration of *cortisol* might influence the effect on cognitive performance. An external application of *cortisol* could lead to different effects than a natural increase in *cortisol* level. A meta-analysis investigating the effects of *cortisol* on different EFs supports this hypothesis and suggest that the time-delay of *cortisol* administration and the impact on EF performance might contribute to different effects[112]. To sum up, the influence of stress might play an essential role in test situations. Therefore, situation-specific stress influences, as well as the general mental situation and workload of the participant, should be taken into account. Moreover, beside different coping strategies of men and women, the concrete type of the evaluated EF test variable should be considered since EF tasks based on reaction times might be differently affected by stress than accuracy-based measurements.

5.2 The potential of a differentiated analysis of the VF task

The relationship of VF and EFs, as well as individual differences in cognitive performance, has also been investigated in the past[81,82]. However, having a more detailed look at the features of VF performance, reveals considerable changes in research approaches. The present studies could take advantages of advanced computational linguistic analysis.

Comprehensive features

Despite the common use of the sum score serving as a basis to interpret EF performances, previous studies showed the advantages of more fine-grained analyses to draw conclusion on specific EF competencies. This thesis particularly highlighted the advantage of a comprehensive VF analysis to assess working memory performance.

While in study 1 and 2 the sum of words was analysed to assess VF performance, study 3 applied a broad set of VF features covering mainly lexical and semantic information. The potential of a comprehensive analysis of VF features was identifiable comparing the results of study 2 and 3, both investigating the relationship of VF and EFs performance. In detail, study 3 investigated the prediction of EF test scores based on both, the sum scores only and a comprehensive set of VF features. Results revealed lower prediction performance with sum score features compared to the full set of comprehensive VF features. In detail, a few main variables of the EF tests were predicted but mainly variables of processing speed and reaction times were predicted successfully. In contrast, the comprehensive VF feature set predicted the main variables of the respective EF test. Additionally, the comprehensive feature set

particularly outperformed the sum score analysis in the domain of cognitive flexibility. Interestingly, study 2 did not reveal a relationship between working memory performance and the VF sum score, while in study 3 working memory scores were predicted successfully.
Having identified the advantages of a comprehensive VF feature set, an additional aim of study 3 was to draw conclusions on the importance of different VF feature types influencing specific domains of EF performance. Here, all types of VF features, representing error types, speech breaks and semantic relatedness contributed similarly to prediction results. The only recognizable pattern of dominating VF feature types was identified within the domain of cognitive flexibility. In particular, error types were found to play an important role for cognitive flexibility test scores. In contrast, error types did not influence prediction results within the EF domain of inhibition.

The general importance of a differentiated analysis of the VF task was investigated in a large number of studies, assessing clustering and switching methods[59,113], error types[68,114,115] and speech breaks between produced words[82]. The present findings of the power of comprehensive VF features to assess cognitive flexibility performance are mainly consistent with previous research, particularly in more recent studies using computational linguistic approaches. In this context, *Pakhomov* et al.[116] investigated fine-grained semantic characteristics in the semantic VF task and revealed a close relationship between semantic VF features, cognitive flexibility and attention, while less associations were found in memory tasks. Further research even demonstrated that the use of computational semantic systems outperforms manual evaluation with respect to characterizing individual differences in Parkinson's patients[117]. Besides the general use of computational semantic systems, the underlying semantic network plays a crucial role in displaying cognitive functioning. *Pakhomov* et al. suggested the use of distributional semantic networks rather than ontology-based approaches[118]. In detail, distributional semantic networks are based on co-occurrences, are scalable, and better capture semantic relatedness than just semantic similarity[118].

We assume that the missing link between VF performance and specific EF subdomains in previous literature might be due to lack of information that could potentially be provided by a more comprehensive analysis of the semantic VF task. The findings of all three studies highlight the need to analyse the VF task with comprehensive features instead of evaluating solely the sum of words. Fine-grained VF features seem to better display the complex involvement of EFs in the VF tasks.

Types of VF tasks

In the present three studies different types of VF tasks i.e., lexical and semantic VF tasks including a switching component were investigated. While study 1 addressed the more general aspect of EFs and the VF task including the lexical and semantic VF test, study 2 and 3 further elaborated the concrete involvement of different EF domains in the semantic VF task only.

The missing link between working memory and EFs in study 2 is consistent with a study investigating the relationship of working memory and VF tasks in children and adults[119]. While the authors identified a relationship of the lexical VF task and working memory, they did not find a link between the semantic VF task and working memory[119]. Moreover, the relationship of the lexical VF task and working memory was only identified in adults and not in children. Azuma et al.[120] investigated the sensitivity of the lexical and semantic VF task as well as the influence of different categories within the semantic VF task in Parkinson's patients. Interestingly, they found that the actual type of the VF task (lexical, semantic) is not the key aspect of differentiating Parkinson's patients from healthy controls but rather the difficulty level of the assessed category of the semantic VF is important. Moreover, they found varying influences of the mental status on the lexical and semantic VF tasks[120]. Similarly, *Obeso et al.*[71] highlighted the influence of the stage of illness and educational level rather than the difference in the actual type of the VF task.

