



Preprint

2023

Public access

This version of the publication is provided by the author(s) and made available in accordance with the copyright holder(s).

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

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

How to cite

YANGUEZ ESCALERA, Marc et al. In search of better practice in executive functions assessment: methodological issues and potential solutions. 2023, p. 85. doi: 10.1037/rev0000434

This publication URL: <https://archive-ouverte.unige.ch/unige:170310>

Publication DOI: [10.1037/rev0000434](https://doi.org/10.1037/rev0000434)

© The author(s). This work is licensed under a Creative Commons Attribution (CC BY 4.0)

<https://creativecommons.org/licenses/by/4.0>

Last deposit update in Archive ouverte UNIGE on 11.04.2024 17:42

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

3

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

6

7 Marc Yangüez^{1,2,3a}, Benoit Bediou^{1,3,a}, Julien Chanal^{1,2,b} and Daphne Bavelier^{1,3,b}

8 ¹ Faculty of Psychology, University of Geneva, Switzerland

9 ² Distance Learning University, Brig, Switzerland

10 ³ Campus Biotech, University of Geneva, Switzerland

11 ^aShared first authorship, ^bshared last authorship

12

13 **Author Note**

14 Corresponding author: Daphne Bavelier (daphne.bavelier@unige.ch), [https://orcid.org/0000-](https://orcid.org/0000-0002-5904-1240)
15 [0002-5904-1240](https://orcid.org/0002-5904-1240); Campus Biotech, Chemin des Mines 9, CH-1202 Geneva

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

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

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

19

20 We have no conflict of interest to disclose.

21 Data and R-code are available in the OSF website (<https://osf.io/yvcj7/>).

22 This study was not preregistered.

23 Support for this project was provided by the Foundation Ernest Boninchi (CH) and by the
24 Jacobs Foundation (CH).

25

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

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

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

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

37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57

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.

58 Executive functions are considered an «umbrella term» for a number of cognitive
59 processes relying on frontal lobe functioning (Barkley, 2012), which are fundamental for
60 school readiness (e.g., Blair & Razza, 2007; Morrison, Ponitz, & McClelland, 2010), scholastic
61 performance (e.g., Duncan et al., 2007; Jacob & Parkinson, 2015; St Clair-Thompson &
62 Gathercole, 2006), job success (e.g., Bailey, 2007), or mental health (e.g., Baler & Volkow,
63 2006; Gardiner & Iarocci, 2018; Penadés et al., 2007; Taylor Tavares et al., 2007). The
64 importance of executive functions for such variety of life aspects explains, in part, the growing
65 interest in psychological and neuroscience research around these functions in the last decades.
66 Unfortunately, the proliferation of studies about executive functions has been coupled with
67 important methodological differences in their conceptualization and measurement, which
68 hinders our understanding of the psychological and theoretical mechanisms of executive
69 functions (Baggetta & Alexander, 2016; Barkley, 2012; Karr et al., 2018; McCabe et al., 2010;
70 Packwood et al., 2011). Furthermore, recently several publications have expressed their
71 concern about the poor psychometric properties of many executive functions measures, a
72 critical point for individual differences research (Draheim et al., 2019; Hedge et al., 2018; Paap
73 & Sawi, 2016; Rouder & Haaf, 2019).

74 To understand what executive functions are, what components form them, and how
75 they organize, first, it is essential to determine how to best operationalize ability in executive
76 functions tasks to better capture the latent processes under assessment. Traditionally, the field
77 has relied on the use of single tasks to assess executive functions components (Baggetta &
78 Alexander, 2016; Chan et al., 2008); however, this approach fails to recognize that no task is
79 process pure, as each task necessarily involves processes other than the intended one (Conway
80 et al., 2005; Miyake, Friedman, et al., 2000; Shah & Miyake, 1996), a phenomenon commonly
81 known in the literature as *task impurity*. Accordingly, it has been widely documented through
82 psychometric analyses that the use of a single task to characterize one or multiple executive

83 functions components suffers from both validity and reliability issues (e.g., Kane et al., 2004;
84 Shah & Miyake, 1996; Yang & Green, 2011). Executive functions assessment remains
85 challenging due to the complexity of the constructs they encompass. This is in part because
86 commonalities between the components that form it (and the tasks used to measure them) do
87 exist, and yet some diversity across components is also noted (Friedman et al., 2008; Frischkorn
88 & von Bastian, 2021; Karr et al., 2018; Miyake, Friedman, et al., 2000). To address the
89 commonalities and diversity between executive functions components, researchers have
90 exploited the use of multiple tasks coupled with confirmatory factor analysis (CFA), a special
91 form of structural equation modeling (SEM) technique, which enables to define and estimate
92 measurement models to analyze the relationship between manifested variables (or indicators)
93 and the latent variables that form the models (MacCallum & Austin, 2000). CFA is a powerful
94 tool for psychometric evaluation and construct validation (Brown & Moore, 2012). Thus, this
95 approach has been very useful to mitigate both the *task impurity* problem, enabling a better
96 evaluation of the cognitive components when the tasks used are not process pure, and the
97 *measurement error* problem, removing the unique variance from each task (Engle et al., 1999;
98 Kane et al., 2004; Miyake, Emerson, et al., 2000; Miyake, Friedman, et al., 2000; Shah &
99 Miyake, 1996). The present study builds on this approach to investigate how different methods
100 proposed in the literature to operationalize executive functions impact the assessment of the
101 underlying structure of executive functions when using CFA.

