In search of better practice in executive functions assessment: methodological issues and potential solutions

Yanguez Escalera, Marc; Bediou, Benoît; Chanal, Julien; Bavelier, Daphné

Note: This manuscript is a preprint that has been submitted for publication and is currently under review.

In search of better practice in executive functions assessment: methodological issues and potential solutions

Marc Yangüez [1,2,3a], Benoit Bediou[1,3,a], Julien Chanal[1,2,b] and Daphne Bavelier[1,3,b]

[1] Faculty of Psychology, University of Geneva, Switzerland

[2] Distance Learning University, Brig, Switzerland

[3] Campus Biotech, University of Geneva, Switzerland

[a]Shared first authorship, [b]shared last authorship

Author Note

Corresponding author: Daphne Bavelier (daphne.bavelier@unige.ch), https://orcid.org/0000-0002-5904-1240; Campus Biotech, Chemin des Mines 9, CH-1202 Geneva

Marc Yangüez (marc.yanguezescalera@unige.ch), https://orcid.org/0000-0001-9513-4543

Benoit Bediou (benoit.bediou@unige.ch), https://orcid.org/0000-0002-3477-7948

Julien Chanal (Julien.chanal@unige.ch), https://orcid.org/0000-0002-9670-1340

We have no conflict of interest to disclose.
Data and R-code are available in the OSF website (https://osf.io/yvcj7/).
This study was not preregistered.
Support for this project was provided by the Foundation Ernest Boninchi (CH) and by the Jacobs Foundation (CH).

Part of this work has been presented in the following conferences:

Yangüez, M., Bediou, B., Chanal, J., & Bavelier, D. (2022). When cognitive modeling meets latent variable methods: impact of RT, accuracy, or drift Rate in measurement models of executive functions' structure. Psychonomic society 63rd annual meeting, Boston, MA (November 17 –20). Conference talk.

Yangüez, M., Bediou, B., Chanal, J., & Bavelier, D. (2022). In search of best practice in executive functions' assessment: a latent variable approach. European Society for Cognitive Psychology (ESCOP), Lille, France (August 29 – September 1). Conference talk.

Yangüez, M., Bediou, B., Chanal, J., & Bavelier, D. (2021). Drift Rate Improves the Psychometric Modeling of Executive Functions. Poster presented at the Psychonomic Society Annual Meeting (November 4-7).

Abstract

The multi-component nature of executive functions (EF) has long been recognized, pushing for a better understanding of both the commonalities and the diversity between EF components. Despite the advances made, the operationalization of performance in EF tasks remains rather heterogeneous, and the structure of EF as modelled by confirmatory factor analyses (CFA) is still a topic of debate (Karr et al., 2018). The present work demonstrates these two issues are related, showing how different operationalizations in task-based performance indicators impact the resulting models of EF structure with CFA.

Using bootstrapped data from 182 children (8-12 years old) and nine EF tasks (tapping inhibition, working memory and cognitive flexibility), we first show improved model convergence and acceptance when operationalizing EF through single tasks' scores (e.g., incongruent trials, Flanker task) relative to difference scores (e.g., incongruent minus congruent trials, Flanker task). Furthermore, we show that reaction times exhibit poor model convergence and acceptance compared not only to accuracy, but also drift rate. The latter, a well-known indicator in drift-diffusion models, is found to present the best psychometric properties to model EF with CFA. Finally, we examine how various operationalizations of performance in EF tasks impact CFA model comparison in the assessment of EF structure and discuss the theoretical foundations for these results.

KEYWORDS: executive functions; confirmatory factor analyses; latent variable analysis; diffusion model; assessment; indicators.

Executive functions are considered an «umbrella term» for a number of cognitive processes relying on frontal lobe functioning (Barkley, 2012), which are fundamental for school readiness (e.g., Blair & Razza, 2007; Morrison, Ponitz, & McClelland, 2010), scholastic performance (e.g., Duncan et al., 2007; Jacob & Parkinson, 2015; St Clair-Thompson & Gathercole, 2006), job success (e.g., Bailey, 2007), or mental health (e.g., Baler & Volkow, 2006; Gardiner & Iarocci, 2018; Penadés et al., 2007; Taylor Tavares et al., 2007). The importance of executive functions for such variety of life aspects explains, in part, the growing interest in psychological and neuroscience research around these functions in the last decades. Unfortunately, the proliferation of studies about executive functions has been coupled with important methodological differences in their conceptualization and measurement, which hinders our understanding of the psychological and theoretical mechanisms of executive functions (Baggetta & Alexander, 2016; Barkley, 2012; Karr et al., 2018; McCabe et al., 2010; Packwood et al., 2011). Furthermore, recently several publications have expressed their concern about the poor psychometric properties of many executive functions measures, a critical point for individual differences research (Draheim et al., 2019; Hedge et al., 2018; Paap & Sawi, 2016; Rouder & Haaf, 2019).

To understand what executive functions are, what components form them, and how they organize, first, it is essential to determine how to best operationalize ability in executive functions tasks to better capture the latent processes under assessment. Traditionally, the field has relied on the use of single tasks to assess executive functions components (Baggetta & Alexander, 2016; Chan et al., 2008); however, this approach fails to recognize that no task is process pure, as each task necessarily involves processes other than the intended one (Conway et al., 2005; Miyake, Friedman, et al., 2000; Shah & Miyake, 1996), a phenomenon commonly known in the literature as *task impurity*. Accordingly, it has been widely documented through psychometric analyses that the use of a single task to characterize one or multiple executive functions components suffers from both validity and reliability issues (e.g., Kane et al., 2004; Shah & Miyake, 1996; Yang & Green, 2011). Executive functions assessment remains challenging due to the complexity of the constructs they encompass. This is in part because commonalities between the components that form it (and the tasks used to measure them) do exist, and yet some diversity across components is also noted (Friedman et al., 2008; Frischkorn & von Bastian, 2021; Karr et al., 2018; Miyake, Friedman, et al., 2000). To address the commonalities and diversity between executive functions components, researchers have exploited the use of multiple tasks coupled with confirmatory factor analysis (CFA), a special form of structural equation modeling (SEM) technique, which enables to define and estimate measurement models to analyze the relationship between manifested variables (or indicators) and the latent variables that form the models (MacCallum & Austin, 2000). CFA is a powerful tool for psychometric evaluation and construct validation (Brown & Moore, 2012). Thus, this approach has been very useful to mitigate both the *task impurity* problem, enabling a better evaluation of the cognitive components when the tasks used are not process pure, and the *measurement error* problem, removing the unique variance from each task (Engle et al., 1999; Kane et al., 2004; Miyake, Emerson, et al., 2000; Miyake, Friedman, et al., 2000; Shah & Miyake, 1996). The present study builds on this approach to investigate how different methods proposed in the literature to operationalize executive functions impact the assessment of the underlying structure of executive functions when using CFA.

Executive functions latent variable studies

Prior to the seminal article by Miyake et al (2000), which introduced the use of CFA to assess executive functions components and its multidimensional structure, the earliest models already viewed executive functions as a higher-order global construct that managed lower-level cognitive processes (Baddeley & Hitch, 1974; Norman & Shallice, 1986). Although these models did not include the term executive functions explicitly, they form the foundation for subsequent models of executive functions, which have described the construct as multidimensional, including higher-order and lower-order cognitive processes, which are moderately to strongly correlated (Diamond, 2013; Engle, 2018; Friedman & Miyake, 2017; Miyake, Friedman, et al., 2000). The existence of different models of executive functions illustrates, in part, the complexity of defining what are executive functions or in other words, the challenge of characterizing which combination of cognitive processes they encompass. There is general agreement, however, that executive functions is a multidimensional construct, which involves core components such as (i) inhibition, (ii) cognitive flexibility and (iii) working memory, which are fundamental for higher-order processes, such as planning, reasoning or goal-directed behavior (Baggetta & Alexander, 2016; Diamond, 2013; Hughes, 2011). Inhibition is globally defined as the ability to suppress the processing of irrelevant stimuli or the outcome of impulsive reactions (MacLeod, 2007); cognitive flexibility, also termed shifting, refers to the capacity to swiftly change focus whether in terms of task goals or attention distribution (Best & Miller, 2010; Ionescu, 2012); finally, working memory refers to the ability to keep information active in mind and mentally manipulate it (Diamond, 2013; Kane et al., 2004). Note that working memory is a multidimensional construct, with components that can be distinguished based on their emphasis on (i) content material (e.g., verbal and visuospatial) or (ii) constituent processes (e.g., updating and maintenance) (Smith & Jonides, 1997; Waris et al., 2017). These three components form the core of most studies investigating executive functions (Karr et al., 2018; Packwood et al., 2011). Note that in part of the literature, and in the present work, the terms shifting and cognitive flexibility are used interchangeably, as well as the terms updating and working memory (Baggetta & Alexander, 2016; Diamond, 2013).

CFA has thus been helpful in mitigating the *task impurity* problem, as it has highlighted time and time again that the predictive validity and reliability of a latent construct is greater than that of the single measures from which it is derived (Conway et al., 2005; Kane et al., 2004; Willoughby et al., 2017). Furthermore, SEM techniques are a powerful statistical tool for individual differences research in the fields of working memory (Engle et al., 1999; Kane & Engle, 2002; Ørskov et al., 2021; Rey-Mermet et al., 2019; Shah & Miyake, 1996) and executive functions (Arán Filippetti & Richaud, 2017; Friedman et al., 2008; Schmidt et al., 2017; Sluis et al., 2007; Spencer et al., 2020). It comes thus as no surprise that in the last decade many studies have used SEM techniques to investigate executive functions development, their structure, as well as their neural organization (Alfonso & Lonigan, 2021; Brydges et al., 2014; Cirino et al., 2018; Huizinga et al., 2006; Lambek & Shevlin, 2011; Lee et al., 2013; Lerner & Lonigan, 2014; Monette et al., 2015; Montroy et al., 2019; Ritchie et al., 2019; Rose et al., 2012; Usai et al., 2014; Wiebe et al., 2011; Willoughby et al., 2012; Xu et al., 2013). For example, latent variable studies suggest that executive functions develop and differentiate from a rather unitary structure to a multidimensional structure throughout childhood and adolescence. Below 8 years of age, most studies document either a unitary structure (Brydges et al., 2014; Shing et al., 2010; Wiebe et al., 2008, 2011; Willoughby et al., 2012) or a two factor structure (Lerner & Lonigan, 2014; M. R. Miller et al., 2012; Monette et al., 2015; Usai et al., 2014). In that age range, the three basic processes of executive functions seem to be initially undifferentiated, and then inhibition is often reported as emerging first. Studies in middle childhood and adolescence show a gradual differentiation of executive functions to the three-factor structure most often described in young adults (e.g., Brydges et al., 2014; Lehto et al., 2003; Rose et al., 2012; Shing et al., 2010).

Recently, in an extensive literature review of studies that applied CFA to assess the structure of executive functions, Karr et al. (2018) pointed out several weaknesses when applying CFA to executive functions research. They performed a literature search resulting in 40 articles, 17 of which provided sufficient data for re-analysis through bootstrapping methods.

This literature search also identified seven different measurement models of executive functions being the most commonly found in the literature. As displayed in Figure 1, these are: one unidimensional model merging all three factors of inhibition, cognitive flexibility, and working memory; three two-factor models (merging the three factors above two by two, so inhibition with working memory, inhibition with cognitive flexibility, and working memory with cognitive flexibility); one three-factor model (inhibition, cognitive flexibility, and working memory); one nested factor model (a common factor, plus two specific orthogonal factors of cognitive flexibility, and working memory); and one bifactor model (a common factor, plus three specific orthogonal factors of inhibition, cognitive flexibility, and working memory).

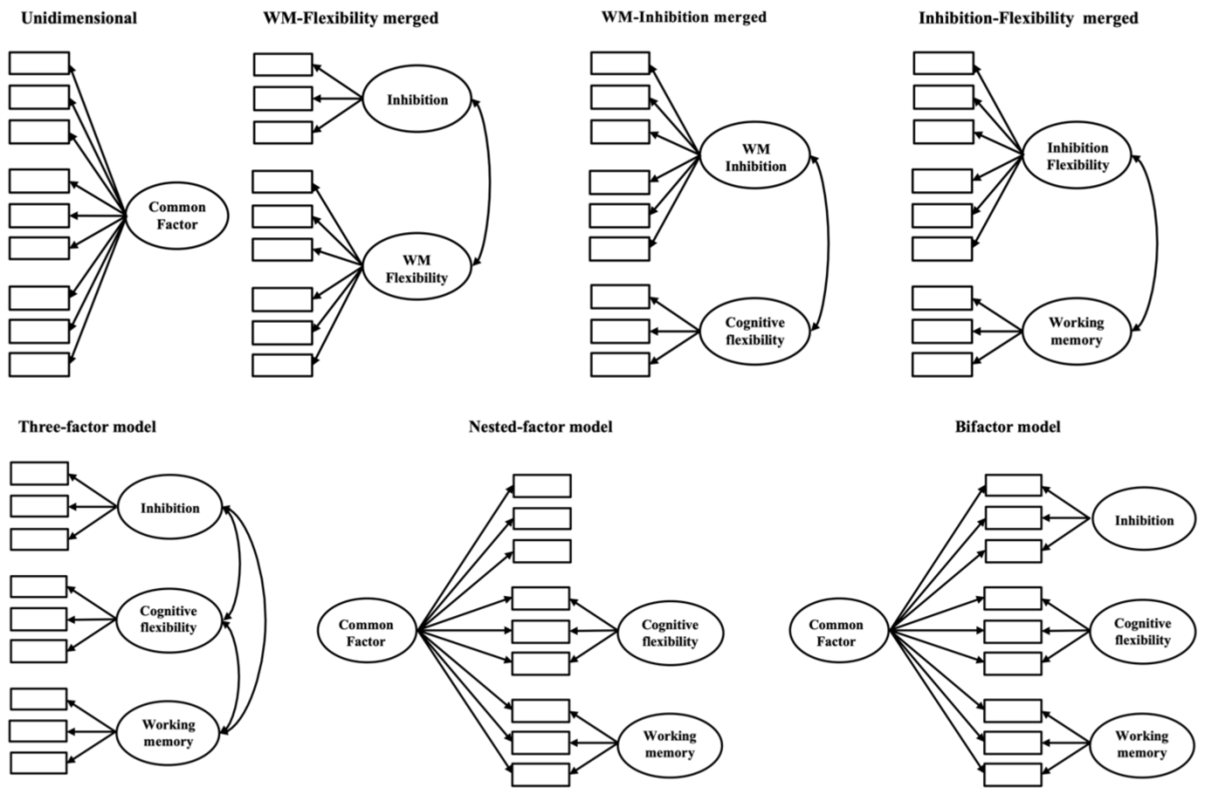

Figure 1. Measurement models of executive functions (adapted from Karr et al., 2018).

For each of the 17 studies that provided sufficient data for re-analysis, these seven measurement models were fitted to 5000 data sets generated through bootstrapping. Their main objective was to determine the replicability of the measurement models of executive functions published up to date, and to evaluate which published model best fitted the data across bootstrapped samples. Strikingly, none of the seven measurement models of executive functions consistently both converged and fitted well the data. Moreover, no model was consistently selected as the best model when it was directly compared with other models. Karr et al. (2018) concluded that the observed low rates of acceptance and selection might be due to a possible bias towards well-fitting models tested with underpowered samples, but also, that one of the core challenges for the field remains in identifying the methodological choices that enhance the consistency of executive functions measurement with CFA. The present work addresses this challenge. In the next section, we review the heterogeneity of methods proposed in the literature to measure and operationalize executive functions, with a specific emphasis on the tasks' conditions and indicators used.

Heterogeneity of tasks' conditions in executive functions studies: single versus difference of conditions

Executive functions tasks have been traditionally designed by contrasting two task conditions with the aim of disentangling both executive and non-executive processes involved during task performance. This approach has its origin in Donders' subtraction method (Donders, 1868), and has been hugely successful in the early days of experimental psychology. Accordingly, some of the most used executive functions tasks' paradigms are built following this approach since it gives robust experimental effects (Eriksen & Eriksen, 1974; Rogers & Monsell, 1995; Simon & Rudell, 1967; Stroop, 1935). Yet, the subtraction method has also been under criticism as the additivity of processing time is largely unsubstantiated (Gomez et al., 2007; Ulrich, 1999; Wundt, 1880). Given our interest in the use of latent variable models for executive functions research, we focus our discussion of the heterogeneity of tasks' conditions across studies to those that have used a CFA approach to assess executive functions; yet this same issue applies throughout the large executive function literature. Table S1 (online supplementary material) provides a list of studies that applied CFA to investigate the structure of executive functions in pre-school and school-aged children, and information about the tasks, tasks' conditions and indicators used to measure executive functions on each study.

Classical inhibition tasks, such as the Flanker task or the Simon task, contrast congruent trials, hypothesized to tap non-executive abilities, with incongruent trials, intended to tap both non-executive abilities and the core executive process of inhibition. Studies have commonly operationalized performance on these tasks either through performance on incongruent trials (e.g., Lee et al., 2013; Van der Ven et al., 2013), or through the difference in performance between congruent and incongruent trials (Bender et al., 2016; Friedman & Miyake, 2004; Unsworth et al., 2009), also known as interference or difference score.

Cognitive flexibility is typically measured through task-switching paradigms (Monsell, 2003). This sort of tasks frequently includes two types of blocks: a homogeneous condition (blocks of trials requiring a response only to a feature of the stimuli, for instance, the color), and a heterogeneous-mixed condition (mixed rule-set of cues to flexibly shift attention towards the correct target feature - for instance, arms down indicates respond to the color; arms up respond to the shape). Performance on task-switch paradigms is commonly operationalized with three different methods: (i) global switch cost (e.g., Miyake, Friedman, et al., 2000), which is the difference in performance between the heterogeneous-mixed condition (i.e., blocks including switch and non-switch trials) and the homogeneous condition (blocks including only non-switch trials); (ii) local switch cost (e.g., Ambrosini et al., 2019; Friedman & Miyake, 2004), which is the difference between switch and non-switch trials in the heterogeneous-mixed condition; and (iii) performance on switch trials from the heterogeneous-mixed condition (e.g., Huizinga et al., 2006; Lee et al., 2013). Again, depending on the task condition(s) selected to operationalize cognitive flexibility, either a single task condition or a difference between conditions may be considered.