Based on the present findings reported in this thesis as well as in previous literature, we suggest focussing on the difficulty level of the respective VF task and assess different categories or letters representing increasing cognitive demands. According to the finding of this thesis that VF performance involves more simple aspects of EFs, an individually adapted difficulty level might better allow for capturing all domains of EFs than assessing different types of the VF task with a similar level of difficulty.

5.3 Advantages of machine learning methods

ML methods are applied in different domains of clinical research e.g., to predict diseases or disease progressions. Specific algorithms are able to detect underlying mechanisms and common structures to create models that even capture complex dependencies. In the present thesis, ML approaches were shown to better reflect the variety of EF domains involved in the VF task compared to classical statistical approaches as well as providing generalisability of results.

Complex relationships between EF and VF tasks

In detail, non-linear relationships as well as valuable predictors were discovered that could not be identified by classical statistical methods. Accordingly, the present three studies highlight the added value of ML methods compared to classical statistical approaches.

Study 1 compared results of ML methods with classical factorization approaches with the aim to investigate common structures of different EF tasks. The results of the different approaches varied in the number and common structures of EF factors. While some classical statistical approaches tended to subdivide EFs in a higher number of factors, representing a heterogenous construct of EFs, the ML approach resulted in a two-factor model. In detail, one factor mainly included EF tests that require switching and monitoring abilities, while the second factor represented complex tasks involving abstraction, problem-solving and abstract thinking. Moreover, study 1 demonstrated that the classical statistical approach was less capable of generalizing results to subsamples.

Study 2 applied a direct comparison of classical and advanced methods investigating the relationship of VF and EFs. In a first step, simple correlation analyses were used to determine linear correlations of specific EF tests and the VF sum score. The correlation analysis revealed significant relationships between the VF sum score and tests of cognitive flexibility, inhibition, and attention. Additionally, *cortisol* was found to correlate with VF performance. However, only the ML approach identified the tremendous influence of *cortisol* on VF performance. Moreover, the ML approach further revealed relationships of working memory scores and VF performance that were not found using correlation analyses. These results indicate that the ML approach was better able capturing complex relationships.

These results are fully in line with literature investigating the use and advantages of ML in studies dealing with the potential of speech characteristics. It has been suggested that the prediction performance of the specific ML approach depends on the study type and the underlying data structure[121]. For example, *Petti* et al.[122] investigated the current potential to predict Alzheimer's disease from speech and language features. In their systematic review the authors described multiple methods that were found to successfully distinguish Alzheimer's patients from healthy controls but also from patients with mild cognitive impairments, e.g. applying support vector machines, neural nets and decision trees[122]. Similarly to our results, studies investigating speech parameters in different diseases defined a complex conglomerate rather than specific speech features as the most influencing parameters[47,123].

In general, we suggest that a fine-grained analysis of the VF task provides enormous potential to reflect cognitive performance in healthy subjects as well as in neurologic or psychiatric patients. Compared to the analysis of the VF sum score only, the high number of extracted features allows for the analysis with ML methods. However, a high number of subjects is needed to make use of cross-validation options and avoid *overfitting* of the model.

Generalisability

Due to the divergent results of previous studies investigating the involvement of the different domains of EFs in the VF task, a lack of generalisation seems to represent a critical aspect in behavioural study results. On the one hand, a limited amount and variety of EF tests lead to a low representation of the complex construct of EF or specific domains. However, the assessment of extensive test batteries is rarely possible due to limited time resources and a decreasing motivation level of the patient. On the other hand, the low number of available speech data sets including EF test results lead to missing replication studies investigated in similar study populations to test generalizability in an independent dataset. Similarly, the present studies were not able to validate results in an independent dataset. However, here the applied cross-validation procedures in all studies provide a solid solution allowing for generalizability. While classical hypothesis-driven approaches build the model based on the whole dataset, cross-validation methods use unseen data. Thus, this approach plays an important role when further developing models for clinical applications. However, clinical research is needed integrating ML-methods into clinical trials to further elaborate patient-specific models taking individual differences into account.

5.4 Conclusion

EF plays an important role in many neurological and psychiatric diseases. Testing the different domains of EFs is often time-consuming and lacks in ecological validity. Consequently, there is a need to develop alternative and more natural EF test. This gap could be filled with specific speech tasks reflecting cognitive performance. The VF task is a common tool to capture first insights into the patient's EF performance. However, only a few studies already investigated the specific predictive potential of the VF task with respect to EF performance. The three studies presented here applied ML methods to investigate the underlying structure of EFs *per se* as well as the relationship of EFs and VF to finally describe the predictive power of the VF task.