102 **Executive functions latent variable studies**

103 Prior to the seminal article by Miyake et al (2000), which introduced the use of CFA to
104 assess executive functions components and its multidimensional structure, the earliest models
105 already viewed executive functions as a higher-order global construct that managed lower-level
106 cognitive processes (Baddeley & Hitch, 1974; Norman & Shallice, 1986). Although these
107 models did not include the term executive functions explicitly, they form the foundation for

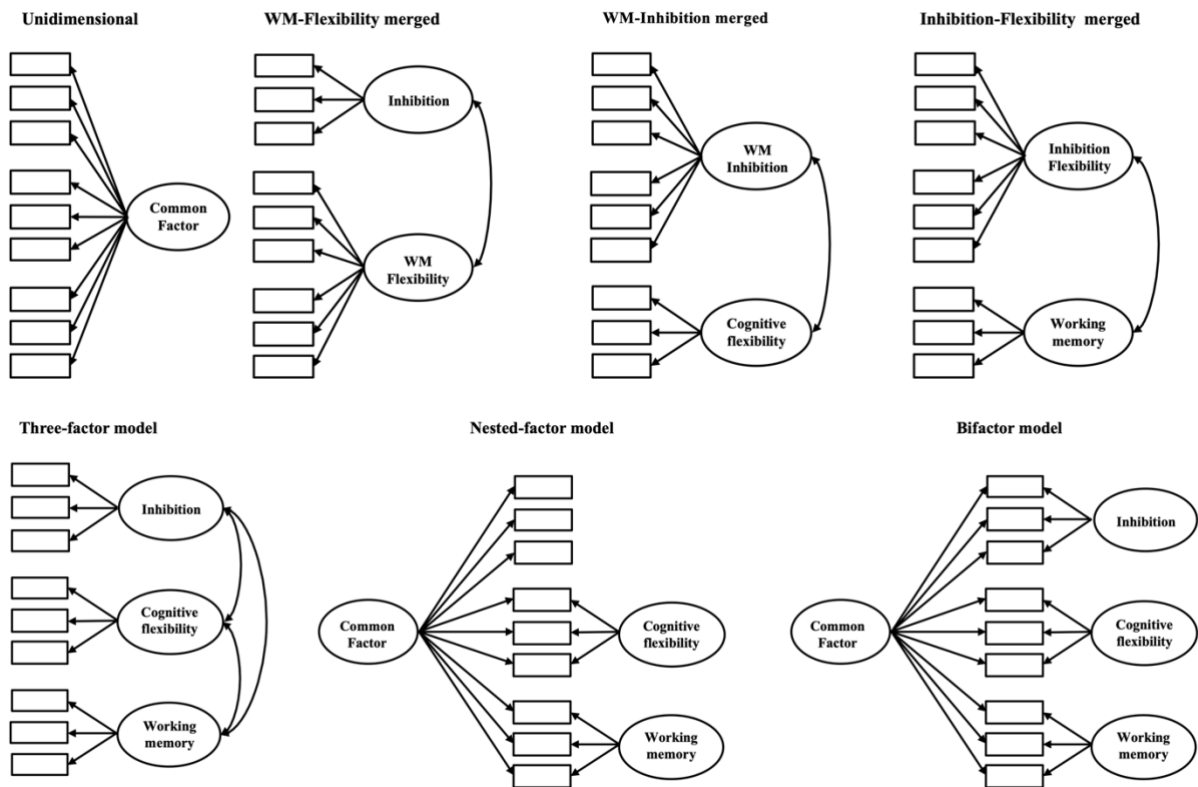
108 subsequent models of executive functions, which have described the construct as
109 multidimensional, including higher-order and lower-order cognitive processes, which are
110 moderately to strongly correlated (Diamond, 2013; Engle, 2018; Friedman & Miyake, 2017;
111 Miyake, Friedman, et al., 2000). The existence of different models of executive functions
112 illustrates, in part, the complexity of defining what are executive functions or in other words,
113 the challenge of characterizing which combination of cognitive processes they encompass.
114 There is general agreement, however, that executive functions is a multidimensional construct,
115 which involves core components such as (i) inhibition, (ii) cognitive flexibility and (iii)
116 working memory, which are fundamental for higher-order processes, such as planning,
117 reasoning or goal-directed behavior (Baggetta & Alexander, 2016; Diamond, 2013; Hughes,
118 2011). Inhibition is globally defined as the ability to suppress the processing of irrelevant
119 stimuli or the outcome of impulsive reactions (MacLeod, 2007); cognitive flexibility, also
120 termed shifting, refers to the capacity to swiftly change focus whether in terms of task goals or
121 attention distribution (Best & Miller, 2010; Ionescu, 2012); finally, working memory refers to
122 the ability to keep information active in mind and mentally manipulate it (Diamond, 2013;
123 Kane et al., 2004). Note that working memory is a multidimensional construct, with
124 components that can be distinguished based on their emphasis on (i) content material (e.g.,
125 verbal and visuospatial) or (ii) constituent processes (e.g., updating and maintenance) (Smith
126 & Jonides, 1997; Waris et al., 2017). These three components form the core of most studies
127 investigating executive functions (Karr et al., 2018; Packwood et al., 2011). Note that in part
128 of the literature, and in the present work, the terms shifting and cognitive flexibility are used
129 interchangeably, as well as the terms updating and working memory (Baggetta & Alexander,
130 2016; Diamond, 2013).