Finally, working memory tasks typically use only a single score that captures the number of items that can be held or manipulated, as per standard span tasks for example (Conway et al., 2005). The n-back task departs from span tasks by allowing working memory assessment under different levels of memory load. Of note, neuro-imaging studies of working memory often analyze differences in brain activity between two different loads of the n-back task (e.g., 2-back minus 0-back condition, Braver et al., 1997; Yaple & Arsalidou, 2018); yet, purely behavioral studies rarely apply the subtractions method to analyze n-back task performance (e.g., use of only the 2-back condition, Duan et al., 2010; Waris et al., 2017).

In sum, researchers that aim to model performance on this sort of tasks must choose between operationalizing ability (i) through performance on those trials that require greater amounts of executive control (i.e., incongruent trails, switch trials), termed thereafter single task condition or (ii) through a difference score, subtracting the performance in one task condition from another, termed thereafter conditions difference. Such operationalization differences between studies are likely to result in different performance assessments, although the sub-processes evaluated are similarly labelled. For example, the factor termed ‘inhibition’ can refer to performance on incongruent trials, as well as to the performance difference between incongruent and congruent trials. The present work highlights that this state of affair is not just introducing a possible source of confusion in the field, but that the use of single versus difference scores may have a major impact in the convergence and acceptance of measurement models of executive functions, which in turn may affect their replicability.

Heterogeneity of indicators in executive functions studies

The issue of the heterogeneity of indicators concerns mainly reaction time-based tasks, in which both speed and accuracy are relevant indicators of task performance. For instance, the Stroop task (Stroop, 1935) has been used by several latent variable studies to assess inhibition, which differed on how to operationalize performance on this task. Bridges et al. (2014) subtracted the difference in reaction time (RT) between congruent and incongruent trials of the Stroop task; van der Sluis et al. (2007) used instead the number of correct items per second on incongruent trials, whereas van der Ven et al. (2013) operationalize performance on the task through the accuracy in incongruent trials. Similarly, studies can differ remarkably in the indicators used to operationalize performance in cognitive flexibility tasks. For instance, both Rose et al. (2012) and van der Sluis et al. (2007) used the Trail Making Test (Reitan, 1971) to assess this construct. Rose et al. (2012) subtracted the time in seconds to complete both task conditions (i.e., Trail-B – Trail-A), whereas van der Sluis et al. (2007) used instead the number of seconds to complete the Trail-B test, which is the task condition intended to tap cognitive flexibility. Such heterogeneity in indicators is even observed within the same study for different tasks tapping the same construct. For instance, Friedman et al. (2008) estimated the inhibition construct using three different tasks that each allow measurement of speed and of accuracy. Yet, for the Antisaccade task, the indicator used was accuracy; for the Stop-signal task, the indicator was mean RT on the stop-signal condition; whereas for the Stroop task, the indicator was RT difference between congruent and incongruent trials.

As illustrated in Table S1, accuracy, RT, and capacity measures (e.g., maximum number of items correctly recalled) are most often used indicators in the literature, at least for those works using CFA as tabulated here. For working memory, performance is standardly operationalized in terms of span capacity or accuracy (Wilhelm et al., 2013), creating less variability in the indicators used. For inhibition and cognitive flexibility, performance is more standardly assessed via RT-based tasks, leading to the possibility of using speed, accuracy, or a combination thereof as indicators. We turn below to the psychometric issues raised by such varied operationalizations of RT-based tasks, with a special focus on the tasks' conditions or the measures used to derive an indicator.

Psychometric issues associated with RT-based measures

RT and RT differences are two of the most popular indicators used to study the speed and efficiency of mental processes in psychology and neuroscience research (Draheim et al., 2019). Despite their widespread use in experimental and individual differences research, several studies have shown that both indicators suffer from reliability and validity issues.

It has been argued that the subtraction method increases the error variance, since it removes part of the common variance between the two mental processes from which the RT difference score is calculated (Hedge et al., 2018). As an illustration of this issue, Paap and Sawi (2016) assessed the test-retest reliability of (i) single RT (e.g., mean RT in congruent or in incongruent trials from inhibition tasks; mean RT in switching or in non-switching trials in cognitive flexibility tasks) and (ii) RT difference between task conditions on four classical executive functions tasks (i.e., Antisaccade, Flanker, Simon & Color-shape switching) in a sample of undergraduate students (N = 81). Their results indicate that single RT is a more reliable behavioral indicator (.71-.89, test-retest reliability range) than RT difference scores (.43-.62). Importantly, such results are in line with those reported in other studies (e.g., Hughes, Linck, Bowles, Koeth, & Bunting, 2014; Salthouse, Fristoe, McGuthry, & Hambrick, 1998; Siegrist, 1997).

RT measures are sensitive to speed-accuracy trade-off, whereby participants as they are told to react faster will show greater error rates, and vice-versa (Fitts, 1966; Ratcliff & Rouder, 1998; Stone, 1960). Despite the effort to instruct participants to give a similar weight to both dimensions, participants tend to adopt different response strategies (Starns & Ratcliff, 2012). Importantly, the literature has shown consistently that age-related differences exist in speed-accuracy trade-off strategies. For example, when comparing older adults with younger adults, the first tend to put more weight on accuracy over speed in order to ensure a higher accuracy, whereas the latter tend to take more risk speeding up their responses at the cost of making more errors (e.g., Forstmann et al., 2011; Hertzog, Vernon, & Rypma, 1993; Smith & Brewer, 1995; Starns & Ratcliff, 2012). Given the complex interaction between speed and accuracy (see Heitz, 2014, for a review), the point has been made that analyses based solely on RT cannot fully account for individual differences in cognition; this is particularly the case of studies with heterogeneous samples, such as on developmental or aging studies, which inevitably will include participants with different response strategies (Draheim et al., 2016; Hertzog et al., 1993; Hughes et al., 2014; Ratcliff et al., 2016; Yang et al., 2015).

Several efforts have been made to address this speed-accuracy trade-off issue, through the development of indicators such as the inverse efficiency score (IES: Townsend & Ashby, 1978), the linear-integrated speed-accuracy score (LISAS: Vandierendonck, 2017, 2018), the rate-correct score (RCS: Woltz & Was, 2006) or the balanced integration score (BIS: Liesefeld, Fu, & Zimmer, 2015). These measures provide several benefits over traditional RT- or accuracy-based measures, such as (i) mitigating speed-accuracy tradeoffs, or (ii) containing more information about individuals' ability than RT and accuracy separately. However, there is debate about the weight that such measures give to speed over accuracy (or vice-versa) to generate a reliable integrated measure of speed and accuracy (Draheim et al., 2019; Liesefeld & Janczyk, 2019). Such composite measures have rarely been used in the context of executive functions latent variable research, although there are some exceptions, such as Gärtner & Strobel (2021) and Yangüez et al., (2021), who used the IES (or RT divided by response accuracy) to operationalize performance on different RT-based executive functions tasks. In the next section, we review another approach to explain patterns of RTs and choices, the Drift Diffusion Model (DDM).

Applications of the DDM to executive functions research

The DDM is likely the most well-known psychometric model of decision making. Responses in choice tasks are understood as generated through sequential sampling of Brownian diffusing signals up to a decision boundary. In this view, evidence accumulates for each of the possible choices until the decision boundary of a given choice is hit, triggering the execution of the response corresponding to that choice (Ratcliff & McKoon, 2008; Shadlen & Kiani, 2013). By modelling the underlying generative events that give rise to decision processes, the DDM provides a natural way of accounting for speed/accuracy tradeoffs (Ratcliff, 1978; Ratcliff et al., 2016).

The initial diffusion model was developed for two-choice RT paradigms, although it can be generalized to paradigms that include more than two choices (Ratcliff et al., 2016; Tajima et al., 2019). When the DDM is applied to a two-choice RT paradigm (represented in Figure 2), stimulus presentation triggers the decision process, and in particular information accumulation until one of the two decision boundaries is reached (*0* or *a, for the two-choice model in Figure 2*). The drift rate (v) represents the average rate of evidence, with a larger drift rate meaning a faster accumulation of evidence, and vice-versa. The model assumes that drift rate varies across trials following a Normal or Gaussian distribution according to $\xi \sim N(v,n)$, since the accumulation process is subject to moment-to-moment, Brownian variability. The distance between the two decision boundaries will affect an individual's speed-accuracy trade-off. Larger values of the boundary parameter a represent a conservative response strategy, as a larger a means more information needs to be accumulated before a decision can be made (resulting in longer RTs and higher accuracy). The z parameter can be added into the model to examine whether an individual has an a priori bias towards one of the two response options, before the stimulus is experienced. The DDM also enables one to estimate the non-decision time or the time to execute a motor response, *Ter*. For a full description of the standard diffusion model, the reader is referred to Ratcliff et al. (2004). Here we focus on drift rate as it characterizes the efficiency of information processing and is thus most promising to characterize core executive functions processes, in contrast to decision boundary, which is related to strategic factors, and to non-decision time, which captures additive processes such as preparation and motor execution (Ratcliff, 1978; Ratcliff et al., 2016).

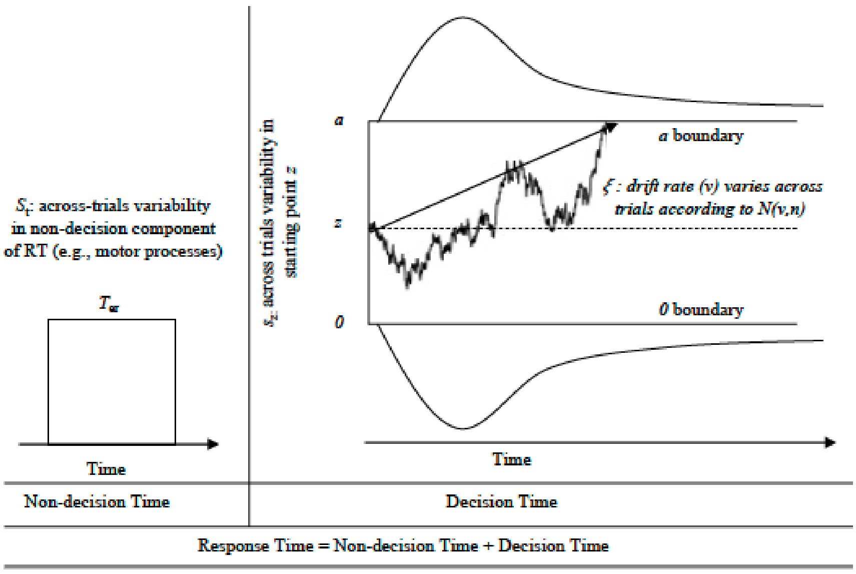

Figure 2. Diffusion model account of evidence accumulation (image adapted from Wagenmakers et al., 2007).

Diffusion models have been applied to some of the most well-known executive functions tasks, such as the Flanker (Ong et al., 2017; Servant & Evans, 2020; White et al., 2011), Simon (McIntosh & Mehring, 2017; McIntosh & Sajda, 2020; Servant et al., 2014) Go/No-Go (Gomez et al., 2007; Ratcliff et al., 2018), Stroop (Fennell & Ratcliff, 2019; Gajewski et al., 2020), task-switch (Ging-Jehli & Ratcliff, 2020; Schmitz & Voss, 2012; Weeda et al., 2014), and n-back tasks (Thurm et al., 2018). Some of these studies have also successfully applied DDM, and in particular drift rate, to investigate individual differences in executive functions (Gajewski et al., 2020, Ging-Jehli et al., 2020, Ong et al., 2017, Servant & Evans, 2020, Ratcliff et al., 2018, Weeda et al., 2014). Yet, the use of drift rate as an indicator for modeling executive functions with SEM techniques, such as CFA, remains largely untested (for an exception, see Rey-Mermet et al., 2021).

In the present work, we take advantage of the EZ-diffusion model (Wagenmakers et al., 2007) to systematically assess how indicator choices, including drift rate, may affect CFA measurement models of executive functions, in terms of convergence and acceptance. The EZ-diffusion model only estimates the three most relevant parameters of the DDM to characterize the decision-making process: (i) the drift rate *v*, (ii) the decision boundary *a* and (iii) the non-decision time *Ter*. The EZ-diffusion model is a solution when tasks do not have enough trials to estimate all the parameters from the DDM simplifying remarkably the standard fitting procedure (Palmer et al., 2005; Ratcliff, 1978, 2002). This is often the case of executive functions CFA studies, which use multiple tasks to measure each of the constructs included in their models, and therefore, the tasks often are not designed to fit more parameters, as is required by the standard DDM or more recent DDM versions with collapsing bound (Drugowitsch et al., 2012; Fudenberg et al., 2018).

Aims of the present study

The present study aims to test the impact of different operationalizations of executive functions in terms of the tasks' conditions and indicators used, on the modeling of executive functions with CFA. This investigation focuses not only on the comparison of single versus difference scores in task's conditions, but also on the choice of indicators between more standard measures, such as RT and accuracy, and less common ones, such as the IES and drift rate. Indeed, our final aim is to determine the impact of such different methodological practices on executive functions modeling, especially the metrics related to model fitting, such as convergence and acceptance (i.e., how well a model fits to the data). Latent variable models are likely sensitive to such operationalization (MacCallum & Austin, 2000), as it is expected that some indicators will show greater common variance than others. Thus, beyond the choice of tasks and conditions, how a researcher operationalizes performance in that very task will have important implications for the modeling of the construct evaluated. In particular, these choices may lead to different results when modeling the structure of executive functions (e.g., model convergence and acceptance, level of factor-loadings, number of latent constructs of the model, inter-factor correlations, etc.). In line with Karr et al. (2018), the rate of model convergence and acceptance (or how often, when a model converges, it meets a fitting threshold) will be examined as tasks' conditions and indicators vary. In addition, whether the likelihood of model selection varies depending on the task conditions and indicators used will also be assessed, to the extent the model converges. In this way, the present work establishes not only which operationalizations of executive functions should be preferred, but also which measurement models of executive functions (e.g., one-factor, two-factor, three-factor, etc.) are most likely representative of executive functions structure during middle childhood.

Methods

Participants and Procedure

This dataset was already used in a previous study (Yangüez et al., 2021). Below we briefly describe participants' characteristics and the data collection procedure. The sample included 182 children (92 females, mean age 10.53, *SD* = 1.17, range 8-12.75 years) recruited from primary schools in Geneva (Switzerland). Data collection was conducted by trained research assistants in a quiet room within the schools' facilities, in groups of two to four children. All procedures were in accordance with the Declaration of Helsinki about ethical principles regarding human experimentation, and were approved by the Ethics Committee of the University of Geneva. For more details, see Yangüez et al. (2021). The present study was not preregistered.

Measures

We used nine tasks to assess three components of executive functions (i.e., inhibition, cognitive flexibility and working memory). All tasks were computer-based, except for the Trail Making Test (Reitan, 1971), which was administered in paper-pencil format. A complete description of the tasks can be found in Yangüez et al. (2021). 123 children completed the nine tasks (67.6% of the total sample, *n* =182), and all the children completed at least six tasks (two per construct). Table S2 (online supplementary material) reports for each of the nine tasks, the number of children that completed each of them.

Description of executive functions tasks

Inhibition tasks. Flanker task - Modified (Eriksen & Eriksen, 1974; Pontifex et al., 2013). In this task, an array of five fishes is presented with the central fish pointing either in the same direction (congruent trials) or in the opposite direction (incongruent trials) as the other fish. The task is to determine the direction the middle, target fish, is facing. Incongruent trials in this task require the greatest inhibitory control demands due to perceptual interference.

Simon task - Modified (Morton & Harper, 2007). In this task, children press the appropriate response key whether a blue or red square appears on the left or right side of the screen. In congruent trials, the square appears on the same side of the screen as the response key to which it is associated (e.g., left-left); whereas in incongruent trials, it appears on the opposite side (e.g., left-right). Incongruent trials in this task require the greatest inhibitory control demands due to response conflict.

Go/No-Go task (Kamijo et al., 2012). In the first part of this task (sustained attention condition) children must press the response button to rare-target stimuli (picture of a lion, 0.2 probability), and to withhold their response to frequent non-target stimuli (picture of a tiger, 0.8 probability). Then children perform the second part of the task (impulsivity condition), in which they must press a button to frequent-target stimuli (tiger, 0.8 probability), and to withhold from responding to rare-non-target stimuli (lion, 0.2 probability). That task order is fixed to induce greater conflict and thus, need for inhibition during the “impulsivity condition”.

Cognitive flexibility tasks. Color-shape switch Task (Espy, 1997). This task requires children to judge the color (blue or green) or shape (circle or square) of the stimulus presented and press the appropriate response button. This task includes two types of blocks. The homogeneous blocks were made of trials requiring a response only to the color of the stimuli, or alternatively of trials requiring a response only to the shape of the stimuli. The heterogeneous-mixed block contained a mixed ruleset of cues to flexibly switch attention towards the correct target feature. For example, arms down indicated to respond to the color of the stimulus, and arms up to respond to the shape of the stimulus. Importantly, the heterogeneous mixed-block includes both switch and non-switch trials, whereby the rule set changes from trial n to n+1 or stays the same. Switch trials require the greatest cognitive flexibility demands, as individuals need to switch task goals.

Gender-Smile switch task - Modified (Huizinga et al., 2006). In this task the stimuli are schematic faces (male or female, happy or sad), appearing in a 2 × 2 grid. At the beginning of the task (homogeneous condition), the children must answer regarding either gender or expression in separate blocks. In the third block (heterogeneous-mixed condition), the stimuli move clockwise through the grid and children must respond regarding the gender, when the face appears in one of the two upper quadrants, and regarding the expression of the face, when it appears in one of the two lower quadrants. As in the color-shape task, the heterogeneous mixed-block includes both switch and non-switch trials, which will be used to derive single versus difference scores. Unlike the color-shape task, repetitions and switch trials in the Gender-Smile task follow a predefined sequence and are thus predictable.

Trail Making Test (Reitan, 1971). In Trail A, children are asked to draw lines connecting numbers by numerical order (numbers from 1-25 are distributed randomly across the test-sheet). In Trail B, the test-sheet contains numbers and letters, and children have to connect numbers and letters by alternating the sequence (i.e., 1-A-2-B-3-C, etc), requiring continuously switching between two different task sets. Trail-B is the most demanding experimental condition in terms of cognitive flexibility.

Working memory tasks. Letter-Memory task – Modified (Tamnes et al., 2010). This is a running memory task, where letters are presented serially in the center of the computer screen. Children's task is to recall the last three letters presented in each list. The number of letters presented (5, 7, 9, or 11) varies randomly across trials to limit strategies and enforce attention across most material.

Backwards digit-span task (Wechsler, 1991). In this task children must recall the numbers they have just heard (from the computer) in reverse order. The task starts with three series of a two digits sequence, and the number of digits increases progressively until reaching children's span capacity. The task ends when, within a series of digits (e.g., six digits sequence), the child gives the wrong answer in two out of three trials of the series.