Across all three studies, the close relationship of EF tests and the VF task became clear by using ML methods. On the one hand results indicated that the VF tasks requires more simple aspects of EF. On the other hand, the comprehensive and fine-grained analysis highlighted the potential of the VF task to be used as a digitalized screening tool to gain insights into the different domains of EFs in healthy participant and patients. In contrast to the common use of the VF sum score, the present results demonstrated the dominance of a fine-grained and objectively evaluated VF task. However, study results also highlighted the strong influence of individual difference such as sex- and stress related hormones or general attention that should be considered in research and clinical context.

The results provided arguments for a superiority of ML methods compared to classical statistical approaches with respect to a detailed analysis of VF performance. In detail, these methods provided insights into the complex and mainly non-linear relationship of EF performance and VF test scores.

Taken together, these results can be generalized to new data, potentially even to compare patient's data with healthy cognitive performance or track a patient´s decline during disease progression. During the last years, there is an increasing interest in investigating VF with the use of ML methods, particularly to predict neurological diseases[124,125]. Thus, the present studies, as well as further research, contribute to the use of a time-efficient and ecologically valid speech test as a screening for EF performance in a high number of diseases, e.g., as an additional diagnostic instrument in disease monitoring diagnostics.

6 References

1. Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S. & Weis, S. Executive functions predict verbal fluency scores in healthy participants. *Sci. Rep.* **10**, 11141 (2020).
2. Amunts, J. *et al.* Comprehensive verbal fluency features predict executive function performance. *Sci. Rep.* **11**, 6929 (2021).
3. Camilleri, J. A. *et al.* A machine learning approach for the factorization of psychometric data with application to the Delis Kaplan Executive Function System. *Sci. Rep.* **11**, 16896 (2021).
4. Anderson, S. *Languages: A Very Short Introduction*. (Oxford University Press, 2012). doi:10.1093/actrade/9780199590599.001.0001.
5. Schwering, S. C. & MacDonald, M. C. Verbal Working Memory as Emergent from Language Comprehension and Production. *Front. Hum. Neurosci.* **14**, 68 (2020).
6. Levelt, W. J. Accessing words in speech production: stages, processes and representations. *Cognition* **42**, 1–22 (1992).
7. Harris, C. L. Language and Cognition. in *Encyclopedia of Cognitive Science* (American Cancer Society, 2006). doi:10.1002/0470018860.s00559.
8. Hagoort, P. MUC (Memory, Unification, Control) and beyond. *Front. Psychol.* **4**, (2013).
9. Baddeley, A. Working memory and language: an overview. *J. Commun. Disord.* **36**, 189–208 (2003).
10. Alvarez, J. A. & Emory, E. Executive Function and the Frontal Lobes: A Meta-Analytic Review. *Neuropsychol. Rev.* **16**, 17–42 (2006).
11. Diamond, A. Executive functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
12. Tavares, J. V. T. *et al.* Distinct profiles of neurocognitive function in unmedicated unipolar depression and bipolar II depression. *Biol. Psychiatry* **62**, 917–924 (2007).
13. Miller, H. V., Barnes, J. C. & Beaver, K. M. Self-control and health outcomes in a nationally representative sample. *Am. J. Health Behav.* **35**, 15–27 (2011).
14. Duncan, G. J. *et al.* School readiness and later achievement. *Dev. Psychol.* **43**, 1428–1446 (2007).
15. Moffitt, T. E. *et al.* A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci.* **108**, 2693–2698 (2011).
16. Stopford, C. L., Thompson, J. C., Neary, D., Richardson, A. M. T. & Snowden, J. S. Working memory, attention, and executive function in Alzheimer's disease and frontotemporal dementia. *Cortex* **48**, 429–446 (2012).
17. Martel, M., Nikolas, M. & Nigg, J. T. Executive Function in Adolescents With ADHD. *J. Am. Acad. Child Adolesc. Psychiatry* **46**, 1437–1444 (2007).
18. C. Delis, D., Kaplan, E. & H. Kramer, J. *Delis-Kaplan executive function system (D-KEFS)*. (2001).
19. *Wiener Testsystem*. (SCHUHFRIED GmbH, 2016).
20. Berg, E. A. A Simple Objective Technique for Measuring Flexibility in Thinking. *J. Gen. Psychol.* **39**, 15–22 (1948).
21. Brown, R. R. & Partington, J. E. Short Articles and Notes: The Intelligence of the Narcotic Drug Addict. *J. Gen. Psychol.* **26**, 175–179 (1942).
22. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: Individual

differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
23. Baddeley, A. *Working memory, thought, and action*. xviii, 412 (Oxford University Press, 2007). doi:10.1093/acprof:oso/9780198528012.001.0001.
24. Wright, H. H., Downey, R. A., Gravier, M., Love, T. & Shapiro, L. P. Processing distinct linguistic information types in working memory in aphasia. *Aphasiology* **21**, 802–813 (2007).
25. Cragg, L. & Nation, K. Go or no-go? Developmental improvements in the efficiency of response inhibition in mid-childhood. *Dev. Sci.* **11**, 819–27 (2008).
26. Verbruggen, F. & Logan, G. D. Response inhibition in the stop-signal paradigm. *Trends Cogn. Sci.* **12**, 418–424 (2008).
27. Stroop, J. R. Studies of interference in serial verbal reactions. *J. Exp. Psychol.* **18**, 643–662 (1935).
28. Chan, R., Shum, D., Toulopoulou, T. & Chen, E. Assessment of executive functions: Review of instruments and identification of critical issues. *Arch. Clin. Neuropsychol.* **23**, 201–216 (2008).
29. Creavin, S. T. *et al.* Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *Cochrane Database Syst. Rev.* CD011145 (2016) doi:10.1002/14651858.CD011145.pub2.
30. Reitan, R. M. Validity of the Trail Making Test as an Indicator of Organic Brain Damage. *Percept. Mot. Skills* **8**, 271–276 (1958).
31. Duncan, J., Emslie, H., Williams, P., Johnson, R. & Freer, C. Intelligence and the Frontal Lobe: The Organization of Goal-Directed Behavior. *Cognit. Psychol.* **30**, 257–303 (1996).
32. Burgess, P. Theory and Methodology in Executive Function Research. *Burgess PW 1997 Theory Methodol. Exec. Funct. Res. Rabbitt P Ed Theory Methodol. Front. Exec. Funct. Psychol. Press East Sussex UK Pp81 - 116 ISBN 9780863774857* (1997).
33. Lezak, M. D. The Problem of Assessing Executive Functions. *Int. J. Psychol.* **17**, 281–297 (1982).
34. Zelazo, P. D., Carter, A., Reznick, J. S. & Frye, D. Early Development of Executive Function: A Problem-Solving Framework. *Rev. Gen. Psychol.* **1**, 198–226 (1997).
35. Mendoza-Halliday, D., Torres, S. & Martinez-Trujillo, J. Chapter 13 - Working Memory Representations of Visual Motion along the Primate Dorsal Visual Pathway. in *Mechanisms of Sensory Working Memory* (eds. Jolicoeur, P., Lefebvre, C. & Martinez-Trujillo, J.) 159–169 (Academic Press, 2015). doi:10.1016/B978-0-12-801371-7.00013-2.
36. Miyake, A. *et al.* The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognit. Psychol.* **41**, 49–100 (2000).
37. Tiego, J., Testa, R., Bellgrove, M. A., Pantelis, C. & Whittle, S. A Hierarchical Model of Inhibitory Control. *Front. Psychol.* **9**, 1339 (2018).
38. Zokaei, N. & Husain, M. Working Memory in Alzheimer's Disease and Parkinson's Disease. in *Processes of Visuospatial Attention and Working Memory* (ed. Hodgson, T.) 325–344 (Springer International Publishing, 2019). doi:10.1007/7854_2019_103.
39. Chen, S. T., Sultzer, D. L., Hinkin, C. H., Mahler, M. E. & Cummings, J. L. Executive Dysfunction in Alzheimer's Disease. *J. Neuropsychiatry Clin. Neurosci.* **10**, 426–432 (1998).

40. Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J. & Pakaski, M. Speaking in Alzheimer's Disease, is That an Early Sign? Importance of Changes in Language Abilities in Alzheimer's Disease. *Front. Aging Neurosci.* **7**, (2015).
41. Trapp, W. *et al.* Speed and capacity of working memory and executive function in schizophrenia compared to unipolar depression. *Schizophr. Res. Cogn.* **10**, 1–6 (2017).
42. Çokal, D. *et al.* The language profile of formal thought disorder. *Npj Schizophr.* **4**, 1–8 (2018).
43. Kuperberg, G. R. Language in schizophrenia Part 1: an Introduction. *Lang. Linguist. Compass* **4**, 576–589 (2010).
44. Cohen, A. S. & Elvevåg, B. Automated computerized analysis of speech in psychiatric disorders. *Curr. Opin. Psychiatry* **27**, 203–209 (2014).
45. Boschi, V. *et al.* Connected Speech in Neurodegenerative Language Disorders: A Review. *Front. Psychol.* **8**, (2017).
46. Mundt, J. C., Vogel, A. P., Feltner, D. E. & Lenderking, W. R. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biol. Psychiatry* **72**, 580–587 (2012).
47. Fraser, K. C., Meltzer, J. A. & Rudzicz, F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *J. Alzheimers Dis. JAD* **49**, 407–422 (2016).
48. Cottingham, M. E. & Hawkins, K. A. Verbal fluency deficits co-occur with memory deficits in geriatric patients at risk for dementia: Implications for the concept of mild cognitive impairment. *Behav. Neurol.* **22**, 73–79 (2010).
49. Hurford, I. M., Ventura, J., Marder, S. R., Reise, S. P. & Bilder, R. M. A 10-minute measure of global cognition: Validation of the Brief Cognitive Assessment Tool for Schizophrenia (B-CATS). *Schizophr. Res.* **195**, 327–333 (2018).
50. Kogan, E. *et al.* Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med. Inform. Decis. Mak.* **20**, 8 (2020).
51. Yassuda, M. S. *et al.* Normative data for the Brief Cognitive Screening Battery stratified by age and education. *Dement. Neuropsychol.* **11**, 48–53 (2017).
52. Aschenbrenner, S., Tucha, O. & Lange, K. W. *Regensburger Wortflüssigkeits-Test: RWT*. (Hogrefe, Verlag für Psychologie, Göttingen).
53. Diaz, M., Sailor, K., Cheung, D. & Kuslansky, G. Category size effects in semantic and letter fluency in Alzheimer's patients. *Brain Lang.* **89**, 108–114 (2004).
54. Sauzéon, H., Lestage, P., Raboutet, C., N'Kaoua, B. & Claverie, B. Verbal fluency output in children aged 7-16 as a function of the production criterion: qualitative analysis of clustering, switching processes, and semantic network exploitation. *Brain Lang.* **89**, 192–202 (2004).
55. Rosen, V. M. & Engle, R. W. The role of working memory capacity in retrieval. *J. Exp. Psychol. Gen.* **126**, 211–227 (1997).
56. Hirshorn, E. A. & Thompson-Schill, S. L. Role of the left inferior frontal gyrus in covert word retrieval: Neural correlates of switching during verbal fluency. *Neuropsychologia* **44**, 2547–2557 (2006).
57. Troyer, A. K., Moscovitch, M. & Winocur, G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology* **11**, 138–146 (1997).
58. Unsworth, N., Spillers, G. J. & Brewer, G. A. Variation in verbal fluency: A latent

variable analysis of clustering, switching, and overall performance. *Q. J. Exp. Psychol.* **64**, 447–466 (2011).
59. Paula, J. J. de, Paiva, G. C. de C. & Costa, D. de S. Use of a modified version of the switching verbal fluency test for the assessment of cognitive flexibility. *Dement. Neuropsychol.* **9**, 258–264 (2015).
60. Fisk, J. E. & Sharp, C. A. Age-Related Impairment in Executive Functioning: Updating, Inhibition, Shifting, and Access. *J. Clin. Exp. Neuropsychol.* **26**, 874–890 (2004).
61. Hedden, T. & Yoon, C. Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology* **20**, 511–528 (2006).
62. Weiss, E. M. *et al.* Sex differences in clustering and switching in verbal fluency tasks. *J. Int. Neuropsychol. Soc.* **12**, (2006).
63. Fournier-Vicente, S., Larigauderie, P. & Gaonac'h, D. More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. *Acta Psychol. (Amst.)* **129**, 32–48 (2008).
64. Whiteside, D. M. *et al.* Verbal Fluency: Language or Executive Function Measure? *Appl. Neuropsychol. Adult* **23**, 29–34 (2016).
65. Shao, Z., Janse, E., Visser, K. & Meyer, A. S. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Front. Psychol.* **5**, (2014).
66. Marchetta, N. D. J., Hurks, P. P. M., Krabbendam, L. & Jolles, J. Interference control, working memory, concept shifting, and verbal fluency in adults with attention-deficit/hyperactivity disorder (ADHD). *Neuropsychology* **22**, 74–84 (2008).
67. Schwartz, S. & Baldo, J. Distinct patterns of word retrieval in right and left frontal lobe patients: a multidimensional perspective. *Neuropsychologia* **39**, 1209–1217 (2001).
68. Azuma, T. Working Memory and Perseveration in Verbal Fluency. *Neuropsychology* **18**, 69–77 (2004).
69. Henry, J. D. & Crawford, J. R. Verbal fluency deficits in Parkinson's disease: a meta-analysis. *J. Int. Neuropsychol. Soc. JINS* **10**, 608–622 (2004).
70. Friedman, N. P. *et al.* Greater Attention Problems During Childhood Predict Poorer Executive Functioning in Late Adolescence. *Psychol. Sci.* **18**, 893–900 (2007).
71. Obeso, I., Casabona, E., Bringas, M. L., Alvarez, L. & Jahanshahi, M. Semantic and phonemic verbal fluency in Parkinson's disease: Influence of clinical and demographic variables. *Behav. Neurol.* **25**, 111–118 (2012).
72. Elgamal, S. A., Roy, E. A. & Sharratt, M. T. Age and Verbal Fluency: The Mediating Effect of Speed of Processing. *Can. Geriatr. J. CGJ* **14**, 66–72 (2011).
73. Lanting, S., Haugrud, N. & Crossley, M. The effect of age and sex on clustering and switching during speeded verbal fluency tasks. *J. Int. Neuropsychol. Soc. JINS* **15**, 196–204 (2009).
74. Hidalgo-Lopez, E. & Pletzer, B. Interactive Effects of Dopamine Baseline Levels and Cycle Phase on Executive Functions: The Role of Progesterone. *Front. Neurosci.* **11**, (2017).
75. Stauble, M. R., Thompson, L. A. & Morgan, G. Increases in cortisol are positively associated with gains in encoding and maintenance working memory performance in young men. *Stress* **16**, 402–410 (2013).
76. Scheuringer, A. & Pletzer, B. Sex Differences and Menstrual Cycle Dependent Changes in Cognitive Strategies during Spatial Navigation and Verbal Fluency. *Front.*