131 CFA has thus been helpful in mitigating the *task impurity* problem, as it has highlighted
132 time and time again that the predictive validity and reliability of a latent construct is greater

133 than that of the single measures from which it is derived (Conway et al., 2005; Kane et al.,
134 2004; Willoughby et al., 2017). Furthermore, SEM techniques are a powerful statistical tool
135 for individual differences research in the fields of working memory (Engle et al., 1999; Kane
136 & Engle, 2002; Ørskov et al., 2021; Rey-Mermet et al., 2019; Shah & Miyake, 1996) and
137 executive functions (Arán Filippetti & Richaud, 2017; Friedman et al., 2008; Schmidt et al.,
138 2017; Sluis et al., 2007; Spencer et al., 2020). It comes thus as no surprise that in the last decade
139 many studies have used SEM techniques to investigate executive functions development, their
140 structure, as well as their neural organization (Alfonso & Lonigan, 2021; Brydges et al., 2014;
141 Cirino et al., 2018; Huizinga et al., 2006; Lambek & Shevlin, 2011; Lee et al., 2013; Lerner &
142 Lonigan, 2014; Monette et al., 2015; Montroy et al., 2019; Ritchie et al., 2019; Rose et al.,
143 2012; Usai et al., 2014; Wiebe et al., 2011; Willoughby et al., 2012; Xu et al., 2013). For
144 example, latent variable studies suggest that executive functions develop and differentiate from
145 a rather unitary structure to a multidimensional structure throughout childhood and
146 adolescence. Below 8 years of age, most studies document either a unitary structure (Brydges
147 et al., 2014; Shing et al., 2010; Wiebe et al., 2008, 2011; Willoughby et al., 2012) or a two
148 factor structure (Lerner & Lonigan, 2014; M. R. Miller et al., 2012; Monette et al., 2015; Usai
149 et al., 2014). In that age range, the three basic processes of executive functions seem to be
150 initially undifferentiated, and then inhibition is often reported as emerging first. Studies in
151 middle childhood and adolescence show a gradual differentiation of executive functions to the
152 three-factor structure most often described in young adults (e.g., Brydges et al., 2014; Lehto et
153 al., 2003; Rose et al., 2012; Shing et al., 2010).

154 Recently, in an extensive literature review of studies that applied CFA to assess the
155 structure of executive functions, Karr et al. (2018) pointed out several weaknesses when
156 applying CFA to executive functions research. They performed a literature search resulting in
157 40 articles, 17 of which provided sufficient data for re-analysis through bootstrapping methods.

158 This literature search also identified seven different measurement models of executive
 159 functions being the most commonly found in the literature. As displayed in Figure 1, these are:
 160 one unidimensional model merging all three factors of inhibition, cognitive flexibility, and
 161 working memory; three two-factor models (merging the three factors above two by two, so
 162 inhibition with working memory, inhibition with cognitive flexibility, and working memory
 163 with cognitive flexibility); one three-factor model (inhibition, cognitive flexibility, and
 164 working memory); one nested factor model (a common factor, plus two specific orthogonal
 165 factors of cognitive flexibility, and working memory); and one bifactor model (a common
 166 factor, plus three specific orthogonal factors of inhibition, cognitive flexibility, and working
 167 memory).
 168



169
 170 **Figure 1.** Measurement models of executive functions (adapted from Karr et al., 2018).
 171

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

186 **Heterogeneity of tasks' conditions in executive functions studies: single versus difference** 187 **of conditions**

188 Executive functions tasks have been traditionally designed by contrasting two task
189 conditions with the aim of disentangling both executive and non-executive processes involved
190 during task performance. This approach has its origin in Donders' subtraction method
191 (Donders, 1868), and has been hugely successful in the early days of experimental psychology.
192 Accordingly, some of the most used executive functions tasks' paradigms are built following
193 this approach since it gives robust experimental effects (Eriksen & Eriksen, 1974; Rogers &
194 Monsell, 1995; Simon & Rudell, 1967; Stroop, 1935). Yet, the subtraction method has also
195 been under criticism as the additivity of processing time is largely unsubstantiated (Gomez et
196 al., 2007; Ulrich, 1999; Wundt, 1880). Given our interest in the use of latent variable models

197 for executive functions research, we focus our discussion of the heterogeneity of tasks'
198 conditions across studies to those that have used a CFA approach to assess executive functions;
199 yet this same issue applies throughout the large executive function literature. Table S1 (online
200 supplementary material) provides a list of studies that applied CFA to investigate the structure
201 of executive functions in pre-school and school-aged children, and information about the tasks,
202 tasks' conditions and indicators used to measure executive functions on each study.