Spatial n-back task – Modified (Drollette et al., 2012). On each trial of this task, a schematic yellow happy face appears pseudo-randomly inside one of the six boxes. This task included three conditions, 0-back, 1-back, and 2-back. The latter is the experimental condition that requires the greatest working memory demands. On 2-back trials participants are instructed to press the right-button if the schematic face appears in the same box as two trials back, otherwise they must press the left-button.

Dependent measures derived from each task

Single-condition indicators represent performance on those trials (or task's conditions) that require greater amounts of executive control (e.g., incongruent trials, switch trials, 2-back trials), whereas condition-difference indicators represent the performance difference between that single task condition with the greater executive control demands and a baseline task condition with low executive control demands (e.g., incongruent minus congruent trials in the Flanker task; switch minus non-switch trials in the Color-Shape switch task; and 2-back minus 0-back in the Spatial n-back task). Table 1 summarizes the task conditions and indicators used for each of the nine tasks. For a description of how the four indicators were computed, see *statistical procedures*.

Inhibition tasks. On the Flanker and Simon tasks, RT and accuracy were recorded, and in addition, we computed IES and drift rate. Single-condition indicators were derived from incongruent trials (RT, accuracy, IES, and drift rate). Condition-difference indicators were derived by subtracting the score difference between the tasks' conditions (incongruent minus congruent trials) for each indicator (i.e., RT difference, accuracy difference, IES difference, and drift rate difference).

On Go/No-Go tasks, RT, although measured, is not considered as a proper measure of inhibition, as it is only collected from Go trials or from errors on No-Go trials. Rather, accuracy measures, such as error rates (i.e., false alarms), have been historically the primary variables of interest to measure inhibition on Go/No-Go Tasks, as they inform about the proportion of responses that individuals fail to withheld (Wright et al., 2014). Therefore, a pure RT measure could not be derived. To compute IES and drift rate, response accuracy was computed collapsing performance in Go (i.e., hits & misses) and No-Go trials (i.e., correct rejects & false alarms); RT was derived solely from correct Go trials. Furthermore, single-condition indicators were extracted from the impulsivity condition (i.e., accuracy, IES, and drift rate); whereas condition-difference indicators were derived from the difference between the impulsivity minus the sustained attention condition.

Table 1 *List of tasks, tasks' conditions and indicators derived per task*

EF task	Tasks' conditions		Indicators			
	Single[4]	Condition difference	RT	Acc	IES	DR
Flanker[1]	Incongruent	Incongruent - Congruent	*	*	*	*
Simon[1]	Incongruent	Incongruent - Congruent	*	*	*	*
Go/No-Go[1]	Impulsivity	Impulsivity – SA		*	*	*
Color-Shape[1]	Switch	Switch – Non-switch[5]	*	*	*	*
Gender-Smile[1]	Switch	Switch – Non-switch[5]	*	*	*	*
Trail Making Test[2]	Trail-B	Trail-B – Trail-A	*			
Spatial n-back[1]	2-back	2-back – 0-back	*	*	*	*
Backwards digit-span[3]	N/A	N/A		*		
Letter-Memory[3]	N/A	N/A		*		

Note. EF: executive functions. RT: response time; Acc: accuracy; IES: inverse efficiency score; DR: drift rate; SA: sustained attention; N/A: not applicable; [1]Computer reaction time-based task, RT and response accuracy are recorded; [2]Paper-pencil, time to completion task; [3]Computer accuracy-based task, only response accuracy is recorded. [4]Single task condition with greatest EF demands. [5]Switch – Non-switch trials mixed-block.

Cognitive flexibility tasks. On the Color-Shape and Gender-Smile switch tasks RT and accuracy were recorded; in addition, IES and drift rate were computed. Single-condition indicators were derived from the switch trials in the heterogeneous-mixed block; difference-condition indicators were derived from the difference between switch and non-switch trials in the heterogeneous-mixed block. Finally, because the Trail Making Test is a time to completion task, only response time is recorded, preventing the use of accuracy, IES or drift rate for that task. The single-condition indicator was derived from the time to completion of the Trail-B test. The condition-difference indicator was derived from the score difference was between task's conditions Trail-B minus Trail-A.

Working memory tasks. The Letter-Memory task is an accuracy-based measured, where response accuracy is collapsed across all trials (single-condition indicator). On the Backwards digit-span task, the longest sequence that was remembered correctly (e.g., 5 digits) was used as a measure of working memory span (single-condition indicator). As these two accuracy-based tasks are not built to contrast performance between different task conditions, a difference score could not be derived. On the Spatial n-back task, RT and accuracy were recorded, and in addition, IES and drift rate were computed. Single-condition indicators were derived from the 2-back condition (i.e., RT, accuracy, IES, and drift rate); difference score indicators were derived by considering performance in 2-back minus 0-back condition.

Statistical procedures

Raw data cleaning

For computer tasks in which response-time was recorded (see Table 1), trials with RT below 200 milliseconds were considered anticipatory responses and removed. For each participant, trials with a RT beyond ± 2.5 *SD* from within-subject's mean were removed.

Raw data transformation to indicators

RT. On RT-based tasks, mean RT was computed from correct trials per task condition.

Accuracy. On RT-based tasks, the proportion of correct responses was computed per task condition separately. For the Letter-Memory task, accuracy was computed across all trials, whereas for the Backward digit span, capacity was derived from the longest sequence correctly recalled.

IES (Townsend & Ashby, 1978). For RT-based tasks, individual IES scores were computed dividing **RT** by **Accuracy**.

Drift rate (Ratcliff, 1978). For RT-based tasks DDM parameters (drift rate, decision boundary, non-decision time) were computed for each task condition separately using the equations from the EZ-diffusion model (R code provided in, Wagenmakers et al., 2007 - Using **RT**, **Accuracy**, and RT *SD*).

Then, univariate analyses on each indicator were conducted to remove outlier data before fitting the models to the data. Values ± 3 *SD* from the sample mean were excluded from the analyses, this affected less than 1.5% of observations (Table S2 reports the % of missing data for each task after removing outlier values). No other data cleaning procedure was conducted.

Structural equation models

Indicator-based models. Two types of indicator-based models were considered. Single condition indicator-based models represent performance in those tasks' conditions requiring the greatest demands for executive control (see Table 2A, single-condition indicator-based models), for the four indicators examined in the present study (RT-based model, accuracy-based model, IES-based model, and drift rate-based model). Condition-difference indicator-based models represent the score difference between tasks' conditions (see Table 2B, condition-difference indicator-based models) for each of these four indicators (i.e., RT difference-based, accuracy difference-based, IES difference-based, drift rate difference-based). Note that for the remainder of the article, when we compare these two types of models, we will refer to them either as (i) single-condition indicator-based models, or as (ii) condition-difference indicator-based models, respectively.

Note that it was not possible to have models with the same indicator for all tasks because as described above, for some tasks RT was either not collected or collected only for some conditions (i.e., Letter-Memory, Backwards digit-span, Go/No-Go) and for others, time to completion was collected preventing proper assessment of accuracy and RT (i.e., Trail Making Test). Yet, although not homogeneous, each of the eight models has a dominant indicator across tasks, which was used to name the model.

Table 2A *List of tasks, single-condition indicator-based models and indicators per model*

Task	RT–based model	Accuracy–based model	IES–based model	Drift Rate–based model
Flanker[1]	RT incongruent trials	Accuracy incongruent trials	IES incongruent trials	DR incongruent trials
Go/No-Go[1]	Accuracy impulsivity trials	Accuracy impulsivity trials	IES impulsivity trials	DR impulsivity trials
Simon[1]	RT incongruent trials	Accuracy incongruent trials	IES incongruent trials	DR incongruent trials
Color-Shape[1]	RT switch trials	Accuracy switch trials	IES switch trials	DR switch trials
Gender-Smile[1]	RT switch trials	Accuracy switch trials	IES switch trials	DR switch trials
Trail Making Test[2]	Trails B seconds	Trails B seconds	Trails B seconds	Trails B seconds
Backwards Digit Span[3]	Span length	Span length	Span length	Span length
Spatial n-back[1]	RT 2-back trials	Accuracy 2-back trials	IES 2-back trials	DR 2-back trials
Letter Memory[3]	Accuracy	Accuracy	Accuracy	Accuracy

Note. [1]Computer reaction time-based task, RT and response accuracy are recorded; [2]Paper-pencil, time to completion task; [3]Computer task, only response accuracy is recorded. DR: drift rate.

Table 2B *List of tasks, condition-difference indicator-based models and indicators per model*

Task	RT difference – based model	Accuracy difference – based model	IES difference – based model	Drift Rate difference – based model
Flanker	RT difference (incongruent – congruent trials)	Accuracy difference (incongruent – congruent trials)	IES difference (incongruent – congruent trials)	DR difference (incongruent – congruent trials)
Go/No-Go	Accuracy difference (impulsivity block – sustained attention block)	Accuracy difference (impulsivity block – sustained attention block)	IES (impulsivity block – sustained attention block)	DR difference (impulsivity block – sustained attention block
Simon	RT difference (incongruent – congruent trials)	Accuracy difference (incongruent – congruent trials)	IES difference (incongruent – congruent trials)	DR difference (incongruent – congruent trials)
Color-Shape	RT difference (switch – non-switch trials)	Accuracy difference (switch – non-switch trials)	IES difference (switch – non-switch trials)	DR difference (switch – non-switch trials)
Gender-Smile	RT difference (switch – non-switch trials)	Accuracy difference (switch – non-switch trials)	IES difference (switch – non-switch trials)	DR difference (switch – non-switch trials)
Trail Making Test	Trail B – Trail A	Trail B–Trail A	Trail B–Trail A	Trail B – Trail A
Backwards Digit Span	Span length	Span length	Span length	Span length
Spatial n-back	RT difference (2-back trials – 0-back trials)	Accuracy difference (2-back trials – 0-back trials)	IES difference (2-back trials – 0-back trials)	DR difference (2-back trials – 0-back trials)
Letter Memory	Accuracy	Accuracy	Accuracy	Accuracy

Bootstrap resampling

Parametric bootstrap resampling with replacement was conducted to generate 5000 data sets of equal sample size and mean age to that of the original data set (*n* = 182). Then, for each of the eight indicator-based models described above, seven measurement models of executive functions (k = 7, Figure 3) were generated and fitted to the data. Note that these seven models are the same ones that Karr et al. (2018) tested in their simulation study. Thus, 56 different models (8 indicator-based models * 7 measurement models of executive functions) were fitted to each of the simulated 5’000 data sets. In total, 280’000 models (56* 5’000) were generated and analyzed. Fit indices were calculated for models that converged without any errors or warnings, also termed improper solutions (e.g., variance-covariance matrix not positive definite, negative residual variances, correlations larger than 1.0). The bootstrap analysis was conducted in R (version 3.6.1). The Lavaan package (Rosseel, 2012), was used to fit all the latent factor models to the data. Missing data was estimated with full information maximum likelihood method.

Bootstrap analysis – Model convergence, acceptance, and selection

The present study analyzed the simulated data in two different ways, following the procedure conducted in Karr et al (2018). First, we looked at the rate of model convergence, as well as the rate of model acceptance. The rate of model convergence represents the percent of models that converged across the 5’000 data sets, regardless of the fit indices. The rate of model acceptance analyzes the percent of models meeting the fitting thresholds (i.e., lenient and strict), among the models that converged. This first step enables to estimate the frequency that any of the models tested in the present study would (i) converge without any errors or (ii) meet the fitting thresholds proposed. Therefore, this first analysis enabled to determine the most suitable indicator-based model(s), in terms of rate of model convergence and model acceptance, to model executive functions with latent variable methods. Importantly, we aimed not only to identify the indicator(s) with the best psychometric properties to model executive functions, but also, we examined potential differences between single-condition indicator-based models and condition-difference indicator-based models.

Second, we investigated which measurement model of executive functions is preferred, among the seven measurement models tested for each of the different indicator-based models. More precisely, we looked at the probability that a given model is selected as the best model over alternative models, based on the direct comparison of their fit indices.

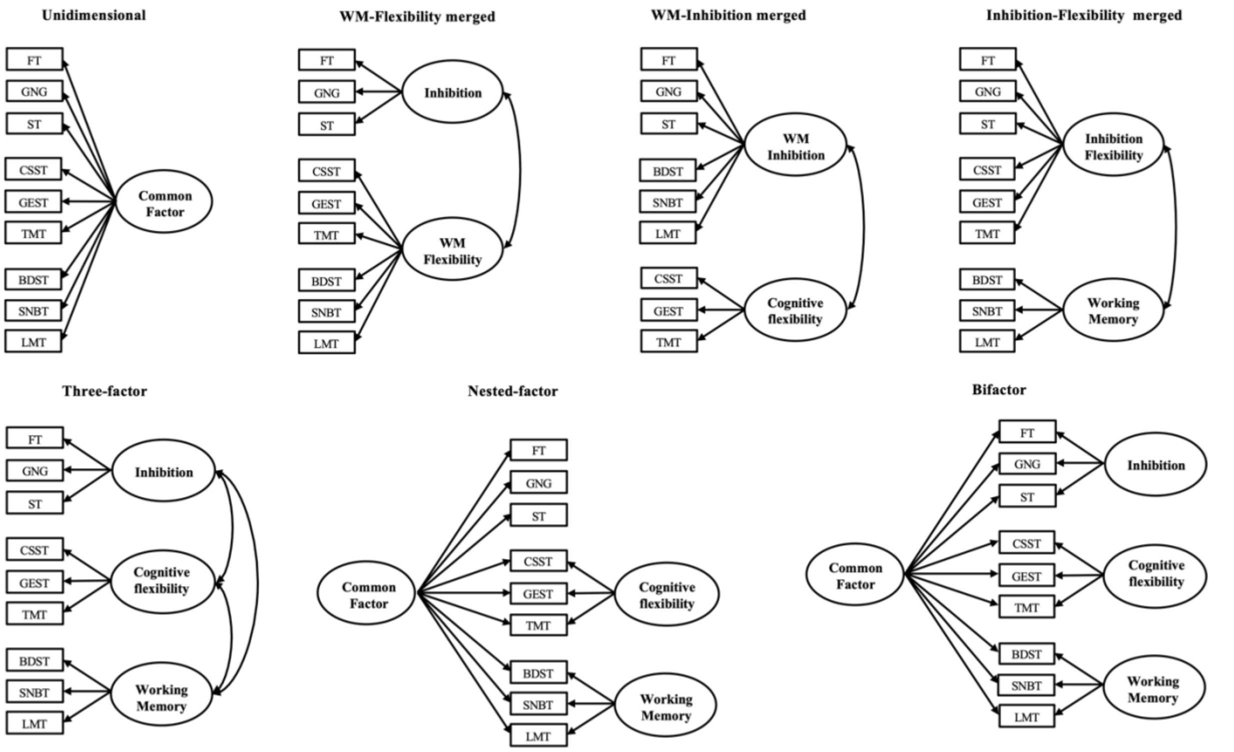

Figure 3. Illustration of measurement models of executive functions tested.
Note. Tasks' acronyms: FT: Flanker; GNG: Go/No-Go; ST: Simon; CSST: Color-Shape Switch; GEST: Gender-Smile Switch; TMT: Trail Making Test; BDST: Backwards digit-span; SNBT: Spatial n-back; LMT: Letter-Memory.

Model fit interpretation

Model acceptance. To determine the rate of model acceptance, goodness of fit to the data of the models tested was evaluated using the comparative fit index (CFI) and root mean square error of approximation (RMSEA). These two fit indices have a common metric and provide complementary information, as the CFI is an incremental fit index that compares an hypothesized model with a baseline model (i.e., a model in which no items covary) in terms of goodness of fit, whereas the RMSEA is an absolute fit index that assesses how far an hypothesized model is from a perfect model (Xia & Yang, 2019). Furthermore, the RMSEA favors parsimony since it penalizes model complexity, unlike the CFI (Hooper et al., 2008).

Importantly, these indices provide cutoffs thresholds that enable to determine whether a model has poor, acceptable or good fit to the data, and also can be interpreted in terms of their absolute fit value. Following Hu and Bentler (1999) recommendations, a model has acceptable fit to the data, if it has a CFI ≥ .90 and RMSEA ≤ .08 (lenient thresholds), and good fit to the data if it has a CFI ≥ .95 and RMSEA ≤ .05 (strict thresholds).

Model selection. To determine the probability of model selection, we assessed the fit of the models with two different indices, Akaike's information criterion (AIC; Akaike, 1973), and the Bayesian information criterion (BIC; Schwartz, 1978). Both indices are measures of comparative fit, which are meaningful only when used to compare different models (Kenny, 2015). Models with lower values indicate a better fit to the data. Both indices balance goodness-of-fit and complexity. Lack of parsimony is penalized according to the number of parameters of the model. The AIC index applies a linear penalty of two for every parameter estimated, whereas the BIC applies a bigger penalty to model complexity, since the BIC increases the penalty exponentially as model complexity increases (Vrieze, 2012).

The model selection analysis was conducted based on the estimation of the relative AIC weight (AIC_w) and BIC weight (BIC_w) of each model, a method proposed by Wagenmakers and Farrell (2004) for model selection. First, for each data set in which the seven measurement models of executive functions converged without warning/errors, we computed the difference in fit, ΔAIC and ΔBIC, between the best fitting model (indicated by the lowest AIC and BIC value) and the other models. Therefore, the best fitting model always has a ΔAIC (or ΔBIC) = 0, and the other models a ΔAIC (or ΔBIC) > 0. Then, we computed the AIC_w and BIC_w for each model using the equations provided in Wagenmakers & Farrell (2004). This method enables to estimate the probability that model M_i is the best model given data and the candidates models tested. Note that the relative model probabilities are normalized by dividing by the sum of the probabilities of all the models.

Data Availability Statement

Data and R code are available in the OSF website (Yangüez et al., 2022). Note that we provide the original preprocessed data (z-transformed), the bootstrapped data (i.e., 5000 data sets), and the R code. To reproduce the exact results published in the manuscript, use the bootstrapped data file.

Results

Model convergence

Single-condition vs Condition-difference indicator-based models' comparison

First, we examined potential differences in model convergence, as a function of the tasks' conditions that can be used to operationalize performance in executive functions tasks. Thus, indicator-based models were grouped as (i) single-condition indicator-based models versus (ii) condition-difference indicator-based models. Collapsed across all seven models, single-condition indicator-based models showed a remarkably higher rate of convergence ($\bar{x} =$ 66.63%, min = 2.2%, max = 100%), compared to condition-difference indicator-based models ($\bar{x} =$ 31.56%, min = 0.98%, max = 94.92%).