Psychol. **8**, 381 (2017).
77. Marczinski, C. A. & Kertesz, A. Category and letter fluency in semantic dementia, primary progressive aphasia, and Alzheimer's disease. *Brain Lang.* **97**, 258–265 (2006).
78. Benjamin, M. J., Cifelli, A., Garrard, P., Caine, D. & Jones, F. W. The role of working memory and verbal fluency in autobiographical memory in early Alzheimer's disease and matched controls. *Neuropsychologia* **78**, 115–121 (2015).
79. Suhr, J. A. & Jones, R. D. Letter and semantic fluency in Alzheimer's, Huntington's, and Parkinson's dementias. *Arch. Clin. Neuropsychol. Off. J. Natl. Acad. Neuropsychol.* **13**, 447–454 (1998).
80. Lundin, N. B. *et al.* Semantic Search in Psychosis: Modeling Local Exploitation and Global Exploration. *Schizophr. Bull. Open* **1**, (2020).
81. Pauselli, L. *et al.* Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia. *Psychiatry Res.* **263**, 74–79 (2018).
82. Wolters, M. K., Kim, N., Kim, J.-H., MacPherson, S. E. & Park, J. C. Prosodic and Linguistic Analysis of Semantic Fluency Data: A Window into Speech Production and Cognition. in 2085–2089 (2016). doi:10.21437/Interspeech.2016-420.
83. Patra, A., Bose, A. & Marinis, T. Performance difference in verbal fluency in bilingual and monolingual speakers. *Biling. Lang. Cogn.* 1–15 (2019) doi:10.1017/S1366728918001098.
84. Oh, S. J., Sung, J. E., Choi, S. J. & Jeong, J. H. Clustering and Switching Patterns in Semantic Fluency and Their Relationship to Working Memory in Mild Cognitive Impairment. *Dement. Neurocognitive Disord.* **18**, 47 (2019).
85. Chen, J. *et al.* Neurobiological substrates of the positive formal thought disorder in schizophrenia revealed by seed connectome-based predictive modeling. *NeuroImage Clin.* **30**, 102666 (2021).
86. Kohoutová, L. *et al.* Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* **15**, 1399–1435 (2020).
87. Lezak, M. D., Howieson, D. B., Bigler, E. D. & Tranel, D. *Neuropsychological Assessment*. (Oxford University Press, 2012).
88. Andreou, G. & Trott, K. Verbal fluency in adults diagnosed with attention-deficit hyperactivity disorder (ADHD) in childhood. *ADHD Atten. Deficit Hyperact. Disord.* **5**, 343–351 (2013).
89. Faraone, S. V. *et al.* Attention-deficit/hyperactivity disorder in adults: an overview. *Biol. Psychiatry* **48**, 9–20 (2000).
90. Planton, M. *et al.* The role of neuropsychological assessment in adults with attention deficit/hyperactivity disorders. *Rev. Neurol. (Paris)* **177**, 341–348 (2021).
91. Eysenck, M. W. & Calvo, M. G. Anxiety and Performance: The Processing Efficiency Theory. *Cogn. Emot.* **6**, 409–434 (1992).
92. Pessoa, L. How do emotion and motivation direct executive control? *Trends Cogn. Sci.* **13**, 160–166 (2009).
93. Sato, Y. *et al.* Neural Basis of Impaired Cognitive Flexibility in Patients with Anorexia Nervosa. *PLOS ONE* **8**, e61108 (2013).
94. PAAST, N., KHOSRAVI, Z., MEMARI, A. H., SHAYESTEHFAR, M. & ARBABI, M. Comparison of cognitive flexibility and planning ability in patients with obsessive compulsive disorder, patients with obsessive compulsive personality disorder, and healthy