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

210 Cognitive flexibility is typically measured through task-switching paradigms (Monsell,
211 2003). This sort of tasks frequently includes two types of blocks: a homogeneous condition
212 (blocks of trials requiring a response only to a feature of the stimuli, for instance, the color),
213 and a heterogeneous-mixed condition (mixed rule-set of cues to flexibly shift attention towards
214 the correct target feature - for instance, arms down indicates respond to the color; arms up
215 respond to the shape). Performance on task-switch paradigms is commonly operationalized
216 with three different methods: (i) global switch cost (e.g., Miyake, Friedman, et al., 2000), which
217 is the difference in performance between the heterogeneous-mixed condition (i.e., blocks
218 including switch and non-switch trials) and the homogeneous condition (blocks including only
219 non-switch trials); (ii) local switch cost (e.g., Ambrosini et al., 2019; Friedman & Miyake,
220 2004), which is the difference between switch and non-switch trials in the heterogeneous-
221 mixed condition; and (iii) performance on switch trials from the heterogeneous-mixed

222 condition (e.g., Huizinga et al., 2006; Lee et al., 2013). Again, depending on the task
223 condition(s) selected to operationalize cognitive flexibility, either a single task condition or a
224 difference between conditions may be considered.

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

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

245 **Heterogeneity of indicators in executive functions studies**

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

265 As illustrated in Table S1, accuracy, RT, and capacity measures (e.g., maximum
266 number of items correctly recalled) are most often used indicators in the literature, at least for
267 those works using CFA as tabulated here. For working memory, performance is standardly
268 operationalized in terms of span capacity or accuracy (Wilhelm et al., 2013), creating less
269 variability in the indicators used. For inhibition and cognitive flexibility, performance is more
270 standardly assessed via RT-based tasks, leading to the possibility of using speed, accuracy, or

271 a combination thereof as indicators. We turn below to the psychometric issues raised by such
272 varied operationalizations of RT-based tasks, with a special focus on the tasks' conditions or
273 the measures used to derive an indicator.

274 **Psychometric issues associated with RT-based measures**

275 RT and RT differences are two of the most popular indicators used to study the speed
276 and efficiency of mental processes in psychology and neuroscience research (Draheim et al.,
277 2019). Despite their widespread use in experimental and individual differences research,
278 several studies have shown that both indicators suffer from reliability and validity issues.

279 It has been argued that the subtraction method increases the error variance, since it
280 removes part of the common variance between the two mental processes from which the RT
281 difference score is calculated (Hedge et al., 2018). As an illustration of this issue, Paap and
282 Sawi (2016) assessed the test-retest reliability of (i) single RT (e.g., mean RT in congruent or
283 in incongruent trials from inhibition tasks; mean RT in switching or in non-switching trials in
284 cognitive flexibility tasks) and (ii) RT difference between task conditions on four classical
285 executive functions tasks (i.e., Antisaccade, Flanker, Simon & Color-shape switching) in a
286 sample of undergraduate students ($N = 81$). Their results indicate that single RT is a more
287 reliable behavioral indicator (.71-.89, test-retest reliability range) than RT difference scores
288 (.43-.62). Importantly, such results are in line with those reported in other studies (e.g., Hughes,
289 Linck, Bowles, Koeth, & Bunting, 2014; Salthouse, Fristoe, McGuthry, & Hambrick, 1998;
290 Siegrist, 1997).

291 RT measures are sensitive to speed-accuracy trade-off, whereby participants as they
292 are told to react faster will show greater error rates, and vice-versa (Fitts, 1966; Ratcliff &
293 Rouder, 1998; Stone, 1960). Despite the effort to instruct participants to give a similar weight
294 to both dimensions, participants tend to adopt different response strategies (Starns & Ratcliff,
295 2012). Importantly, the literature has shown consistently that age-related differences exist in

296 speed-accuracy trade-off strategies. For example, when comparing older adults with younger
297 adults, the first tend to put more weight on accuracy over speed in order to ensure a higher
298 accuracy, whereas the latter tend to take more risk speeding up their responses at the cost of
299 making more errors (e.g., Forstmann et al., 2011; Hertzog, Vernon, & Rypma, 1993; Smith &
300 Brewer, 1995; Starns & Ratcliff, 2012). Given the complex interaction between speed and
301 accuracy (see Heitz, 2014, for a review), the point has been made that analyses based solely on
302 RT cannot fully account for individual differences in cognition; this is particularly the case of
303 studies with heterogeneous samples, such as on developmental or aging studies, which
304 inevitably will include participants with different response strategies (Draheim et al., 2016;
305 Hertzog et al., 1993; Hughes et al., 2014; Ratcliff et al., 2016; Yang et al., 2015).