When we looked more in detail at differences between both groups across measurement models of executive functions, we observed that single-condition indicator-based models showed systematically higher rates of convergence compared to condition-difference indicator-based models (Figure 4). As a case example, the three-factor models showed a much higher rate of convergence on single-condition indicator-based models ($\bar{x} = 51.58\%$, min = 3.4%, max = 96.26%) compared to condition-difference indicator-based models ($\bar{x} = 11.01\%$, min = 2.82%, max = 19.24%). The same pattern was observed across the seven measurement models of executive functions (Figure 4).

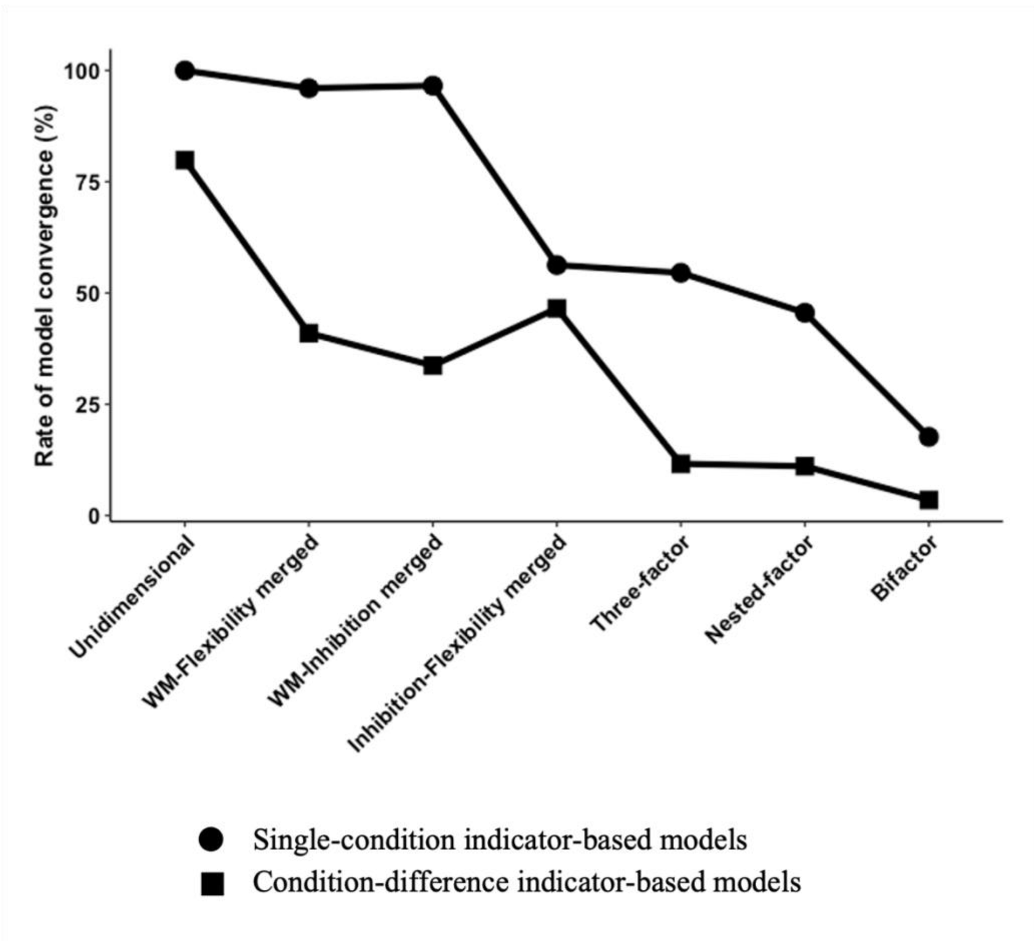

Figure 4. Average rate of convergence across single-condition vs. condition-difference indicator-based models, as a function of measurement models of executive functions.

Given this striking difference, the remainder of our analyses will focus exclusively on single-condition indicator-based models, which showed the best psychometric properties when it comes to convergence, that is, they showed a much lower percentage of improper solutions. Full results of model convergence and acceptance of condition-difference indicator-based models are provided in Table S3 (online supplementary material) for the interested reader. Importantly, Table S3 confirms that the poor performance of condition-difference indicator-based models does not hide a tradeoff between convergence and acceptance. Single-condition indicator-based models perform better based on their enhanced rates of acceptance when both CFI and RMSEA fit indices are taken into account. Furthermore, for the interested reader, Table S4 (online supplementary material) reports the percentage of measurement models that (i) converged (without warning), (ii) did not converge, and (iii) that converged with a warning message (e.g., negative variance; variance-covariance matrix not positive definite, etc.)

Impact of the indicator used

Table 3 lists the mean percent of models that converged across the 5’000 data sets, as a function of the indicator-based model tested for each of the seven measurement models of executive functions models. Within each measurement model, there were remarkable differences in model convergence depending on the indicator used. While unidimensional models all converged regardless of the indicator used, two factor models also showed high rate of convergence overall, but less so when RT was used as an indicator. As model complexity increased convergence rates not only decreased as expected, but surprisingly this effect was much more marked for RT-based and IES-based models than for accuracy-based and drift rate-based models. This information is illustrated in Figure 5 (thick black line). Furthermore, averaging across measurement models of executive functions, we observed remarkable differences between single-condition indicator-based models in model convergence. On average, accuracy-based models ($\bar{x} = 85.54\%$, min = 28.88%, max = 100%) showed the

greatest rate of model convergence followed closely by drift rate-based models ($\bar{x} = 81.93\%$, min = 24.12%, max = 100%), whereas IES-based models ($\bar{x} = 53.36\%$, min = 12.44%, max = 100%), and RT-based models ($\bar{x}$ = 45.70%, min = 2.2%, max = 99.98%) showed much lower convergence rates, as often failed to converge.

Table 3 *Percent of Models that Converged for 5,000 Bootstrapped data sets & Percent of Models that Meet CFI and RMSEA Lenient and Strict Criteria among Converged Models*

	Indicator-based model	Converged (%)	CFI ≥ .90 (%)	CFI ≥ .95 (%)	RMSEA ≤ .08 (%)	RMSEA ≤ .05 (%)
Unidimensional	RT- based	99.98	0	0	0.02	0
	ACC-based	100	0.28	0	1.56	0
	IES-based	100	1.56	0	2	0
	DR-based	100	46.28	6.18	38.94	2.72
	Mean	99.99	12.03	1.55	10.63	0.68
	Median	100	0.92	0	1.78	0
WM-Flexibility merged	RT- based	80.62	0.02	0	0.02	0
	ACC-based	99.94	20.29	2.4	30.34	2.3
	IES-based	97.7	4.03	0.1	3.73	0.02
	DR-based	99.88	83	33.9	74.89	19.02
	Mean	94.54	35.77	12.13	36.32	7.11
	Median	98.79	20.29	2.4	30.34	2.3
WM-Inhibition merged	RT- based	93.86	0	0	0.04	0
	ACC-based	99.88	4.91	0.2	10.01	0.22
	IES-based	98.06	3	0.02	2.9	0
	DR-based	98.92	65.93	16.24	55.44	7.3
	Mean	97.68	24.61	5.49	22.78	2.51
	Median	98.49	4.91	0.2	10.01	0.22
Inhibition-Flexibility merged	RT- based	2.2	0	0	0.91	0
	ACC-based	95.68	3.72	0.13	8.13	0.13
	IES-based	15.98	2.75	0	2.25	0
	DR-based	98.16	59.15	11.61	48.43	4.69
	Mean	53.01	21.87	3.91	19.60	1.61
	Median	55.83	3.72	0.13	8.13	0.13
Three-factor	RT- based	3.4	0	0	0	0
	ACC-based	96.26	51.96	12.2	54.87	8.48
	IES-based	12.44	7.88	0.16	4.98	0
	DR-based	94.2	91.42	47.32	80.83	26.09
	Mean	51.58	50.42	19.89	46.89	11.52
	Median	53.32	51.96	12.2	54.87	8.48
Nested-factor	RT- based	34.4	0.06	0	0	0
	ACC-based	78.16	34.34	3.79	27.71	1.54
	IES-based	34.72	7.32	0.06	2.36	0
	DR-based	58.22	86.95	36.1	63.21	12.47
	Mean	51.38	42.87	13.32	31.09	4.67
	Median	46.47	34.34	3.79	27.71	1.54
Bifactor	RT- based	5.44	0	0	0	0
	ACC-based	28.88	64.75	16.9	42.04	6.44
	IES-based	14.6	23.7	1.23	4.79	0.14
	DR-based	24.12	97.43	65.51	77.11	27.11
	Mean	18.3	62.0	27.9	41.3	11.2
	Median	19.36	64.75	16.9	42.04	6.44
Average	RT- based	45.70	0.01	0.00	0.14	0.00
	ACC-based	85.54	25.75	5.09	24.95	2.73
	IES-based	53.36	7.18	0.22	3.29	0.02
	DR-based	81.93	75.74	30.98	62.69	14.20
	Mean	66.63	27.17	9.07	22.77	4.24
	Median	94.03	6.12	0.15	4.89	0.08

Model acceptance

The second step of the analysis aimed to examine potential differences in model acceptance between single-condition indicator-based models. Thus, we looked at the percent of models meeting either lenient (CFI ≥ .90 or RMSEA ≤ .08) or strict fit thresholds (CFI ≥ .95 or RMSEA ≤ .05), which indicate acceptable and good fit to the data, respectively. Table 3 includes the percentage of models that met lenient and strict fit thresholds for each measurement model of executive functions, as a function of the single-condition indicator-based model tested. Note that the rate of model acceptance was computed among the models that converged. That is, when a model converged, we computed how often the model meets the lenient (or strict, respectively) thresholds for each fit index (CFI and RMSEA) separately. This information is visually represented on Figure 5.

Impact of the indicator used

As seen in Figure 5, the rate of executive functions model acceptance differed remarkably according to the specific indicator used (RT, accuracy, IES or drift rate). Model acceptance was defined based on two fit indices, CFI (in red), and RMSEA (in turquoise), for lenient (solid thin colored line) and strict thresholds (dashed thin colored line).

In sum, averaging across the seven measurement models of executive functions, drift rate-based models (Figure 5D) showed the highest rate of model acceptance based on both lenient (CFI: $\bar{x} = 75.74\%$; RMSEA: $\bar{x} = 62.69\%$) and strict thresholds (CFI: $\bar{x} = 30.98\%$; RMSEA: $\bar{x} = 14.20\%$). These rates are remarkably higher than those of any other indicator-based model. More precisely, RT-based models (Figure 5A) showed extremely low rates of acceptance, for both lenient (CFI: $\bar{x} = 0.01\%$; RMSEA: $\bar{x} = 0.14\%$) and strict thresholds (CFI: $\bar{x} = 0\%$; RMSEA: $\bar{x} = 0\%$). The same pattern was observed on IES-based models (Figure 5C), for both lenient (CFI: $\bar{x} = 7.18\%$; RMSEA: $\bar{x} = 3.29\%$) and strict thresholds (CFI: $\bar{x} = 0.22\%$; RMSEA: $\bar{x} = 0.02\%$). Furthermore, although accuracy-based models (Figure 5B) showed a slightly higher rate of convergence than drift rate-based models (as discussed in the previous section), they showed only modest to low rates of acceptance for both lenient (CFI: $\bar{x}$ = 25.75%; RMSEA: $\bar{x}$ = 24.95%) and strict thresholds (CFI: $\bar{x}$ = 5.09%; RMSEA: $\bar{x}$ = 2.73%).

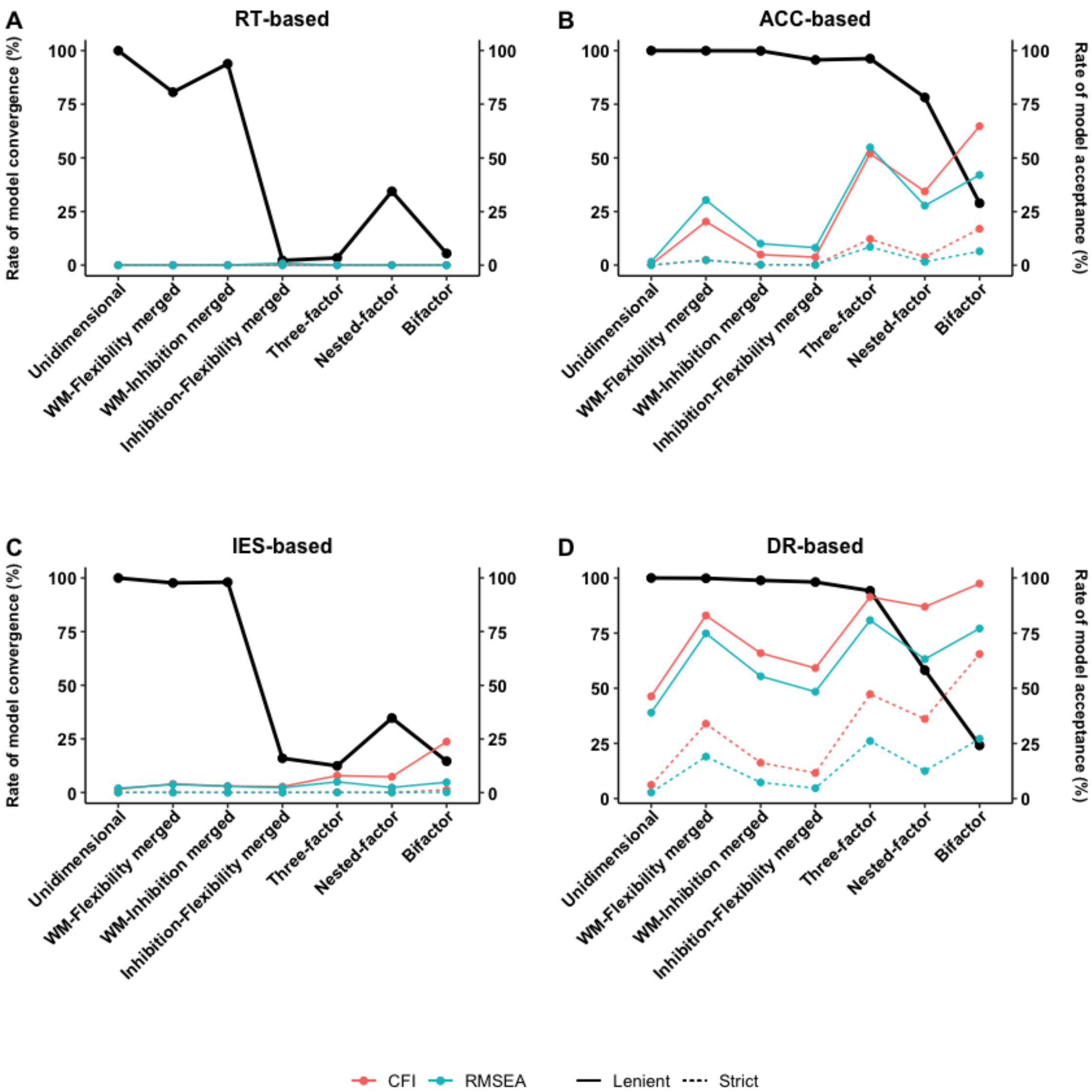

Figure 5. Rate of model convergence and acceptance as a function of single-condition indicator-based models.

Note. Figure 5A: RT-based models; Figure 5B: accuracy-based models; Figure 5C: IES-based models; Figure 5D: drift rate-based models; Rate of convergence: black solid lines; Rate of acceptance (lenient and strict thresholds): CFI >.90 = red solid lines; CFI >.95 = red dashed lines; RMSEA < .08 = turquoise solid lines; RMSEA < .05 = turquoise dashed lines.

Relationship between model convergence, acceptance, and complexity

Figure 5 illustrates a tight relationship between model convergence (thick black line) and acceptance (thin colored lines). As expected, the most complex models (e.g., three-factor, nested-factor and bifactor) converged less often than simpler models (e.g., unidimensional or two-factor); however, when they converged, they showed better fit to the data as indicated by their higher rate of acceptance. This expected trend was seen for both CFI & RMSEA fit indices, as well as for both fit thresholds (see Table 3). This pattern is best illustrated by accuracy-based models (Figure 5B) and drift rate-based models (Figure 5D) due to their higher rates of acceptance; indeed, RT-based (Figure 5A) and IES-based models (Figure 5C) showed too poor rates of acceptance.

To conclude, drift rate-based models and accuracy-based models showed the highest (and similar) rates of convergence across the seven measurement models of executive functions; however, drift rate-based models showed a much higher rate of acceptance and thus, better fit to the data. Therefore, drift rate-based models appear preferable to accuracy-based models, as only the former achieve high rates of both convergence and acceptance. This work also makes clear that both RT-based and IES-based models show overall only modest rates of convergence and very low rates of acceptance, questioning the usefulness of these indicators.

Model selection

The last step of the analysis examined how the choice of indicator may impact which of the seven measurement models of executive functions is preferred. To do so, firstly, we selected those data sets in which the seven measurement models of executive functions converged, to ensure that even models with lesser convergence rates (e.g., nested-factor and bifactor models) be both run on the same samples and equally represented in the model selection analysis. Only drift rate-based (n = 996) and accuracy-based (n = 1320) models provided enough data sets (see Table 4).

AIC weights (AIC_w) and BIC weights (BIC_w) were then computed comparing the seven measurement models of executive functions, using separately accuracy and drift rate as indicator. Figure 6 shows the distribution of these weights. The reader can find in Table S5 (online supplementary material) the mean AIC_w and BIC_w, as a function of the seven measurement models of executive functions.

Table 4 *Data sets where the seven executive functions measurement models converged out of 5000 data sets*

Indicator-based model	Data sets (*n*)	Data sets (%)
RT-based	1	0.02
ACC-based	1320	26.4
IES-based	14	0.28
DR-based	996	19.32
RTdiff-based	12	0.24
ACCdiff-based	36	0.72
IESdiff-based	0	0
DRdiff-based	10	0.2

Note. RT: response time; ACC: accuracy; DR: drift rate; IES: inverse efficiency score; diff: difference

Based on AIC weights, the three-factor model showed the strongest evidence as the best model. More precisely, on drift rate-based models (Figure 6A), the average AIC weight of the three-factor model is of .521, with the next best model, two-factor WM-Flexibility merged, being at .212. Similarly, based on accuracy-based models (Figure 6B), the average AIC weight of the three-factor model is of .711, whereas the weight of the next best model (i.e., nested-factor model) is as low as .108.