controls. *Shanghai Arch. Psychiatry* **28**, 28–34 (2016).
95. Goel, V. & Grafman, J. Are the frontal lobes implicated in 'planning' functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia* **33**, 623–642 (1995).
96. Carpenter, P. A., Just, M. A. & Shell, P. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychol. Rev.* **97**, 404–431 (1990).
97. Pino Escobar, G., Kalashnikova, M. & Escudero, P. Vocabulary matters! The relationship between verbal fluency and measures of inhibitory control in monolingual and bilingual children. *J. Exp. Child Psychol.* **170**, 177–189 (2018).
98. Baddeley, A. D. & Hitch, G. Working Memory. in *Psychology of Learning and Motivation* (ed. Bower, G. H.) vol. 8 47–89 (Academic Press, 1974).
99. Delgado-Álvarez, A. *et al.* Cognitive Processes Underlying Verbal Fluency in Multiple Sclerosis. *Front. Neurol.* **0**, (2021).
100. Troyer, A. K. Normative data for clustering and switching on verbal fluency tasks. *J. Clin. Exp. Neuropsychol.* **22**, 370–378 (2000).
101. Kane, M. J. & Engle, R. W. Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *J. Exp. Psychol. Gen.* **132**, 47–70 (2003).
102. Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W. & Delazer, M. Sex differences in cognitive functions. *Personal. Individ. Differ.* **35**, 863–875 (2003).
103. McAllister-Williams, R. H. & Rugg, M. D. Effects of repeated cortisol administration on brain potential correlates of episodic memory retrieval. *Psychopharmacology (Berl.)* **160**, 74–83 (2002).
104. Torgrimson, B. N. & Minson, C. T. Sex and gender: what is the difference? *J. Appl. Physiol.* **99**, 785–787 (2005).
105. PubChem. Estradiol. https://pubchem.ncbi.nlm.nih.gov/compound/5757.
106. Sundström Poromaa, I. & Gingnell, M. Menstrual cycle influence on cognitive function and emotion processing—from a reproductive perspective. *Front. Neurosci.* **8**, (2014).
107. Miller, K. J., Conney, J. C., Rasgon, N. L., Fairbanks, L. A. & Small, G. W. Mood Symptoms and Cognitive Performance in Women Estrogen Users and Nonusers and Men. *J. Am. Geriatr. Soc.* **50**, 1826–1830 (2002).
108. Leeners, B. *et al.* Cognitive function in association with high estradiol levels resulting from fertility treatment. *Horm. Behav.* **130**, 104951 (2021).
109. Oei, N. Y. L., Everaerd, W. T. A. M., Elzinga, B. M., Well, S. van & Bermond, B. Psychosocial stress impairs working memory at high loads: An association with cortisol levels and memory retrieval. *Stress* **9**, 133–141 (2006).
110. Goldfarb, E. V., Froböse, M. I., Cools, R. & Phelps, E. A. Stress and Cognitive Flexibility: Cortisol Increases Are Associated with Enhanced Updating but Impaired Switching. *J. Cogn. Neurosci.* **29**, 14–24 (2017).
111. Plessow, F., Kiesel, A. & Kirschbaum, C. The stressed prefrontal cortex and goal-directed behaviour: acute psychosocial stress impairs the flexible implementation of task goals. *Exp. Brain Res.* **216**, 397–408 (2012).
112. Shields, G. S., Bonner, J. C. & Moons, W. G. Does cortisol influence core executive functions? A meta-analysis of acute cortisol administration effects on working memory,

inhibition, and set-shifting. *Psychoneuroendocrinology* **58**, 91–103 (2015).
113. Gonçalves, H. A. *et al.* Clustering and switching in unconstrained, phonemic and semantic verbal fluency: the role of age and school type. *J. Cogn. Psychol.* **29**, 670–690 (2017).
114. Galaverna, F., Bueno, A. M., Morra, C. A., Roca, M. & Torralva, T. Analysis of errors in verbal fluency tasks in patients with chronic schizophrenia. *Eur. J. Psychiatry* **30**, 305–320 (2016).
115. Pakhomov, S. V. S., Eberly, L. E. & Knopman, D. S. Recurrent perseverations on semantic verbal fluency tasks as an early marker of cognitive impairment. *J. Clin. Exp. Neuropsychol.* **40**, 832–840 (2018).
116. Pakhomov, S. V. S., Hemmy, L. S. & Lim, K. O. Automated semantic indices related to cognitive function and rate of cognitive decline. *Neuropsychologia* **50**, 2165–2175 (2012).
117. Farzanfar, D., Statucka, M. & Cohn, M. Automated Indices of Clustering and Switching of Semantic Verbal Fluency in Parkinson's Disease. *J. Int. Neuropsychol. Soc.* **24**, 1047–1056 (2018).
118. Pakhomov, S. V. S., Eberly, L. & Knopman, D. Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia* **89**, 42–56 (2016).
119. Kavé, G. & Sapir-Yogev, S. Associations between memory and verbal fluency tasks. *J. Commun. Disord.* **83**, 105968 (2020).
120. Azuma, T. *et al.* Comparing the difficulty of letter, semantic, and name fluency tasks for normal elderly and patients with Parkinson's disease. *Neuropsychology* **11**, 488–497 (1997).
121. Themistocleous, C. *et al.* Automatic Subtyping of Individuals with Primary Progressive Aphasia. *J. Alzheimers Dis.* **79**, 1185–1194 (2021).
122. Petti, U., Baker, S. & Korhonen, A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *J. Am. Med. Inform. Assoc.* **27**, 1784–1797 (2020).
123. Zimmerer, V. C. *et al.* Automated profiling of spontaneous speech in primary progressive aphasia and behavioral-variant frontotemporal dementia: An approach based on usage-frequency. *Cortex* **133**, 103–119 (2020).
124. Balagopalan, A., Eyre, B., Rudzicz, F. & Novikova, J. To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection. *ArXiv200801551 Cs* (2020).
125. Soni, A., Amrhein, B., Baucum, M., Paek, E. J. & Khojandi, A. *Using Verb Fluency, Natural Language Processing, and Machine Learning to Detect Alzheimer's Disease*. (2021). doi:10.13140/RG.2.2.26268.21126.