306 Several efforts have been made to address this speed-accuracy trade-off issue, through
307 the development of indicators such as the inverse efficiency score (IES: Townsend & Ashby,
308 1978), the linear-integrated speed-accuracy score (LISAS: Vandierendonck, 2017, 2018), the
309 rate-correct score (RCS: Woltz & Was, 2006) or the balanced integration score (BIS: Liesefeld,
310 Fu, & Zimmer, 2015). These measures provide several benefits over traditional RT- or
311 accuracy-based measures, such as (i) mitigating speed-accuracy tradeoffs, or (ii) containing
312 more information about individuals' ability than RT and accuracy separately. However, there
313 is debate about the weight that such measures give to speed over accuracy (or vice-versa) to
314 generate a reliable integrated measure of speed and accuracy (Draheim et al., 2019; Liesefeld
315 & Janczyk, 2019). Such composite measures have rarely been used in the context of executive
316 functions latent variable research, although there are some exceptions, such as Gärtner &
317 Strobel (2021) and Yangüez et al., (2021), who used the IES (or RT divided by response
318 accuracy) to operationalize performance on different RT-based executive functions tasks. In
319 the next section, we review another approach to explain patterns of RTs and choices, the Drift
320 Diffusion Model (DDM).

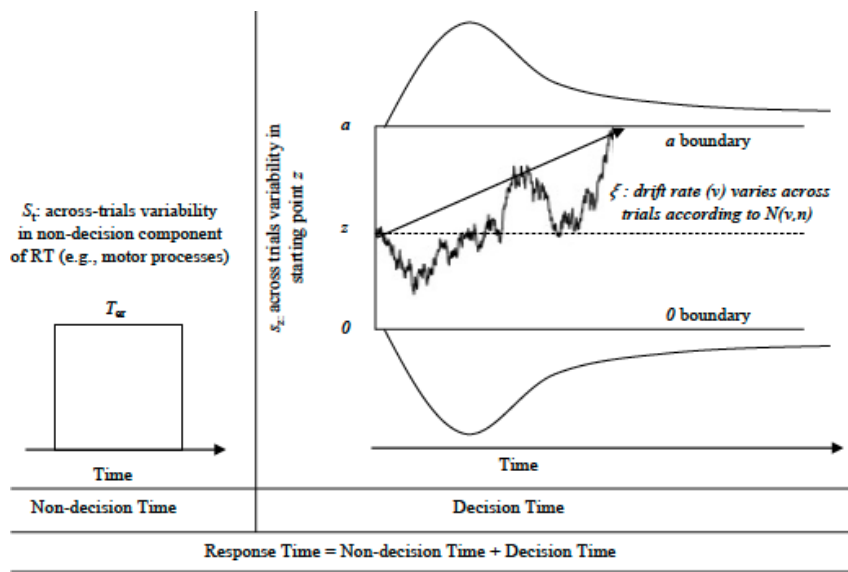
321 **Applications of the DDM to executive functions research**

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

330 The initial diffusion model was developed for two-choice RT paradigms, although it
331 can be generalized to paradigms that include more than two choices (Ratcliff et al., 2016;
332 Tajima et al., 2019). When the DDM is applied to a two-choice RT paradigm (represented in
333 Figure 2), stimulus presentation triggers the decision process, and in particular information
334 accumulation until one of the two decision boundaries is reached (0 or a , for the two-choice
335 model in Figure 2). The drift rate (v) represents the average rate of evidence, with a larger drift
336 rate meaning a faster accumulation of evidence, and vice-versa. The model assumes that drift
337 rate varies across trials following a Normal or Gaussian distribution according to $\xi \sim N(v, n)$,
338 since the accumulation process is subject to moment-to-moment, Brownian variability. The
339 distance between the two decision boundaries will affect an individual's speed-accuracy trade-
340 off. Larger values of the boundary parameter a represent a conservative response strategy, as a
341 larger a means more information needs to be accumulated before a decision can be made
342 (resulting in longer RTs and higher accuracy). The z parameter can be added into the model to
343 examine whether an individual has an a priori bias towards one of the two response options,
344 before the stimulus is experienced. The DDM also enables one to estimate the non-decision
345 time or the time to execute a motor response, T_{er} . For a full description of the standard diffusion

346 model, the reader is referred to Ratcliff et al. (2004). Here we focus on drift rate as it
 347 characterizes the efficiency of information processing and is thus most promising to
 348 characterize core executive functions processes, in contrast to decision boundary, which is
 349 related to strategic factors, and to non-decision time, which captures additive processes such
 350 as preparation and motor execution (Ratcliff, 1978; Ratcliff et al., 2016).