BIC weights point to the two-factor model with WM-Flexibility merged as the one with the greatest evidence for drift rate-based models (.541) followed by the three-factor model (.212), whereas for accuracy-based models the strongest evidence is observed for the three-factor model (.595) followed by the two-factor model with WM-Flexibility merged (.314) (Figure 6C and 6D, respectively).

In sum, despite the AIC and BIC fit indices yielding a slightly different pattern of results, the three-factor model appears, overall, as the best model among the seven measurement models tested, which is also the best fitting model based on CFI and RMSEA indices (see table S6 online supplementary material, *Mean CFI and RMSEA, and percent of measurement models that meet lenient thresholds*).

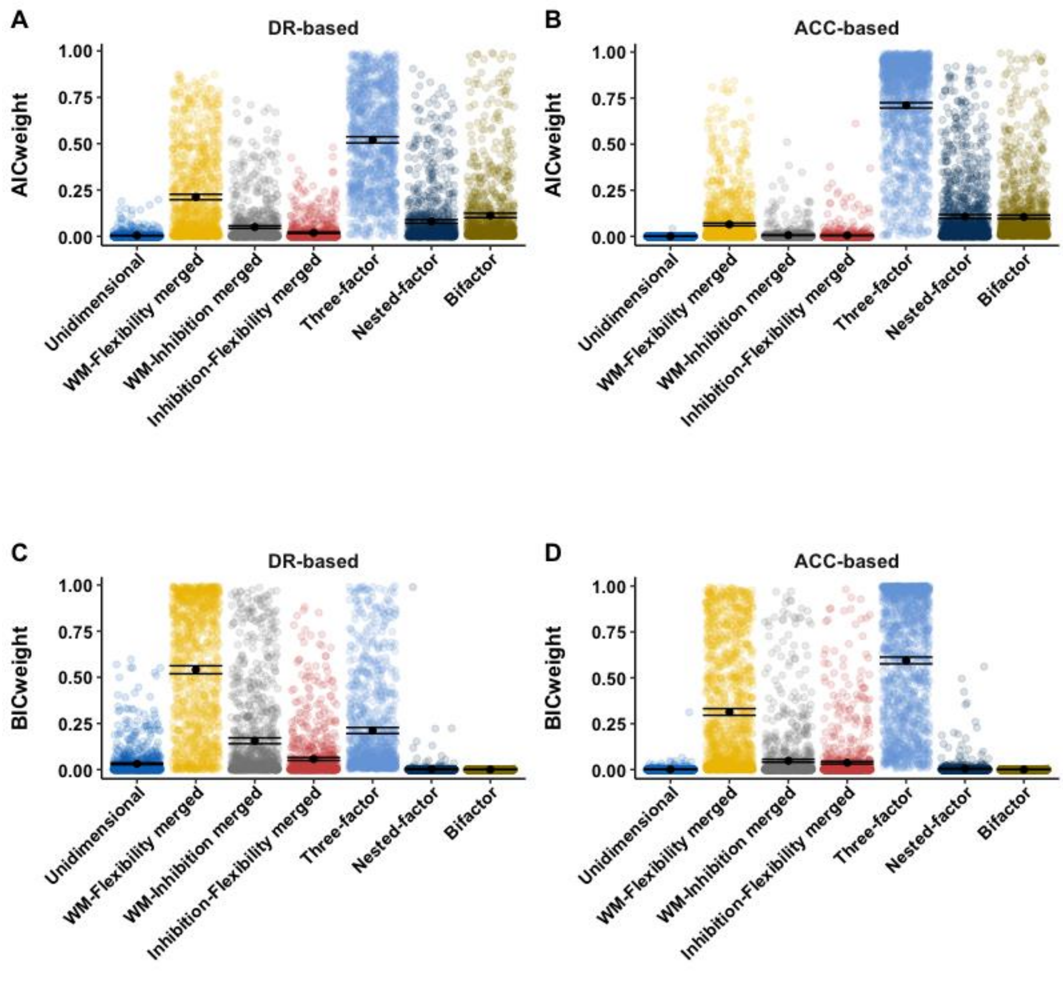

Figure 6. AIC weights and BIC weights of measurement models of executive functions, as a function of drift rate-based and accuracy-based models.

Note. Figure 6A: AIC weights drift rate-based models; Figure 6B: AIC weights accuracy-based models; Figure 6C: BIC weights drift rate-based models; Figure 6D: BIC weights accuracy-based models. Black dots represent mean weights; top and bottom horizontal black bars represent bootstrapped 95% confidence intervals. Note that the relative model probabilities are normalized by dividing by the sum of the probabilities of all the models.

Three-factor model - mean fit indices, inter-factor correlations, and factor-loadings

A recurrent issue when using CFA concerns the appropriateness of the tasks selected to assess the different latent constructs. While we have shown through model selection that the three-factor model is to be preferred, the values of the respective factor loadings from each task provide additional information about the construct validity of each task. Table 5 provides the mean CFI and RMSEA and mean inter-factor correlations of three-factor accuracy-based and drift rate-based models. Table 6 provides the mean factor-loadings for both indicator-based models, as well as two coefficients (i.e., omega ω and H index) that inform about the reliability of the latent factors from both indicator-based models. Furthermore, the reader can find in Table S7 (online supplementary material) the mean factor loadings of the eight indicator-based three-factor models, as well as the omega ω and H index reliability coefficients of the latent factors from each indicator-based model.

The drift rate-based three-factor model on average showed good fit to the data (CFI: $\bar{x} = .95$; RMSEA: $\bar{x} = .06$). The observed correlations between the three latent constructs on average were high, inhibition-cognitive flexibility ($\bar{r} = .75$), inhibition-working memory ($\bar{r} = .76$), working memory-cognitive flexibility ($\bar{r} = .79$), and were stronger compared to those observed in three-factor accuracy-based model. Importantly, all the tasks loaded significantly into their corresponding latent construct, showing moderate-to-strong factor loadings ($\bar{x} = .61$; min = .41; max = .75). The accuracy-based three-factor model on average showed acceptable fit to the data (CFI: $\bar{x} = .90$; RMSEA: $\bar{x} = .08$). The three constructs on average were strongly correlated, inhibition-cognitive flexibility ($\bar{r} = .59$), inhibition-working memory ($\bar{r} = .56$), working memory-cognitive flexibility ($\bar{r} = .62$). All the tasks loaded significantly into their corresponding latent construct and more importantly, they showed moderate-to-strong factor loadings ($\bar{x} = .60$; min = .43; max = .76).

In sum, both indicator-based models achieved satisfactory factor loadings, except for the Backwards digit-span task and the Letter-Memory task (working memory latent construct). In addition, both indicator-based models included latent factors that achieved levels of reliability remarkably higher compared to any other indicator-based model (see Table S7).

Table 5 *Three-factor Drift rate-based and Accuracy-based models. Mean Fit Indices and Inter-Factor Correlations and Standard Deviation.*

Indicator-based model	**CFI**		**RMSEA**		**Inhibition – Cognitive Flexibility**		**Inhibition – Working Memory**		**Cognitive Flexibility – working Memory**	
	$\bar{x}$	*SD*	$\bar{x}$	*SD*	*r*	*SD*	*r*	*SD*	*r*	*SD*
DR-based	0.95	± .03	0.06	± .02	0.75	± .08	0.76	± .10	0.79	± .11
ACC-based	0.90	± .04	0.08	± .02	0.59	± .12	0.56	± .11	0.62	± .15

Note. This information corresponds to three-factor models that converged without warnings/errors out of 5000 data sets. DR-based (n) = 4710 data sets; ACC-based (n) = 4813 data sets.

Table 6 *Three-factor Drift Rate and Accuracy-based models*. Mean Factor Loadings and Latent Factors' Reliability Coefficients*

EF Factor – reliability coefficients	**Task – factor loadings**	**DR-based model (n = 4710)**	**ACC-based model (n = 4813)**
	Flanker	0.71 (± .06)	0.71 (± .07)
Inhibition	Simon	0.65 (± .06)	0.56 (± .07)
	Go/No-Go	0.74 (± .06)	0.74 (± .07)
Omega		0.75 (± .03)	0.71 (± .04)
H index		0.76 (± .04)	0.75 (± .05)
	Color-Shape switch	0.50 (± .08)	0.50 (± .09)
Cognitive flexibility	Gender-Smile switch	0.67 (± .07)	0.60 (± .10)
	Trail Making Test	0.58 (± .07)	0.61 (± .09)
Omega		0.61 (± .05)	0.60 (± .05)
H index		0.63 (± .05)	0.63 (± .06)
	Backwards Digit-Span	0.43 (± .08)	0.46 (± .09)
Working memory	Spatial n-Back	0.75 (± .08)	0.76 (± .10)
	Letter-Memory	0.41 (± .07)	0.43 (± .08)
Omega		0.55 (± .06)	0.57 (± .06)
H index		0.65 (± .08)	0.67 (± .11)
	Mean factor loadings	0.61	0.60
	Median factor loadings	0.63	0.60

Note. * Models that converged with no warning and errors. EF: executive functions; DR: drift rate; ACC: accuracy. In brackets (): standard deviations.

Discussion

The main objective of the present study was to investigate to what extent measurement models of executive functions are sensitive to different methodological practices proposed in the literature to operationalize executive functions. The heterogeneity of methods addressed in the present study concerns both, the indicators (e.g., RT, accuracy, etc.), as well as the tasks' conditions (score in single conditions vs score difference between tasks conditions) employed to operationalize executive functions. By examining the impact of different operationalizations of executive functions on important aspects of model replicability, such as the rate of model convergence and acceptance, the present study documents four notable findings, which provide novel insights regarding better practices for the modeling of the structure of executive functions.

The first striking finding is the remarkable differences in the rate of model convergence, depending on the tasks' conditions used to assess executive functions. More precisely, the use of single-condition indicators, which reflect performance in the tasks' conditions most representative of the component under study (e.g. incongruent condition in a Flanker task, inhibition component), led to a convergence rate remarkably higher compared to condition-difference indicators, which reflect the difference in performance between the task condition representative of the component under study, and a baseline task condition (e.g. incongruent minus congruent condition in a Flanker task). The subtraction method is rooted in the seminal work of Donders and was once argued to be essential to subtract effects of no interest (Donders, 1868; for a review, see Roelofs, 2018). The present work, however, confirms that difference scores have rather poor psychometric properties, in line with previous claims (Draheim et al., 2019; Griffin et al., 1999; Hughes et al., 2014; Miller & Ulrich, 2013; Paap & Sawi, 2016). Our findings suggest that the difference score method should be avoided when modeling executive functions with SEM techniques, such as CFA. The second most striking finding of the present study is that measurement models that included RT-based measures fared quite poorly as compared to measurement models that included accuracy-based measures, which showed much greater acceptance rates. A third finding is that drift rate, modeled by the EZ-diffusion model, was the indicator that showed the best psychometric properties to model executive functions, in terms of both model convergence and acceptance. A fourth main finding is that when considering models that converge, the three-factor model remains the most preferred model based on the direct comparison of the fit indices of the seven measurement models of executive functions tested. Finally, measurement models that included drift rate as the main indicator, showed comparatively moderate to high contribution (i.e., factor loadings) from most of the tasks into their corresponding latent construct, providing a path forward for re-analysis of existing data set or future ones, as the tasks used in the present study tend to be commonly used in the field of executive functions.

The advantage of using single scores over difference scores in measurement models of executive functions

One of the most robust findings of the present study was the remarkable difference in model convergence between single versus condition difference indicator-based models. The advantage of single-condition indicator-based models most likely owes to the fact that latent variables from structural equation models are defined by what their indicators have in common (MacCallum & Austin, 2000). As single scores are expected to show greater common variance than difference scores (for a review about RT and RT difference, see Draheim et al., 2019), the use of the former is seen to result in higher convergence of the models, hence, a lower probability of improper solutions. Although expected, the systematicity of that effect is notable.

The use of the difference score method has produced robust experimental effects, such as the well-known Stroop effect (Stroop, 1935), Simon effect (Simon & Rudell, 1967), Flanker effect (Eriksen & Eriksen, 1974), or task-switch costs (Monsell, 2003). However, robust experimental effects often fail to produce reliable individual differences in cognition (Hedge et al., 2018). This phenomenon likely occurs due to distinct nature of two historical approaches in psychological research, experimental research versus correlational research. The former aims to characterize the cognitive mechanisms underlying responses to different experimental manipulations (e.g., within-subject variance), whereas the latter aims to characterize the cognitive mechanisms underlying inter-individual differences in cognitive processing (e.g., between-subjects variance). Recently, Hedge et al. (2018) assessed in three studies the reliability of seven classical cognitive tasks using different scoring methods (i.e., mean RT, RT difference, accuracy rate, and accuracy difference), with most tasks showing test-retest reliabilities below .70 when performance was operationalized through the difference score method. Thus, the pattern of low convergence of condition-difference indicator-based models is likely due to the psychometric issues of difference scores reported in the literature by Hedge et al. (2018), in line with other claims (Caruso, 2004; Draheim et al., 2016, 2019). Accordingly, several studies have consistently shown that RT difference measures tend to show weaker association with other variables of interest, compared to the single RT measures from which RT differences are computed, as discussed in the introduction (e.g., Hughes et al., 2014; Paap & Sawi, 2016; Salthouse et al., 1998; Siegrist, 1997). The main reason is that the subtraction method removes part of the common variance between the two variables from which the score difference is computed increasing the proportion of error variance of this sort of measures (Cronbach & Furby, 1970; Hedge et al., 2018).

To shed further insight into the psychometric issues of condition-difference indicator-based models, we looked at the factor loadings of the tasks' indicators in the three-factor model, for each indicator-based model (see online supplementary material, Table S7). Factor loadings of condition-difference indicator-based models were, on average rather low (RT difference: $\bar{x}$ = .38); Accuracy difference: $\bar{x}$ = .40; IES difference: $\bar{x}$ = .38; Drift rate difference: $\bar{x}$ =

.34), compared to most single-condition indicator-based models (RT: $\bar{x} = .47$; Accuracy: $\bar{x} = .60$; IES: $\bar{x} = .55$; Drift rate: $\bar{x} = .61$). Lower factor loadings have been associated with convergence issues regardless of sample size (Gagné & Hancock, 2006; Marsh et al., 1998). Thus, the systematic lower factor loadings observed in condition-difference indicator-based models, possibly account for their convergence issues. It is also in line with a recent investigation that reported a pattern of factor loadings that differed between difference scores and single scores, when modeling tasks such as the Stroop, Simon, Flanker, Global-Local ones with CFA (Rey-Mermet et al., 2021). In addition, the latent factors from condition-difference models showed consistently lower reliability (i.e., omega ω and H index) compared to those from single-condition indicator-based models. This is a matter of concern because many latent variable studies still use difference scores, such as RT difference, to operationalize executive functions in their RT-based tasks (e.g., Agostino, Johnson, & Pascual-Leone, 2010; Brydges et al., 2014; Duan, Wei, Wang, & Shi, 2010; Xu et al., 2013). Although less commonly used, accuracy difference has the same psychometric issues than RT difference, but the former is more prone to reduced variance issues due to ceiling effects (Wang et al., 2008).

In sum, many paradigms in psychology still rely on difference score methods following a long tradition started with Donders. While this is certainly a valuable approach for a number of research questions, in the context of latent variable research and CFA, our work demonstrates it is a poor methodological choice, which compromises the quality of analyses downstream.

Impact of indicators on model convergence and acceptance of measurement models of executive functions

The poor psychometric properties of RT-based compared to accuracy-based measures

RT-based models showed moderate rates of convergence, and very low rates of acceptance for both fit index and thresholds. Accuracy-based models, on the other hand, show greater convergence rates and low-to-moderate rates of acceptance, suggesting better psychometric properties than RT-based measures. Note that the relative low rate of acceptance of RT-based and accuracy-based models was to some extent expected given the low rates of acceptance across executive functions models reported by Karr et al. (2018), as most of the studies included in their re-analysis used RT or accuracy-based measures to operationalize executive functions (e.g., Carlson et al., 2014; Lee et al., 2013; Lerner & Lonigan, 2014; Miller et al., 2012; Monette et al., 2015; Rose et al., 2012; Wiebe et al., 2011).

The psychometric issues of RT-based measures are likely due to a combination of factors, such as the low reliability of RT-based measures and their sensitivity to speed-accuracy trade-offs. This is not the first work to highlight such weaknesses. For example, Miller and Ulrich (2013) proposed a model to investigate the psychometric properties of mean RT and RT difference to predict individual differences on these measures. Their model estimated three different parameters underlying RTs, that is, (i) common processing speed across tasks, (ii) processing speed for individual tasks, and (iii) residual differences in RT related to neither general nor task-specific processing speed. Their model showed that mean RT can be reliable across tasks, provided that the model parameters show a reasonable amount of variability. As a consequence, mean RT reliability can come to depend on a parameter of no interest for researchers such as the residual differences in RT, a state of affair which is problematic. Interestingly, their model showed that the mechanisms underlying RT during decision making are far more complex that what is commonly assumed in the literature, in line with recent views (Frischkorn & Schubert, 2018; Ratcliff et al., 2016). A well-known source of that complexity arises from the sensitivity of RT-based measures to speed-accuracy trade-off, whereby individuals adopt different response strategies emphasizing speed over accuracy and vice-versa (Starns & Ratcliff, 2012). Research exploring individual or developmental differences in cognition based on RT analyses might be particularly impacted by such trade-offs (Draheim et al., 2016; Hertzog et al., 1993; Yang et al., 2015), with the potential for misleading conclusions due to the complex interplay between speed and accuracy.