7 Appendix

Additionally to the presented studies in this thesis, a subsequent study was conducted, investigating the predictive potential of voice characteristics in ADHD. Due to the clinical relevance and the voice-focused analysis, this preprint provides a first step towards a patient focused transfer of advanced speech analysis applying ML methods.

Predicting adult Attention Deficit Hyperactivity Disorder (ADHD) using vocal acoustic features

Georg G. von Polier, Eike Ahlers, Julia Amunts, Jörg Langner, Kaustubh R. Patil, Simon B. Eickhoff, Florian Helmhold, Daina Langner

doi: https://doi.org/10.1101/2021.03.18.21253108

Abstract

It is a key concern in psychiatric research to investigate objective measures to support and ultimately improve diagnostic processes. Current gold standard diagnostic procedures for attention deficit hyperactivity disorder (ADHD) are mainly subjective and prone to bias. Objective measures such as neuropsychological measures and EEG markers show limited specificity. Recent studies point to alterations of voice and speech production to reflect psychiatric symptoms also related to ADHD. However, studies investigating voice in large clinical samples allowing for individual-level prediction of ADHD are lacking. To aim of this study was to explore a role of prosodic voice measures as objective marker of ADHD.

1005 recordings were analyzed from 387 ADHD patients, 204 healthy controls, and 100 clinical (psychiatric) controls. All participants (age range 18-59 years, mean age 34.4) underwent an extensive diagnostic examination according to gold standard methods and provided speech samples (3 min in total) including free and given speech. Paralinguistic features were calculated, and random forest based classifications were performed using a 10-fold cross-validation with 100 repetitions controlling for age, sex, and education. Association of voice features and ADHD-symptom severity assessed in the clinical interview were analyzed using random forest regressions.

ADHD was predicted with AUC = 0.76. The analysis of a non-comorbid sample of ADHD resulted in similar classification performance. Paralinguistic features were associated with ADHD-symptom severity as indicated by random forest regression. In female participants, particularly with age < 32 years, paralinguistic features showed the highest classification performance (AUC = 0.86).

Paralinguistic features based on derivatives of loudness and fundamental frequency seem to be promising candidates for further research into vocal acoustic biomarkers of ADHD. Given the relatively good performance in female participants independent of comorbidity, vocal measures may evolve as a clinically supportive option in the complex diagnostic process in this patient group.

Acknowledgements

First of all, I would like to express my sincere thanks to Prof. Simon Eickhoff for providing me the opportunity to become part of his institute and explore such an interesting research topic. His methodological input as well as his scientific farsightedness showed me the link between basics science, machine learning methods and clinical translation. His scientific direction and personal attitude allowed me to gain so much knowledge and always provided me a great feeling working in his institute.

My deeply thanks go to PD Dr. Susanne Weis and Dr. Julia Camilleri for supervising my project and accompanying every step of my PhD. Besides fruitful scientific discussions they were always supportive and understanding. Susanne introduced me to the world of conferences and contributed essentially to make so many thoughts and texts much more concise. She helped me to think and write logically. Julia was there from minute one and was always an anchor for me. I appreciate her calm, professional and sometimes very enthusiastic attitude. She always was a role model for me and provided support whenever needed.
It was a pleasure to work with this excellent supervisor team, that was always so responsive, supportive and became part of my life even beyond my PhD.

I am also thankful for the great atmosphere we had among PhD students. Particularly Lisa Mochalski, Lisa Hahn, Marisa Heckner, Lya Paas and Shammi More were so much more than colleagues and always guaranteed for deep and funny conversations that definitively highlighted my work routine. I will always remember our Karneval action and the fantastic time we had.

Last but definitively not least, I would like to thank my mother for her constant trust in me and my abilities. Despite a hard time at school, she always pushed me further and ultimately enabled me to grow beyond myself, discover new areas of science and finally find my own way. I appreciate very much that she did not want to influence my scientific work but was always there to drink a coffee together in the institute and support me emotionally, especially in personally difficult times.
I also want to thank Markus for supporting me permanently. He was my greatest critic, but also my closest friend and companion. In good times, he was there to share and further develop my enthusiasm about my project and in stressful times he supported me. Sharing this PhD journey with him, gave me the power to keep on going.