351



352

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

355

356 Diffusion models have been applied to some of the most well-known executive
 357 functions tasks, such as the Flanker (Ong et al., 2017; Servant & Evans, 2020; White et al.,
 358 2011), Simon (McIntosh & Mehring, 2017; McIntosh & Sajda, 2020; Servant et al., 2014)
 359 Go/No-Go (Gomez et al., 2007; Ratcliff et al., 2018), Stroop (Fennell & Ratcliff, 2019;
 360 Gajewski et al., 2020), task-switch (Ging-Jehli & Ratcliff, 2020; Schmitz & Voss, 2012; Weeda
 361 et al., 2014), and n-back tasks (Thurm et al., 2018). Some of these studies have also successfully
 362 applied DDM, and in particular drift rate, to investigate individual differences in executive
 363 functions (Gajewski et al., 2020, Ging-Jehli et al., 2020, Ong et al., 2017, Servant & Evans,

364 2020, Ratcliff et al., 2018, Weeda et al., 2014). Yet, the use of drift rate as an indicator for
365 modeling executive functions with SEM techniques, such as CFA, remains largely untested
366 (for an exception, see Rey-Mermet et al., 2021).

367 In the present work, we take advantage of the EZ-diffusion model (Wagenmakers et al.,
368 2007) to systematically assess how indicator choices, including drift rate, may affect CFA
369 measurement models of executive functions, in terms of convergence and acceptance. The EZ-
370 diffusion model only estimates the three most relevant parameters of the DDM to characterize
371 the decision-making process: (i) the drift rate v , (ii) the decision boundary a and (iii) the non-
372 decision time T_{er} . The EZ-diffusion model is a solution when tasks do not have enough trials
373 to estimate all the parameters from the DDM simplifying remarkably the standard fitting
374 procedure (Palmer et al., 2005; Ratcliff, 1978, 2002). This is often the case of executive
375 functions CFA studies, which use multiple tasks to measure each of the constructs included in
376 their models, and therefore, the tasks often are not designed to fit more parameters, as is
377 required by the standard DDM or more recent DDM versions with collapsing bound
378 (Drugowitsch et al., 2012; Fudenberg et al., 2018).

379 **Aims of the present study**

380 The present study aims to test the impact of different operationalizations of executive
381 functions in terms of the tasks' conditions and indicators used, on the modeling of executive
382 functions with CFA. This investigation focuses not only on the comparison of single versus
383 difference scores in task's conditions, but also on the choice of indicators between more
384 standard measures, such as RT and accuracy, and less common ones, such as the IES and drift
385 rate. Indeed, our final aim is to determine the impact of such different methodological practices
386 on executive functions modeling, especially the metrics related to model fitting, such as
387 convergence and acceptance (i.e., how well a model fits to the data). Latent variable models
388 are likely sensitive to such operationalization (MacCallum & Austin, 2000), as it is expected

389 that some indicators will show greater common variance than others. Thus, beyond the choice
390 of tasks and conditions, how a researcher operationalizes performance in that very task will
391 have important implications for the modeling of the construct evaluated. In particular, these
392 choices may lead to different results when modeling the structure of executive functions (e.g.,
393 model convergence and acceptance, level of factor-loadings, number of latent constructs of the
394 model, inter-factor correlations, etc.). In line with Karr et al. (2018), the rate of model
395 convergence and acceptance (or how often, when a model converges, it meets a fitting
396 threshold) will be examined as tasks' conditions and indicators vary. In addition, whether the
397 likelihood of model selection varies depending on the task conditions and indicators used will
398 also be assessed, to the extent the model converges. In this way, the present work establishes
399 not only which operationalizations of executive functions should be preferred, but also which
400 measurement models of executive functions (e.g., one-factor, two-factor, three-factor, etc.) are
401 most likely representative of executive functions structure during middle childhood.

402

403

Methods

404

Participants and Procedure

405

406

407

408

409

410

411

412

413

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.

414 **Measures**

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

422 *Description of executive functions tasks*

423 **Inhibition tasks.** Flanker task - Modified (Eriksen & Eriksen, 1974; Pontifex et al.,
424 2013). In this task, an array of five fishes is presented with the central fish pointing either in
425 the same direction (congruent trials) or in the opposite direction (incongruent trials) as the other
426 fish. The task is to determine the direction the middle, target fish, is facing. Incongruent trials
427 in this task require the greatest inhibitory control demands due to perceptual interference.

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

434 Go/No-Go task (Kamijo et al., 2012). In the first part of this task (sustained attention
435 condition) children must press the response button to rare-target stimuli (picture of a lion, 0.2
436 probability), and to withhold their response to frequent non-target stimuli (picture of a tiger,
437 0.8 probability). Then children perform the second part of the task (impulsivity condition), in
438 which they must press a button to frequent-target stimuli (tiger, 0.8 probability), and to

439 withhold from responding to rare-non-target stimuli (lion, 0.2 probability). That task order is
440 fixed to induce greater conflict and thus, need for inhibition during the “impulsivity condition”.