Accuracy-based models although showing higher rates of convergence, displayed rather low-to-moderate rates of acceptance. Accuracy-based measures suffer from two weaknesses. First, they are often subject to speed-accuracy trade-offs, and in doing so confound information processing efficiency with strategic issues of boundary setting (Drugowitsch et al., 2015; Maris & van der Maas, 2012; Ratcliff & Rouder, 1998; Starns & Ratcliff, 2012). Second, accuracy-based models often suffer from a lack of sensitivity with most individuals operating around a narrow range of high accuracies. It has been noted before that accuracy-based measures are reliable for individual differences research, only if individuals make enough errors during the task and thus, there exist sufficient between subjects' variability in the rate of errors (Wang et al., 2008). Note this is the case of the present study. Although the level of accuracy (%) on our RT-based tasks was relatively high in the most demanding tasks' conditions (e.g., incongruent, switch, 2-back trails), the variability was quite large even after removing univariate outlier values (Flanker: $\bar{x} = 87$ %, range = 56-100 %; Simon: $\bar{x} = 88$%, range = 53-100 %; Go/No-Go: $\bar{x} = 93$ %, range = 74-100 %; Color-Shape switch: $\bar{x} = 82$ %, range = 50-99 %; Gender-Smile switch: $\bar{x} = 87$ %, range = 55-99 %; N-back: $\bar{x} = 76$ %, range = 40-100 %. Nevertheless, despite their low-to-moderate acceptance rates, it is important to point out that the most likely accuracy-based model (i.e., three-factor), on average showed acceptable fitting to the data. Furthermore, all the tasks showed moderate-to-high factor loadings, except for the Backwards digit-span task and Letter-Memory updating task, and more importantly, most of the tasks showed similar factor loadings, which indicates that the three latent factors of the model reflected to a greater or lesser extend common variance across the tasks. This is in line with accuracy-based measure having been useful to investigate CFA measurement models of executive functions (e.g., Agostino et al., 2010; Alfonso & Lonigan, 2021; Brocki & Tillman, 2014; Carlson et al., 2014; Lerner & Lonigan, 2014; Masten et al., 2012; Monette et al., 2015; Willoughby et al., 2012), working memory (Engle, 2018; Luck & Vogel, 2013; Waris et al., 2017; Wilhelm et al., 2013), and of intelligence (Conway et al., 2002; Rey-Mermet et al., 2019). Moreover, accuracy-based and capacity measures have been the measures of excellence in the broad field of working memory (Engle, 2018; Luck & Vogel, 2013; Waris et al., 2017; Wilhelm et al., 2013). This is the case during clinical and educational evaluations with the forward and backward digit span task, as part of standard batteries such as the WAIS (Wechsler, 1981) or the WISC (Wechsler, 1991), as well as in most laboratory experiments, albeit using more sophisticated forms of span tasks, such as the operation span (Kane et al., 2004). The introduction about 20 years ago of change detection tasks to measure working memory capacity point to possible changes. In particular, recently several authors have proposed to go beyond capacity, as measured in terms of memory slots (Rouder et al., 2011; Zhang & Luck, 2008), to rather characterize working memory in terms of processing efficiency (Lew & Vul, 2015; Ma et al., 2014).

In sum, our results confirm that the use of accuracy-based models, despite the two weaknesses described above, is a sound choice to model executive functions based on (i) their high convergence rates, and (ii) the fact that the most likely model of executive functions, which also tends to be the most accepted model in the literature, showed acceptable fitting to the data, and moderate-to-high factor loadings.

The promising psychometric properties of Drift Rate to model executive functions with latent variable methods

An issue when considering just RTs or accuracy is that neither fully capture behavioral performance as illustrated by the speed-accuracy trade-off discussed earlier. To address this issue, other indicators have been developed, such as the IES, which is often used in the developmental or aging literature (Vandierendonck, 2017), or drift rate, a measure rarely considered in the executive functions' literature.

The present work tested the IES, a measure that combines speed and accuracy in a single metric in an effort to mitigate the issue of speed-accuracy trade-offs (Townsend & Ashby, 1978). IES-based models on average showed moderate rates of convergence and very low rates of acceptance. These results were very similar to the ones observed on RT-based models, which was not surprising since an individual's IES can be considered as the RT corrected by the proportion of errors committed.

A unique contribution of the present study is to show that drift rate has excellent psychometric properties to model executive functions with latent variable methods, such as CFA. Drift rate-based models were the most stable in terms of model convergence and acceptance, showing moderate-to-high rates of convergence and acceptance, across all measurement models of executive functions, regardless of fit index (CFI or RMSEA) and thresholds (lenient or strict).

Drift rate is an appealing measure as it can be relatively easily computed from the EZ-diffusion model, provided each participant's response accuracy, mean RT and RT standard deviation is available (Wagenmakers et al., 2007). By acknowledging that speed and accuracy arise from the same underlying generative process of integration to a bound, diffusion models allow to assess the rate at which information processing accumulates, or the quality of information processing, separately from the height of the bound to be reached for a decision to be triggered. In doing so, drift rate provides a better estimation of information processing quality than RT or accuracy separately (Ratcliff et al., 2016). Accordingly, drift rate, by capturing sensitivity to information processing, appears as a more valid representation of the executive processes evaluated than using RT-based or accuracy-based measures, which are conflated with other processes such as response conservativeness or non-decision time.

Although seldom used in executive functions latent variable research, the reliability and stability of DDM parameters has been previously documented. The within-session and between-session reliability of DDM parameters in a lexical decision task was investigated by Yap et al. (2012 - n = 819) and confirmed by Lerche and Voss (2017 - n = 105). More precisely, Yap et al. (2012) reported that the three parameters of the DDM of greatest psychological interest showed excellent within-session reliability (drift rate: .81; boundary separation: .91; non-decision time: .93), and acceptable between-session reliability (drift rate: .69; boundary separation: .71; non-decision time: .72). Moreover, some works point to the criterion validity of diffusion model parameters, in particular drift rate, in the context of individual differences in intelligence. These studies have shown that a drift rate latent factor (derived from non-executive RT-based tasks) is associated with intelligence, thus suggesting that individuals with larger drift rate show greater scores in tests of intelligence (Lerche et al., 2020; Ratcliff et al., 2010; Schmiedek et al., 2007; Schmitz & Wilhelm, 2016). Furthermore, Schmiedek et al. (2007) and Schmitz and Wilhelm (2016) included measures of working memory capacity, and both studies modeled latent constructs for each of the three main parameters of the DDM (i.e., drift rate, boundary separation, and non-decision time). Interestingly, the drift rate latent construct was the main predictor of working memory capacity and intelligence latent constructs, whereas boundary separation and non-decision time showed very low associations with intelligence. Therefore, the better psychometric properties shown by drift rate in the present study are in line with the results from previous studies, which have highlighted the reliability and validity of drift rate for research in cognitive psychology.

Finally, post-hoc analyses, not included in the present study, were performed to examine the convergence and acceptance rates of measurement models of executive functions when drift rate was replaced by, either boundary separation or non-execution time, the other two parameters from the EZ-diffusion model. These models showed extremely poor convergence and acceptance rates, which was not surprising given that drift rate is the main parameter of the DDM capturing the quality of information processing. Accordingly, drift rate has been consistently associated with cognitive ability, such as fluid intelligence or working memory (Lerche et al., 2020; Ratcliff et al., 2010; Schmiedek et al., 2007; Schmitz & Wilhelm, 2016).

The interplay between model complexity and the rate of convergence and acceptance

Our results indicate that the simplest models (i.e., unidimensional, two-factor) despite converging more often, on average showed low rates of model acceptance, whereas the most complex models showed an opposite pattern. These results are expected and in line with the re-analysis of published models from Karr et al. (2018), who observed a similar trade-off between model convergence/acceptance and model complexity in both children/adolescent and adult samples.

The likelihood that a model converges and fits well the data depends on multiple factors, but mainly on a complex interplay between sample size, model parameters, and the level of commonality between the variables that form the model (Gagné & Hancock, 2006; Kyriazos, 2018; MacCallum et al., 1999; Wolf et al., 2013). These are important aspects of the study design, which along with statistical power, must be considered by researchers prior to conducting their study. More complex models, which have more parameters, often need larger samples to converge than simpler models (Green & Yang, 2018; Kline, 2016; Nicolaou & Masoner, 2013). In our case, the two most complex models, the nested-factor and bifactor models, had an important difference compared to the simpler models; they both included a common factor and specific sub-factors. Thus, the indicators had to load at the same time in the common factor and their corresponding sub-specific factor (note that in the nested-factor model, indicators from inhibition tasks loaded only in the common factor). There are several reasons that might explain why nested-factor and bifactor models showed a remarkably lower rate of convergence. A first reason concerns sample size; our study may be underpowered to identify these models with high frequency. Unlike in other domains of psychology, the field of latent variable research has not yet fully tackled how to best determine the sample size prior to conducting a study. There exist recommendations about how to estimate the sample size based on Monte Carlo simulation studies. For instance, Tanaka (1987) suggested a ratio 5:1 between *N* and parameters, whereas Bentler & Chou (1987) suggested a ratio of 10:1. More recently, Wolf et al. (2013), conducted a series of Monte Carlo simulations to understand sample size requirements, as a function of model type, number of factors and indicators, strength of the factor loadings and the amount of missing data. Their results show that the sample size requirements are far more complex than the recommendations from Tanaka (1987) and Bentler & Chou (1987). They observed that sample size requirements ranged from 30 cases (one-factor model with four indicators loading at .80), to 460 cases (two-factor model with three indicators per factor loading at .50). Importantly, they pointed out that inter-indicator correlations and the factor loadings also play an important role on statistical power and model identification, beyond sample size. Thus, our two most complex models, nested-factor and bifactor models, may be difficult to identify due to the complex interplay between sample size, the number of model parameters, the inter-indicator correlations, and the factor loadings from each indicator into both the common factor and their corresponding specific sub-factor.

The rate of acceptance followed an opposite pattern. That is, more complex models, when they converged, tended to have higher rates of acceptance than simpler models. These results are consistent across indicator-based models. The observed higher acceptance of two-factor and three-factor models over unidimensional models was expected given that an important number of studies have shown that executive functions tend to organize as a two-factor or three-factor structure during middle childhood (Brydges et al., 2014; Lehto et al., 2003; Rose et al., 2012). The high acceptance of nested-factor and bifactor models was less expected, as the use of these models is not an extended practice in executive functions research. The nested-factor model has been proposed in adolescents and adults studies (e.g., Friedman, Miyake, Robinson, & Hewitt, 2011; Friedman et al., 2008), and one study in children 9-to-12 years old (Sluis et al., 2007). The bifactor model reported in the present study and in Karr et al. (2018) is uncommon in children's executive functions research; indeed, a bifactor model based on a battery of tasks tapping inhibition, working memory and cognitive flexibility is rarely considered (for an exception see Yangüez et al., 2021). The higher acceptance of these two models must be taken with caution since there are several concerns about the tendency of these models to overfit the data (Bonifay et al., 2017; Karr et al., 2018; Murray & Johnson, 2013; Sellbom & Tellegen, 2019). Indeed, as noted by Hancock and Mueller (2008), a model with better fitting does not necessarily represent the true model for the population, but perhaps it is just a model that captures the data well, thanks in part to their ability to overfit the data (Preacher et al., 2013). Re-analyses with models that have a better balance between convergence and acceptance seems to be a healthy practice to adopt for the field. Unfortunately, rarely papers report on the necessary simulations to estimate the rate of convergence, when applying CFA to more practical ends. The present work calls for a more systematic assessment and report of convergence rates as a given model structure is chosen.

Impact of tasks' operationalization on model selection

A concern raised by the present study is that of the dependence of the most likely model on how executive functions are operationalized. Given the expected differences between indicator-based models in model convergence and acceptance, one could have expected that the preferred model, based on direct comparison of fit indices, would differ across indicator-based models. Only drift rate-based and accuracy-based models provided sufficient data sets to conduct model selection based on the seven measurement models of executive functions tested. This state of affair limits our understanding of the impact of the choice of indicators on model selection, since RT-based or IES-based models, as well as condition-difference indicator-based too often failed to converge and showed poor fitting to the data.

Nevertheless, the results of model selection indicate the preferred model to be the three-factor model, and the next preferred one being the two-factor model (working memory-cognitive flexibility merged). These results are in line with the bulk of the literature on executive functions reporting a three-factor structure on children of a similar age range (Agostino et al., 2010; Arán Filippetti & Richaud, 2017; Duan et al., 2010; Lehto et al., 2003; Rose et al., 2012), although for a sample of 8-to-12 years old, a two-factor structure has been also documented (Brydges et al., 2014; Huizinga et al., 2006; Lee et al., 2013; Monette et al., 2015; Scionti & Marzocchi, 2021; Usai et al., 2014; Van der Ven et al., 2013). That is, when it was directly compared against alternative models, the three-factor model was the best fitting model regardless of which indicator was used (i.e., accuracy or drift rate), except for drift rate-based models when using BIC_w; here, the two-factor model merging working memory and cognitive flexibility was preferred. The discrepancy between AIC_w and BIC_w most likely reflects how both fit indices penalize model complexity with the BIC favoring more parsimony than the AIC (Vrieze, 2012). The finding that the second most preferred model might be a more parsimonious two-factor model, with working memory merged with cognitive flexibility is in line with the view that cognitive flexibility develops later (Diamond, 2013; Karr et al., 2018). It is also well aligned with previous latent variable studies that report a non-differentiated cognitive flexibility (also termed shifting in some works) factor from working memory in pre-school and school-aged children (Monette et al., 2015; Scionti & Marzocchi, 2021; Usai et al., 2014). Indeed, our sample included children with a varied age range extending from 8 years of age all the way to 12 years of age, a period of development during which executive functions undergo rapid developmental changes (Zelazo et al., 2016). Interestingly, not only improvements in efficiency whereby children become faster and more precise have been documented during that age range (e.g., Anderson et al., 2001; Davidson et al., 2006; Huizinga et al., 2006; Lee et al., 2013; Zelazo & Carlson, 2012), but also qualitative changes in the structure and organization of executive functions with greater likelihood of two-factor models in younger samples and of the three-factor model in older samples (e.g., Brydges et al., 2014; Huizinga et al., 2006; Lee et al., 2013; Lehto et al., 2003; Rose et al., 2012; Wiebe et al., 2011).

Last but not least, to confirm the selected model fits well to the data, we looked at the average CFI and RMSEA values shown by accuracy-based and drift rate-based models, which, unlike AIC or BIC, can be interpreted based on their absolute value (Hu & Bentler, 1999). Both fit indices show not only the three-factor model on average fits well the data, but also that it shows better fit to the data than the second model that received further support based on AIC and BIC weights (i.e., two-factor model with cognitive flexibility and working memory merged). Indeed, among the seven measurement models tested, the three-factor model showed the best trade-off between convergence, acceptance, and parsimony. Importantly, this was the case whether all samples were considered (see table 3) or only the subset used for model selection (for the interested reader see table S6). It would seem good practice in future works that apply AIC/BIC model selection as we did here, to also check that the selected models fit well data by looking at indices that can be interpreted based on their absolute values, such as the CFI and RMSEA as used in the present study, or root mean squared residual (SRMR), or the goodness-of-fit index (GFI) among others (for a review see, Hu & Bentler, 1999). Finally, it would also seem important when evaluating best models to check for model convergence more systematically, as one should avoid drawing strong conclusions based on well-fitting models that often show convergence or estimation issues.

Limitations

This study offers a comprehensive empirical evaluation of the sensitivity of cognitive functions latent variable models to different methods proposed in the literature to operationalize such functions. We have provided a list of robust findings based on thousands of models and bootstrapped samples. However, these findings should also be interpreted considering the limitations of this work.

First, we could not derive the same exact indicators for each task because for some tasks, either RT (i.e., Backwards digit-span task, Letter-Memory task) or accuracy (Trail Making Test) were not recorded. Ideally, studies should aim for homogeneous models with the same type of indicators. Indeed, mixed-indicators models that use the most common indicator for each task appear not only weak on theoretical grounds, but also arbitrary in its application. For instance, on RT-based tasks (i.e., Flanker, Simon, Color-Shape switch, and Gender-Smile switch), the most common measure is unclear as both RT and accuracy have been used as relevant indicators of performance in such tasks, as discussed in the introduction. Models that search for the combination of indicators that result in the greatest rates of convergence and acceptance, would result in an explosion of combinations and not be necessarily valid. Therefore, due to the wide variety of possible combinations of indicators across tasks, the present work was systematic in the choice of indicators, by creating the most homogeneous model for each dominant indicator. It is possible that a principled combination of indicators, for instance, combining drift rate as indicator of performance on RT-based tasks (e.g., inhibition and cognitive flexibility), with accuracy-based or capacity-based measures when modeling performance in capacity tasks (e.g., working memory) may be a possible avenue for future research.

Second, the sample size of the present study was modest (n = 182). As discussed above, there is little consensus about how to estimate the required sample size for latent variable models, even if there exist some recommendations (or rules of thumb) from several studies based on the results from Monte Carlo simulations (Bentler & Chou, 1987; MacCallum & Austin, 2000; Tanaka, 1987; Wolf et al., 2013). Thus, given the sensitivity of latent variable models to sample size, model complexity, and the quality of the data, it would be welcome to see the present results confirmed in an even bigger sample. For now, we note that the sample size of this study (*n* = 182) is comparable to the one reported by most of the studies from the field (e.g., Brydges et al., 2014; A Miyake et al., 2000; Rose et al., 2012; Sluis et al., 2007; Usai et al., 2014; Van der Ven et al., 2013).

Third, the present work was not designed to address whether the different operationalizations of executive functions investigated lead to valid representations of the constructs. To the extent that we limited ourselves to well-known tasks that are accepted in the executive functions literature, this work is in line with the view that a key to validity is the choice of tasks, that is, whether the task measures what is intended to measure (Borsboom et al., 2009). To the extent that factor loadings in the preferred single-condition indicator-based models were to a greater or lesser satisfactory (e.g., three-factor model), this work is in line with the existing literature interested in identifying which operationalizations of performance in executive functions tasks, allow to reliably capture individual differences in the underlying cognitive constructs evaluated.

Fourth, the seven measurement models of executive functions tested were identified in a literature review that shows a strong influence by Miyake & Friedman's work, adopting a CFA measurement approach, that is a *reflective* approach, to examine the structure of executive functions (Friedman et al., 2008; Miyake, Friedman, et al., 2000; Shah & Miyake, 1996). Yet, other statistical methods exist to assess cognitive and psychological processes, such as *formative* models (e.g., principal component analysis), which are less affected by the low shared variance between executive functions tasks (Willoughby et al., 2014; Willoughby & Blair, 2016). Exploratory SEM (ESEM) could also be an interesting approach in future works as it overcomes some limitations of CFA, such as model misspecification and misfit (Perry et al., 2015), and the excessive flexibility of exploratory factor analysis (Marsh et al., 2014).

Traditional ESEM, nevertheless, has been challenged since it often lacks parsimony, and might cluster together constructs that are supposed to be separated in relation to theory, specially, when ESEM is applied to complex models and small samples (Marsh et al., 2014). Another interesting extension could be to complement CFA-based model selection with exploratory approaches, such as exploratory factor analysis (EFA), as Waris et al. (2017) did to investigate the structure of working memory. Finally, network modeling, which proposes that cognitive processes are conceptualized as networks of directly related manifested variables, has been proposed to overcome the limitations of *reflective* and *formative* models (Schmittmann et al., 2013). Interestingly, network modeling has proven to be an effective tool to study the differentiation process of executive functions during development (Hartung et al., 2020; Karr et al., 2022).