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

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

462 Trail Making Test (Reitan, 1971). In Trail A, children are asked to draw lines
463 connecting numbers by numerical order (numbers from 1-25 are distributed randomly across

464 the test-sheet). In Trail B, the test-sheet contains numbers and letters, and children have to
465 connect numbers and letters by alternating the sequence (i.e., 1-A-2-B-3-C, etc), requiring
466 continuously switching between two different task sets. Trail-B is the most demanding
467 experimental condition in terms of cognitive flexibility.

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

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

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

484 *Dependent measures derived from each task*

485 Single-condition indicators represent performance on those trials (or task’s conditions)
486 that require greater amounts of executive control (e.g., incongruent trials, switch trials, 2-back
487 trials), whereas condition-difference indicators represent the performance difference between
488 that single task condition with the greater executive control demands and a baseline task

489 condition with low executive control demands (e.g., incongruent minus congruent trials in the
490 Flanker task; switch minus non-switch trials in the Color-Shape switch task; and 2-back minus
491 0-back in the Spatial n-back task). Table 1 summarizes the task conditions and indicators used
492 for each of the nine tasks. For a description of how the four indicators were computed, see
493 *statistical procedures*.

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

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

511

512

513

514

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		*		

515 *Note.* EF: executive functions. RT: response time; Acc: accuracy; IES: inverse efficiency score; DR: drift rate;
 516 SA: sustained attention; N/A: not applicable; ¹Computer reaction time-based task, RT and response accuracy are
 517 recorded; ²Paper-pencil, time to completion task; ³Computer accuracy-based task, only response accuracy is
 518 recorded. ⁴Single task condition with greatest EF demands. ⁵Switch – Non-switch trials mixed-block.
 519

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

529 **Working memory tasks.** The Letter-Memory task is an accuracy-based measured,
 530 where response accuracy is collapsed across all trials (single-condition indicator). On the
 531 Backwards digit-span task, the longest sequence that was remembered correctly (e.g., 5 digits)
 532 was used as a measure of working memory span (single-condition indicator). As these two
 533 accuracy-based tasks are not built to contrast performance between different task conditions, a

534 difference score could not be derived. On the Spatial n-back task, RT and accuracy were
535 recorded, and in addition, IES and drift rate were computed. Single-condition indicators were
536 derived from the 2-back condition (i.e., RT, accuracy, IES, and drift rate); difference score
537 indicators were derived by considering performance in 2-back minus 0-back condition.

538 **Statistical procedures**

539 ***Raw data cleaning***

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

543 ***Raw data transformation to indicators***

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

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

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

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

555 Then, univariate analyses on each indicator were conducted to remove outlier data
556 before fitting the models to the data. Values $\pm 3 SD$ from the sample mean were excluded from
557 the analyses, this affected less than 1.5% of observations (Table S2 reports the % of missing

558 data for each task after removing outlier values). No other data cleaning procedure was
559 conducted.

560 *Structural equation models*

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

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

578

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. ¹Computer reaction time-based task, RT and response accuracy are recorded; ²Paper-pencil, time to completion task;