Finally, we recognize there also exists alternative modeling specification approaches to the seven considered here. For instance, second-order (hierarchical) models, which are more common in intelligence research (Canivez et al., 2019; Reynolds & Keith, 2017; Schneider & Newman, 2015) could be evaluated as recently done by Hartung et al. (2020) and Wolff et al. (2016) in the case of executive functions. Similarly, other bifactor structure could be considered (e.g., inhibition and cognitive flexibility tasks loading into the same specific factor and working memory tasks loading into another specific factor). In short, future works will certainly benefit from combining the strengths of different statistical techniques (e.g., CFA and networking modeling) to further investigate the underlying structure of executive functions; in doing so, the present work highlights the importance of not just the task selection but also the choice of indicators.

Conclusions

The present study confirms the sensitivity of measurement models of executive functions to the different methods proposed in the literature to operationalize executive functions. Below we summarize the highlights from this work:

- Difference scores should be avoided when modeling executive functions with latent variable methods. Measurement models that included difference scores often failed to converge and showed poor fit to the data, but in addition, they showed systematically lower factor loadings compared to measurement models that included single scores.
- RT-based models showed poor fit to the data compared to measurement models including accuracy-based measures.
- Drift rate showed the best psychometric properties for CFA models of executive functions among the four indicators tested (RTs, Accuracy, IES and drift rate). Of note, drift rate can be easily computed from individual's response accuracy, mean RT, and RT deviation through the EZ-diffusion model (Wagenmakers et al., 2007).
- This work highlights the benefit of homogenizing indicators through the use of either accuracy or drift rate to reach acceptable levels of convergence and acceptance, as well as satisfactory factor loadings, when using CFA to model executive functions.

BIBLIOGRAPHY

Agostino, A., Johnson, J., & Pascual-Leone, J. (2010). Executive functions underlying multiplicative reasoning: Problem type matters. *Journal of Experimental Child Psychology*, *105*(4), 286–305. https://doi.org/10.1016/j.jecp.2009.09.006

Alfonso, S. V., & Lonigan, C. J. (2021). Executive function, language dominance and literacy skills in Spanish-speaking language-minority children: A longitudinal study. *Early Childhood Research Quarterly*, *57*, 228–238. https://doi.org/10.1016/j.ecresq.2021.06.005

Ambrosini, E., Arbula, S., Rossato, C., Pacella, V., & Vallesi, A. (2019). Neuro-cognitive architecture of executive functions: A latent variable analysis. *Cortex*, *119*, 441–456. https://doi.org/10.1016/j.cortex.2019.07.013

Anderson, V. A., Anderson, P., Northam, E., Jacobs, R., & Catroppa, C. (2001). Development of Executive Functions Through Late Childhood and Adolescence in an Australian. *Developmental Neuropsychology*, *20*(1), 385–406. https://doi.org/10.1207/S15326942DN2001_5

Arán Filippetti, V., & Richaud, M. C. (2017). A structural equation modeling of executive functions, IQ and mathematical skills in primary students: Differential effects on number production, mental calculus and arithmetical problems. *Child Neuropsychology*, *23*(7), 864–888. https://doi.org/10.1080/09297049.2016.1199665

Baddeley, A. D., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation*, *8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

Baggetta, P., & Alexander, P. A. (2016). Conceptualization and Operationalization of Executive Function. *Mind, Brain, and Education*, *10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Bailey, C. E. (2007). Cognitive accuracy and intelligent executive function in the brain and in business. *Annals of the New York Academy of Sciences*, *1118*, 122–141. https://doi.org/10.1196/annals.1412.011

Baler, R. D., & Volkow, N. D. (2006). Drug addiction: the neurobiology of disrupted self-control. *Trends in Molecular Medicine*, *12*(12), 559–566. https://doi.org/10.1016/j.molmed.2006.10.005

Barkley, R. A. (2012). *Executive Functions: What they are, how they work, and why they evolved*. Guilford Press.

Bender, A. D., Filmer, H. L., Garner, K. G., Naughtin, C. K., & Dux, P. E. (2016). On the relationship between response selection and response inhibition: An individual differences approach. *Attention, Perception, and Psychophysics*, *78*(8), 2420–2432. https://doi.org/10.3758/s13414-016-1158-8

Bentler, P. M., & Chou, C.-P. (1987). Practical Issues in Structural Modeling. *Sociological Methods & Research*, *16*(1), 78–117.

Best, J. R., & Miller, P. H. (2010). <Best_et_al-2010-Child_Development.pdf>. *Child Development*, *81*(6), 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. https://doi.org/10.1177/2167702616657069

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The End of Construct Validity. In R. W. Lissitz (Ed.), *The Concept of Validity: Revisions, new directions, and applications* (pp. 135–170). Charlotte, NC, US: IAP Information Age Publishing.

Braver, T., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A Parametric Study of Prefrontal Cortex Involvement in Human Working Memory. *NeuroImage*, *62*(5), 49–62.

Brocki, K., & Tillman, C. (2014). Mental Set Shifting in Childhood: The Role of Working Memory and Inhibitory Control. *Infant and Child Development*, *23*, 588–604. https://doi.org/10.1002/icd.871

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361–379). The Guilford Press.

Brydges, C. R., Fox, A. M., Reid, C. L., & Anderson, M. (2014). The differentiation of executive functions in middle and late childhood : A longitudinal latent-variable analysis. *Intelligence*, *47*, 34–43. https://doi.org/10.1016/j.intell.2014.08.010

Canivez, G. L., Watkins, M. W., & McGill, R. J. (2019). Construct validity of the Wechsler Intelligence Scale For Children – Fifth UK Edition: Exploratory and confirmatory factor analyses of the 16 primary and secondary subtests. *British Journal of Educational Psychology*, *89*(2), 195–224. https://doi.org/10.1111/bjep.12230

Carlson, S. M., White, R. E., & Davis-unger, A. C. (2014). Cognitive Development Evidence for a relation between executive function and pretense representation in preschool children. *Cognitive Development*, *29*, 1–16. https://doi.org/10.1016/j.cogdev.2013.09.001

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*, *20*(3), 166–171. https://doi.org/10.1027/1015-5759.20.3.166

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology*, *23*(2), 201–216. https://doi.org/10.1016/j.acn.2007.08.010

Cirino, P. T., Ahmed, Y., Miciak, J., Taylor, W. P., Gerst, E. H., & Barnes, M. A. (2018). A framework for executive function in the late elementary years. *Neuropsychology*, *32*(2), 176–189. https://doi.org/10.1037/neu0000427

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. https://doi.org/10.1016/S0160-2896(01)00096-4

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68–80. https://doi.org/10.1037/h0029382

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Donders, F. C. (1868). Over de snelheid van psychische processen (On the speed of mental processes). In *In W. G. Koster, Attention and Performance II* (pp. 412–431).

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining Reaction Time and Accuracy: The Relationship Between Working Memory Capacity and Task Switching as a Case Example. *Perspectives on Psychological Science*, *11*(1), 133–155. https://doi.org/10.1177/1745691615596990

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508–535. https://doi.org/10.1037/bul0000192

Drollette, E. S., Shishido, T., Pontifex, M. B., & Hillman, C. H. (2012). Maintenance of cognitive control during and after walking in preadolescent children. *Medicine and Science in Sports and Exercise*, *44*(10), 2017–2024. https://doi.org/10.1249/MSS.0b013e318258bcd5

Drugowitsch, J., Deangelis, G. C., Angelaki, D. E., & Pouget, A. (2015). Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *ELife*, *4*(JUNE2015), 1–11. https://doi.org/10.7554/eLife.06678

Drugowitsch, J., Moreno-Bote, R. N., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628. https://doi.org/10.1523/JNEUROSCI.4010-11.2012

Duan, X., Wei, S., Wang, G., & Shi, J. (2010). The relationship between executive functions and intelligence on 11- to 12-year- old children. *Psychological Test and Assessment Modeling*, *52*(4), 419–431.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School Readiness and Later Achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science*, *13*(2), 190–193. https://doi.org/10.1177/1745691617720478

Engle, R. W., Laughlin, J. E., Tuholski, S. W., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331.

https://doi.org/10.1037/0096-3445.128.3.309

Eriksen, B., & Eriksen, C. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Espy, K. A. (1997). The Shape School: Assessing executive function in preschool children. *Developmental Neuropsychology*, *13*(4), 495–499. https://doi.org/10.1080/87565649709540690

Fennell, A., & Ratcliff, R. (2019). Does Response Modality Influence Conflict? Modelling Vocal and Manual Response Stroop Interference. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(11), 2098–2119. https://doi.org/10.1037/xlm0000689

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology*, *71*(6), 849–857. https://doi.org/10.1037/h0023232

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E. J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: A structural model-based approach. *Journal of Neuroscience*, *31*(47), 17242–17249. https://doi.org/10.1523/JNEUROSCI.0309-11.2011

Friedman, N. P., & Miyake, A. (2004). The Relations Among Inhibition and Interference Control Functions: A Latent-Variable Analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Friedman, N. P., Miyake, A., Robinson, J. A. L., & Hewitt, J. K. (2011). Developmental Trajectories in Toddlers' Self-Restraint Predict Individual Differences in Executive

Functions 14 Years Later: A Behavioral Genetic Analysis. *Developmental Psychology*, *47*(5), 1410–1430. https://doi.org/10.1037/a0023750

Friedman, N. P., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual Differences in Executive Functions Are Almost Entirely Genetic in Origin. *Journal of Experimental Psychology: General*, *137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Frischkorn, G. T., & Schubert, A. L. (2018). Cognitive models in intelligence research: Advantages and recommendations for their application. *Journal of Intelligence*, *6*(3), 1–22. https://doi.org/10.3390/jintelligence6030034

Frischkorn, G. T., & von Bastian, C. C. (2021). In search of the executive cognitive processes proposed by process-overlap theory. *Journal of Intelligence*, *9*(3). https://doi.org/10.3390/jintelligence9030043

Fudenberg, D., Strack, P., & Strzalecki, T. (2018). Speed, accuracy, and the optimal timing of choices. *American Economic Review*, *108*(12), 3651–3684. https://doi.org/10.1257/aer.20150742

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65–83. https://doi.org/10.1207/s15327906mbr4101_5

Gajewski, P. D., Falkenstein, M., Thönes, S., & Wascher, E. (2020). Stroop task performance across the lifespan: High cognitive reserve in older age is associated with enhanced proactive and reactive interference control. *NeuroImage*, *207*(October 2019). https://doi.org/10.1016/j.neuroimage.2019.116430

Gardiner, E., & Iarocci, G. (2018). Everyday executive function predicts adaptive and internalizing behavior among children with and without autism spectrum disorder. *Autism Research*, *11*(2), 284–295. https://doi.org/10.1002/aur.1877

Gartner, A., & Strobel, A. (2021). Individual differences in inhibitory control: A latent variable analysis. *Journal of Cognition*, *4*(1), 1–18. https://doi.org/10.5334/joc.150

Ging-Jehli, N. R., & Ratcliff, R. (2020). Effects of aging in a task-switch paradigm with the diffusion decision model. *Psychology and Aging*, *35*(6), 850–865. https://doi.org/10.1037/pag0000562

Gomez, P., Ratcliff, R., & Perea, M. (2007). A Model of the Go / No-Go Task. *Journal of Experimental Psychology: General*, *136*(3), 389–413. https://doi.org/10.1037/0096-3445.136.3.389

Green, S., & Yang, Y. (2018). Empirical Underidentification with the Bifactor Model: A Case Study. *Educational and Psychological Measurement*, *78*(5), 717–736. https://doi.org/10.1177/0013164417719947

Griffin, D., Murray, S., & Gonzalez, R. (1999). Difference score correlations in relationship research: A conceptual primer. *Personal Relationships*, *6*(4), 505–518. https://doi.org/10.1111/j.1475-6811.1999.tb00206.x

Hancock, G. R., & Mueller, R. (2008). Best Practices in Structural Equation Modeling. *Best Practices in Quantitative Methods*, 488–510.

Hartung, J., Engelhardt, L. E., Thibodeaux, M. L., Harden, K. P., & Tucker-Drob, E. M. (2020). Developmental transformations in the structure of executive functions. *Journal of Experimental Child Psychology*, *189*, 104681. https://doi.org/10.1016/j.jecp.2019.104681

Hedge, C., Powell, G., & Sumner, P. (2018). *The reliability paradox : Why robust cognitive tasks do not produce reliable individual differences*. 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*(8 JUN), 1–19.

https://doi.org/10.3389/fnins.2014.00150

Hertzog, C., Vernon, M. C., & Rypma, B. (1993). Age differences in mental rotation task performance: The influence of speed/accuracy tradeoffs. *Journals of Gerontology*, *48*(3), 150–156. https://doi.org/10.1093/geronj/48.3.P150

Hooper, D., Coughlan, J., & Mullen, M. (2008). Evaluating model fit: A synthesis of the structural equation modeling literature. *Paper Presented at the 7th European Conference on Research Methodology for Business and Management Studies,*.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis : Conventional criteria versus new alternatives Cutoff Criteria for Fit Indexes in Covariance Structure Analysis : Conventional Criteria Versus New Alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hughes, C. (2011). Changes and Challenges in 20 Years of Research Into the Development of Executive Functions. *Infant and Child Development*, *20*, 251–271. https://doi.org/10.1002/icd.736

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods*, *46*(3), 702–721. https://doi.org/10.3758/s13428-013-0411-5

Huizinga, M., Dolan, C. V., & Van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, *44*(11), 2017–2036. https://doi.org/10.1016/j.neuropsychologia.2006.01.010

Ionescu, T. (2012). Exploring the nature of cognitive flexibility. *New Ideas in Psychology*, *30*(2), 190–200. https://doi.org/10.1016/j.newideapsych.2011.11.001

Jacob, R., & Parkinson, J. (2015). The Potential for School-Based Interventions That Target

Executive Function to Improve Academic Achievement: A Review. *Review of Educational Research*, *85*(4), 512–552. https://doi.org/10.3102/0034654314561338

Kamijo, K., Khan, N. A., Pontifex, M. B., Scudder, M. R., Drollette, E. S., Raine, L. B., Evans, E. M., Castelli, D. M., & Hillman, C. H. (2012). The relation of adiposity to cognitive control and scholastic achievement in preadolescent children. *Obesity*, *20*(12), 2406–2411. https://doi.org/10.1038/oby.2012.112

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review*, *9*(4), 637–671. https://doi.org/10.3758/BF03196323

Kane, M. J., Tuholski, S. W., Hambrick, D. Z., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*, *144*(11), 1147–1185. https://doi.org/10.1037/bul0000160

Karr, J. E., Rodriguez, J. E., Goh, P. K., Martel, M. M., & Rast, P. (2022). The Unity and Diversity of Executive Functions: A Network Approach to Life Span Development. *Developmental Psychology*, *58*(4), 751–767. https://doi.org/10.1037/dev0001313

Kenny, D. A. (2015). *Measuring model fit*. http://www.davidakenny.net/cm/fit.htm

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (Fourth Edi). Guilford Press.

Kyriazos, T. A. (2018). Applied Psychometrics: Sample Size and Sample Power

Considerations in Factor Analysis (EFA, CFA) and SEM in General. *Psychology*, *09*(08), 2207–2230. https://doi.org/10.4236/psych.2018.98126

Lambek, R., & Shevlin, M. (2011). Working memory and response inhibition in children and adolescents: Age and organization issues. *Scandinavian Journal of Psychology*, *52*(5), 427–432. https://doi.org/10.1111/j.1467-9450.2011.00899.x

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental Changes in Executive Functioning. *Child Development*, *84*(6), 1933–1953. https://doi.org/10.1111/cdev.12096

Lehto, J., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning : evidence from children. *British Journal of Developmental Psychology*, *21*, 59–80.

Lerche, V., von Krause, M., Voss, A., Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2020). Diffusion modeling and intelligence: Drift rates show both domain-general and domain-specific relations with intelligence. *Journal of Experimental Psychology: General*, *149*(12), 2207–2249.

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, *81*(3), 629–652. https://doi.org/10.1007/s00426-016-0770-5

Lerner, M. D., & Lonigan, C. J. (2014). *Executive Function Among Preschool Children : Unitary Versus Distinct Abilities*. 626–639. https://doi.org/10.1007/s10862-014-9424-3

Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the Weber noise of relative positions. *Journal of Vision*, *15*(4), 1–14. https://doi.org/10.1167/15.4.10

Liesefeld, H. R., Fu, X., & Zimmer, H. D. (2015). Fast and careless or careful and slow? Apparent holistic processing in mental rotation is explained by speed-accuracy trade-offs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1140–1151.

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, *51*(1), 40–60. https://doi.org/10.3758/s13428-018-1076-x

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

MacLeod, C. M. (2007). The concept of inhibition in cognition. In D. S. Gorfein & C. M. MacLeod (Eds.), *Inhibition in cognition* (pp. 3–23). https://doi.org/10.1037/11587-001

Maris, G., & van der Maas, H. (2012). Speed-Accuracy Response Models: Scoring Rules based on Response Time and Accuracy. *Psychometrika*, *77*(4), 615–633. https://doi.org/10.1007/s11336-012-9288-y

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*(2), 181–220. https://doi.org/10.1207/s15327906mbr3302_1

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*(Mimic), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Masten, A. S., Herbers, J. E., Desjardins, C. D., Cutuli, J. J., Christopher, M., Sapienza, J. K., Long, J. D., & Zelazo, P. D. (2012). Executive Function Skills and School Success in Young Children Experiencing Homelessness. *Educational Researcher*, *41*(9), 375–384. https://doi.org/10.3102/0013189X12459883

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology*, *24*(2), 222–243. https://doi.org/10.1037/a0017619

McIntosh, J. R., & Mehring, C. (2017). Modifying response times in the Simon task with transcranial random noise stimulation. *Scientific Reports*, *7*(1), 1–16. https://doi.org/10.1038/s41598-017-15604-1

McIntosh, J. R., & Sajda, P. (2020). Decomposing Simon task BOLD activation using a drift-diffusion model framework. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-60943-1

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin and Review*, *20*(5), 819–858. https://doi.org/10.3758/s13423-013-0404-5

Miller, M. R., Giesbrecht, G. F., Müller, U., Mcinerney, R. J., & Kerns, K. A. (2012). A Latent Variable Approach to Determining the Structure of Executive Function in Preschool Children A Latent Variable Approach to Determining the Structure of Executive Function in Preschool Children. *Journal of Cognition and Development*, *13*(3). https://doi.org/10.1080/15248372.2011.585478

Miyake, A., Emerson, M. J., & Friedman, N. P. (2000). Assessment of Executive Functions in Clinical Settings: Problems and Recommendations. *Seminars in Speech and Language*, *21*(2), 169–183. https://doi.org/10.1016/B978-0-12-803676-1.00022-2

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal Lobe” tasks: a latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Monette, S., Bigras, M., & Lafrenière, M. (2015). Structure of executive functions in typically developing kindergarteners. *Journal of Experimental Child Psychology*, *140*, 120–139. https://doi.org/10.1016/j.jecp.2015.07.005

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Montroy, J. J., Merz, E. C., Williams, J. M., Landry, S. H., Johnson, U. Y., Zucker, T. A., Assel, M., Taylor, H. B., Lonigan, C. J., Phillips, B. M., Clancy-Menchetti, J., Barnes, M. A., Eisenberg, N., Spinrad, T., Valiente, C., de Villiers, J., & de Villiers, P. (2019). Hot and cool dimensionality of executive function: Model invariance across age and maternal education in preschool children. *Early Childhood Research Quarterly*, *49*, 188–201. https://doi.org/10.1016/j.ecresq.2019.06.011

Morrison, F. J., Ponitz, C. C., & McClelland, M. M. (2010). Self-regulation and academic achievement in the transition to school. *Child Development at the Intersection of Emotion and Cognition.*, 203–224. https://doi.org/10.1037/12059-011

Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental Science*, *10*(6), 719–726. https://doi.org/10.1111/j.1467-7687.2007.00623.x

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407–422. https://doi.org/10.1016/j.intell.2013.06.004

Nicolaou, A. I., & Masoner, M. M. (2013). Sample size requirements in structural equation models under standard conditions. *International Journal of Accounting Information Systems*, *14*(4), 256–274. https://doi.org/10.1016/j.accinf.2013.11.001

Norman, D. A., & Shallice, T. (1986). Attention to Action. *Consciousness and Self-Regulation*, 1–18. https://doi.org/10.1007/978-1-4757-0629-1_1

Ong, G., Sewell, D. K., Weekes, B., McKague, M., & Abutalebi, J. (2017). A diffusion model approach to analysing the bilingual advantage for the Flanker task: The role of attentional control processes. *Journal of Neurolinguistics*, *43*, 28–38. https://doi.org/10.1016/j.jneuroling.2016.08.002

Ørskov, P. T., Norup, A., Beatty, E. L., & Jaeggi, S. M. (2021). Exploring Individual Differences as Predictors of Performance Change During Dual-N-Back Training. *Journal of Cognitive Enhancement*, *5*(4), 499–501. https://doi.org/10.1007/s41465-021-00221-8

Paap, K. R., & Sawi, O. (2016). The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods*, *274*, 81–93. https://doi.org/10.1016/j.jneumeth.2016.10.002

Packwood, S., Hodgetts, H. M., & Tremblay, S. (2011). A multiperspective approach to the conceptualization of executive functions. *Journal of Clinical and Experimental Neuropsychology*, *33*(4), 456–470. https://doi.org/10.1080/13803395.2010.533157

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376–404. https://doi.org/10.1167/5.5.1

Penadés, R., Catalán, R., Rubia, K., Andrés, S., Salamero, M., & Gastó, C. (2007). Impaired response inhibition in obsessive compulsive disorder. *European Psychiatry*, *22*(6), 404–410. https://doi.org/10.1016/j.eurpsy.2006.05.001

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, *19*(1), 12–21. https://doi.org/10.1080/1091367X.2014.952370

Pontifex, M. B., Saliba, B. J., Raine, L. B., Picchietti, D. L., & Hillman, C. H. (2013). Exercise improves behavioral, neurocognitive, and scholastic performance in children with attention-deficit/hyperactivity disorder. *Journal of Pediatrics*, *162*(3), 543–551. https://doi.org/10.1016/j.jpeds.2012.08.036

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the Optimal Number of Factors in Exploratory Factor Analysis: A Model Selection Perspective. *Multivariate Behavioral Research*, *48*(1), 28–56. https://doi.org/10.1080/00273171.2012.710386

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review*, *9*(2), 278–291. https://doi.org/10.3758/BF03196283

Ratcliff, R., Huang-Pollock, C., & McKoon, G. (2018). Modeling individual differences in the go/no-go task with a diffusion model. *Decision*, *5*(1), 42–62. https://doi.org/10.1037/dec0000065

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., McKoon, G., & Gomez, P. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review*, *111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions.

Psychological Science, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127–157. https://doi.org/10.1016/j.cogpsych.2009.09.001

Reitan, R. M. (1971). Trail Making Test results for normal and brain-damaged children. *Perceptual and Motor Skills*, *33*, 575–581.

Rey-Mermet, A., Gade, M., Souza, A. S., von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General*, *148*(8), 1335–1372. https://doi.org/10.1037/xge0000593

Rey-Mermet, A., Singmann, H., & Oberauer, K. (2021). Neither measurement error nor speed-accuracy trade-offs explain the difficulty of establishing attentional control as a psychometric construct: Evidence from a latent-variable analysis using diffusion modeling. *Preprint at PsyArxiv*.

Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fifth Edition: What does it measure? *Intelligence*, *62*, 31–47. https://doi.org/10.1016/j.intell.2017.02.005

Ritchie, S. J., Quinlan, E. B., Banaschewski, T., Bokde, A. L. W., Flor, H., Frouin, V., Garavan, H., Gowland, P., Ittermann, B., Martinot, J., Martinot, M. P., Orfanos, D. P., Paus, T., Poustka, L., Hohmann, S., Millenet, S., Fröhner, J. H., Smolka, M. N., Walter, H., … Consortium, I. (2019). Neuroimaging and genetic correlates of cognitive ability and cognitive development in adolescence. *Preprint at PsyArxiv*.

https://doi.org/%0A10.31234/osf.io/8pwd6

Roelofs, A. (2018). Acta Psychologica One hundred fi fty years after Donders : Insights from unpublished data , a replication , and modeling of his reaction times ☆. *Acta Psychologica*, *191*(October), 228–233. https://doi.org/10.1016/j.actpsy.2018.10.002

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207–231. https://doi.org/10.1037//0096-3445.124.2.207

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2012). Implications of Infant Cognition for Executive Functions at Age 11. *Psychological Science*, *21*(11), 1345–1355. https://doi.org/10.1177/0956797612444902

Rosseel, Y. (2012). lavaan : an R package for structural equation modeling and more Version 0 . 5-12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin and Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin and Review*, *18*(2), 324–330. https://doi.org/10.3758/s13423-011-0055-3

Salthouse, T. A., Fristoe, N., McGuthry, K. E., & Hambrick, D. Z. (1998). Relation of task switching to speed, age, and fluid intelligence. *Psychology and Aging*, *13*(3), 445–461. https://doi.org/10.1037/0882-7974.13.3.445

Schmidt, M., Egger, F., Benzing, V., Jäger, K., Conzelmann, A., Roebers, C. M., & Pesce, C. (2017). Disentangling the relationship between children ' s motor ability , executive function and academic achievement. *PLoS ONE*, *12*(8). https://doi.org/e0182845

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007).

Individual Differences in Components of Reaction Time Distributions and Their Relations to Working Memory and Intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414–429. https://doi.org/10.1037/0096-3445.136.3.414

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Schmitz, F., & Voss, A. (2012). Decomposing task-switching costs with the diffusion model. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 222–250. https://doi.org/10.1037/a0026003

Schmitz, F., & Wilhelm, O. (2016). Modeling mental speed: Decomposing response time distributions in elementary cognitive tasks and correlations with working memory capacity and fluid intelligence. *Journal of Intelligence*, *4*(4), 1–23. https://doi.org/10.3390/jintelligence4040013

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, *25*(1), 12–27. https://doi.org/10.1016/j.hrmr.2014.09.004

Scionti, N., & Marzocchi, G. M. (2021). The dimensionality of early executive functions in young preschoolers: Comparing unidimensional versus bidimensional models and their ecological validity. *Child Neuropsychology*, *27*(4), 491–515. https://doi.org/10.1080/09297049.2020.1868419

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Servant, M., & Evans, N. J. (2020). A diffusion model analysis of the effects of aging in the flanker task. *Psychology and Aging*, *35*(6), 831–849. https://doi.org/10.1037/pag0000546

Servant, M., Montagnini, A., & Burle, B. (2014). Conflict tasks and the diffusion framework: Insight in model constraints based on psychological laws. *Cognitive Psychology*, *72*, 162–195. https://doi.org/10.1016/j.cogpsych.2014.03.002

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Shah, P., & Miyake, A. (1996). The Separability of Working Memory Resources for Spatial Thinking and Language Processing : An Individual Differences Approach. *Journal of Experimental Psychology: General*, *125*(1), 4–27.

Shing, Y. L., Lindenberger, U., Diamond, A., Li, S. C., & Davidson, M. C. (2010). Memory maintenance and inhibitory control differentiate from early childhood to adolescence. *Developmental Neuropsychology*, *35*(6), 679–697. https://doi.org/10.1080/87565641.2010.508546

Siegrist, M. (1997). Test-retest reliability of different versions of the stroop test. *Journal of Psychology: Interdisciplinary and Applied*, *131*(3), 299–306. https://doi.org/10.1080/00223989709603516

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. https://doi.org/10.1037/h0020586

Sluis, S. Van Der, Jong, P. F. De, & Leij, A. Van Der. (2007). Executive functioning in children , and its relations with reasoning , reading , and arithmetic. *Intelligence*, *35*, 427–449. https://doi.org/10.1016/j.intell.2006.09.001

Smith, E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*(1), 5–42. https://doi.org/10.1006/cogp.1997.0658

Smith, G. A., & Brewer, N. (1995). Slowness and Age: Speed-Accuracy Mechanisms. *Psychology and Aging*, *10*(2), 238–247. https://doi.org/10.1037/0882-7974.10.2.238

Spencer, M., Richmond, M. C., & Cutting, L. E. (2020). Considering the Role of Executive Function in Reading Comprehension: A Structural Equation Modeling Approach. *Scientific Studies of Reading*, *24*(3), 179–199. https://doi.org/10.1080/10888438.2019.1643868

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology (2006)*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Starns, J. J., & Ratcliff, R. (2012). Age-related differences in diffusion model boundary optimality with both trial-limited and time-limited tasks. *Psychonomic Bulletin and Review*, *19*(1), 139–145. https://doi.org/10.3758/s13423-011-0189-3

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260. https://doi.org/10.1007/BF02289729

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, *22*(9), 1503–1511. https://doi.org/10.1038/s41593-019-0453-9

Tamnes, C. K., Østby, Y., Walhovd, K. B., Westlye, L. T., Due-Tønnessen, P., & Fjell, A. M. (2010). Neuroanatomical correlates of executive functions in children and adolescents: A magnetic resonance imaging (MRI) study of cortical thickness. *Neuropsychologia*, *48*(9), 2496–2508. https://doi.org/10.1016/j.neuropsychologia.2010.04.024

Tanaka, J. S. (1987). How Big Is Big Enough ?: Sample Size and Goodness of Fit in

Structural Equation Models with Latent Variables. *Child Development*, *58*(1), 134–146.

Taylor Tavares, J. V., Clark, L., Cannon, D. M., Erickson, K., Drevets, W. C., & Sahakian, B. J. (2007). Distinct Profiles of Neurocognitive Function in Unmedicated Unipolar Depression and Bipolar II Depression. *Biological Psychiatry*, *62*(8), 917–924. https://doi.org/10.1016/j.biopsych.2007.05.034

Thurm, F., Zink, N., & Li, S. C. (2018). Comparing effects of reward anticipation on working memory in younger and older adults. *Frontiers in Psychology*, *9*(NOV), 1–16. https://doi.org/10.3389/fpsyg.2018.02318

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In N. J. Castellan & F. Restle (Eds.), *Cognitive theory* (3rd ed., pp. 199–239). Lawrence Erlbaum Associates. https://doi.org/https://doi.org/10.4324/9781315802473

Ulrich, R. (1999). *Donders ' s assumption of pure insertion : an evaluation on the basis of response dynamics*. *102*, 43–75.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychology Science Quarterly*, *51*(4), 388–402. http://p16277.typo3server.info/fileadmin/download/PschologyScience/4-2009/psq_4_2009_388-402.pdf

Usai, M. C., Viterbori, P., Traverso, L., Franchis, V. De, Usai, M. C., Viterbori, P., Traverso, L., & De, V. (2014). Latent structure of executive function in five- and six-year-old children : A longitudinal study. In *European Journal of Developmental Psychology* (Vol. 11, Issue 4, pp. 447–462). Taylor & Francis. https://doi.org/10.1080/17405629.2013.840578

Van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2013). The structure of executive functions in children: A closer examination of inhibition, shifting, and updating. *British Journal of Developmental Psychology*, *31*(1), 70–87. https://doi.org/10.1111/j.2044-835X.2012.02079.x

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5

Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *Journal of Cognition*, *1*(1), 1–16. https://doi.org/10.5334/joc.6

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wagenmakers, E. J., van der Maas, H., & Grasman, R. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, *43*(3), 476–496. https://doi.org/10.1080/00273170802285941

Waris, O., Soveri, A., Ahti, M., Hoffing, R. C., Ventus, D., Jaeggi, S. M., Seitz, A. R., & Laine, M. (2017). *A Latent Factor Analysis of Working Memory Measures Using Large-Scale Data*. *8*(June), 1–14. https://doi.org/10.3389/fpsyg.2017.01062

Wechsler, D. (1981). The psychometric tradition: Developing the wechsler adult intelligence scale. *Contemporary Educational Psychology*, *6*(2), 82–85. https://doi.org/10.1016/0361-476X(81)90035-7

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (T. P. Corporation (ed.); Third

Edit). The Psychological Corporation.

Weeda, W. D., Van der Molen, M. W., Barceló, F., & Huizinga, M. (2014). A diffusion model analysis of developmental changes in children's task switching. *Journal of Experimental Child Psychology*, *126*, 178–197. https://doi.org/10.1016/j.jecp.2014.05.001

White, C. N., Ratcliff, R., & Starns, J. J. (2011). Diffusion models of the flanker task: Discrete versus gradual attentional selection. *Cognitive Psychology*, *63*(4), 210–238. https://doi.org/10.1016/j.cogpsych.2011.08.001

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using Confirmatory Factor Analysis to Understand Executive Control in Preschool Children: I. Latent Structure. *Developmental Psychology*, *44*(2), 575–587. https://doi.org/10.1037/0012-1649.44.2.575

Wiebe, S. A., Sheffield, T., Mize, J., Clark, C. A. C., Chevalier, N., & Andrews, K. (2011). Journal of Experimental Child The structure of executive function in 3-year-olds. *Journal of Experimental Child Psychology*, *108*(3), 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*(JUL), 1–22. https://doi.org/10.3389/fpsyg.2013.00433

Willoughby, M. T., & Blair, C. B. (2016). Measuring executive function in early childhood: A case for formative measurement. *Psychological Assessment*, *28*(3), 319–330. https://doi.org/10.1037/pas0000152

Willoughby, M. T., Blair, C., Wirth, R. J., & Greenbereg, M. (2012). The Measurement of Executive Function at Age 5: Psychometric Properties and Relationship to Academic Achievement. *Psychological Assessment*, *24*(1), 226–239. https://doi.org/10.1037/a0025361

Willoughby, M. T., Holochwost, S. J., Blanton, Z. E., & Blair, C. B. (2014). Executive Functions: Formative Versus Reflective Measurement. *Measurement*, *12*(3), 69–95. https://doi.org/10.1080/15366367.2014.929453

Willoughby, M. T., Kuhn, L. J., Blair, C. B., Samek, A., & List, J. A. (2017). The test–retest reliability of the latent construct of executive function depends on whether tasks are represented as formative or reflective indicators. *Child Neuropsychology*, *23*(7), 822–837. https://doi.org/10.1080/09297049.2016.1205009

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *73*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wolff, M., Krönke, K. M., Venz, J., Kräplin, A., Bühringer, G., Smolka, M. N., & Goschke, T. (2016). Action versus state orientation moderates the impact of executive functioning on real-life self-control. *Journal of Experimental Psychology: General*, *145*(12), 1635–1653. https://doi.org/10.1037/xge0000229

Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory and Cognition*, *34*(3), 668–684. https://doi.org/10.3758/BF03193587

Wright, L., Lipszyc, J., Dupuis, A., Thayapararajah, S. W., & Schachar, R. (2014). Response inhibition and psychopathology: A meta-analysis of Go/No-Go task performance. *Journal of Abnormal Psychology*, *123*(2), 429–439. https://doi.org/10.1037/a0036295

Wundt, W. (1880). *Grundz• uge der physiologischen Psychologie (Foundations of physiological psychology)* (2nd ed.).

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods.

Behavior Research Methods, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Xu, F., Han, Y., Sabbagh, M. A., Wang, T., Ren, X., & Li, C. (2013). Developmental differences in the structure of executive function in middle childhood and adolescence. *PloS One*, *8*(10). https://doi.org/10.1371/journal.pone.0077770

Yang, Y., Bender, A. R., & Raz, N. (2015). Neuropsychologia Age related differences in reaction time components and diffusion properties of normal-appearing white matter in healthy adults. *Neuropsychologia*, *66*, 246–258. https://doi.org/10.1016/j.neuropsychologia.2014.11.020

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392. https://doi.org/10.1177/0734282911406668

Yangüez, M., Bediou, B., Chanal, J., & Bavelier, D. (2022). *In search of better practice in executive functions assessment: methodological issues and potential solutions*. Retrieved from osf.io/yvcj7

Yangüez, M., Bediou, B., Hillman, C. H., Bavelier, D., & Chanal, J. (2021). The Indirect Role of Executive Functions on the Relationship between Cardiorespiratory Fitness and School Grades. In *Medicine & Science in Sports & Exercise: Vol. Publish Ah*. https://doi.org/10.1249/mss.0000000000002630

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. https://doi.org/10.1037/a0024177

Yaple, Z., & Arsalidou, M. (2018). N-back Working Memory Task: Meta-analysis of Normative fMRI Studies With Children. *Child Development*, *89*(6), 2010–2022.

https://doi.org/10.1111/cdev.13080

Zelazo, P. D., Blair, C. B., & Willoughby, M. T. (2016). *Executive Function: Implications for Education (NCER 2017-2000)* (pp. 1–148). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/

Zelazo, P. D., & Carlson, S. M. (2012). Hot and Cool Executive Function in Childhood and Adolescence: Development and Plasticity. *Child Development Perspectives*, *6*(4), 354–360. https://doi.org/10.1111/j.1750-8606.2012.00246.x

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860