³Computer 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

626 ***Bootstrap resampling***

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

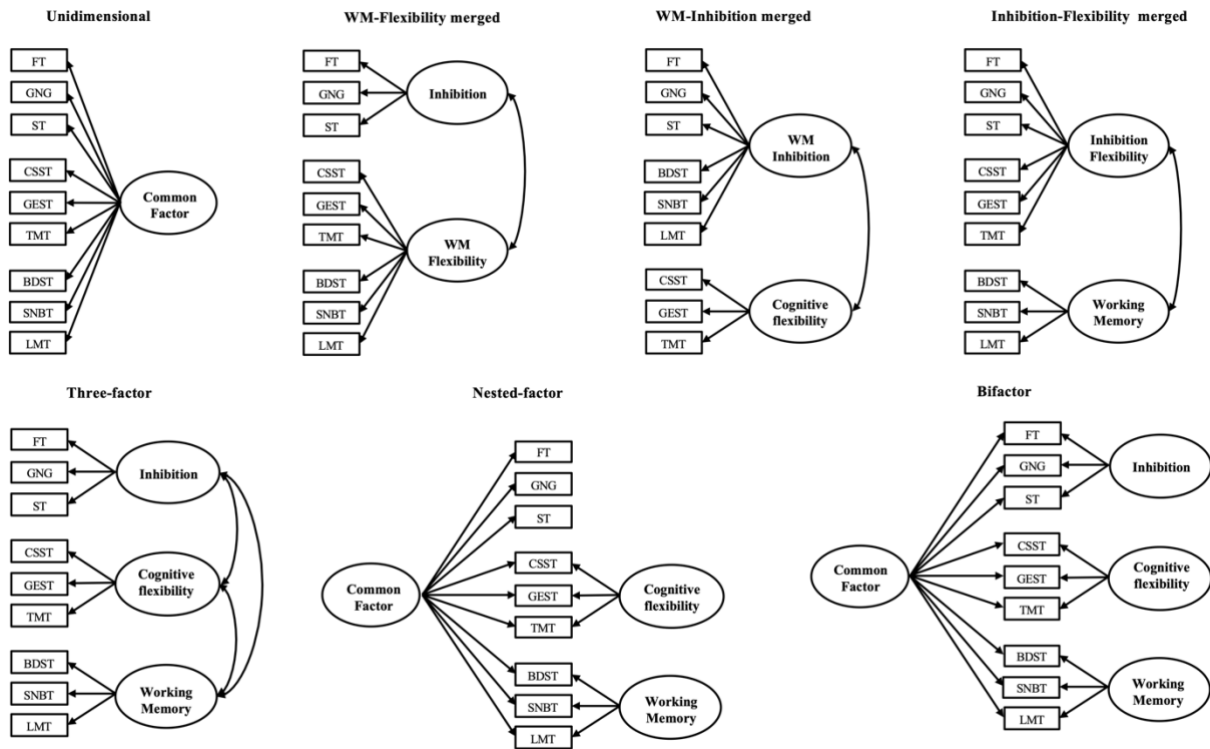
640 ***Bootstrap analysis – Model convergence, acceptance, and selection***

641 The present study analyzed the simulated data in two different ways, following the
642 procedure conducted in Karr et al (2018). First, we looked at the rate of model convergence, as
643 well as the rate of model acceptance. The rate of model convergence represents the percent of
644 models that converged across the 5'000 data sets, regardless of the fit indices. The rate of model
645 acceptance analyzes the percent of models meeting the fitting thresholds (i.e., lenient and
646 strict), among the models that converged. This first step enables to estimate the frequency that
647 any of the models tested in the present study would (i) converge without any errors or (ii) meet
648 the fitting thresholds proposed. Therefore, this first analysis enabled to determine the most
649 suitable indicator-based model(s), in terms of rate of model convergence and model
650 acceptance, to model executive functions with latent variable methods. Importantly, we aimed

651 not only to identify the indicator(s) with the best psychometric properties to model executive
 652 functions, but also, we examined potential differences between single-condition indicator-
 653 based models and condition-difference indicator-based models.

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

658
 659



660

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

665

666 **Model fit interpretation**

667 **Model acceptance.** To determine the rate of model acceptance, goodness of fit to the
 668 data of the models tested was evaluated using the comparative fit index (CFI) and root mean

669 square error of approximation (RMSEA). These two fit indices have a common metric and
 670 provide complementary information, as the CFI is an incremental fit index that compares an
 671 hypothesized model with a baseline model (i.e., a model in which no items covary) in terms
 672 of goodness of fit, whereas the RMSEA is an absolute fit index that assesses how far an
 673 hypothesized model is from a perfect model (Xia & Yang, 2019). Furthermore, the RMSEA
 674 favors parsimony since it penalizes model complexity, unlike the CFI (Hooper et al., 2008).

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

680 **Model selection.** To determine the probability of model selection, we assessed the fit
 681 of the models with two different indices, Akaike's information criterion (AIC; Akaike, 1973),
 682 and the Bayesian information criterion (BIC; Schwartz, 1978). Both indices are measures of
 683 comparative fit, which are meaningful only when used to compare different models (Kenny,
 684 2015). Models with lower values indicate a better fit to the data. Both indices balance goodness-
 685 of-fit and complexity. Lack of parsimony is penalized according to the number of parameters
 686 of the model. The AIC index applies a linear penalty of two for every parameter estimated,
 687 whereas the BIC applies a bigger penalty to model complexity, since the BIC increases the
 688 penalty exponentially as model complexity increases (Vrieze, 2012).

689 The model selection analysis was conducted based on the estimation of the relative AIC
 690 weight (AIC_w) and BIC weight (BIC_w) of each model, a method proposed by Wagenmakers
 691 and Farrell (2004) for model selection. First, for each data set in which the seven measurement
 692 models of executive functions converged without warning/errors, we computed the difference
 693 in fit, ΔAIC and ΔBIC , between the best fitting model (indicated by the lowest AIC and BIC

694 value) and the other models. Therefore, the best fitting model always has a ΔAIC (or ΔBIC) =
695 0, and the other models a ΔAIC (or ΔBIC) > 0 . Then, we computed the AIC_w and BIC_w for
696 each model using the equations provided in Wagenmakers & Farrell (2004). This method
697 enables to estimate the probability that model M_i is the best model given data and the candidates
698 models tested. Note that the relative model probabilities are normalized by dividing by the sum
699 of the probabilities of all the models.

700 ***Data Availability Statement***

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

705

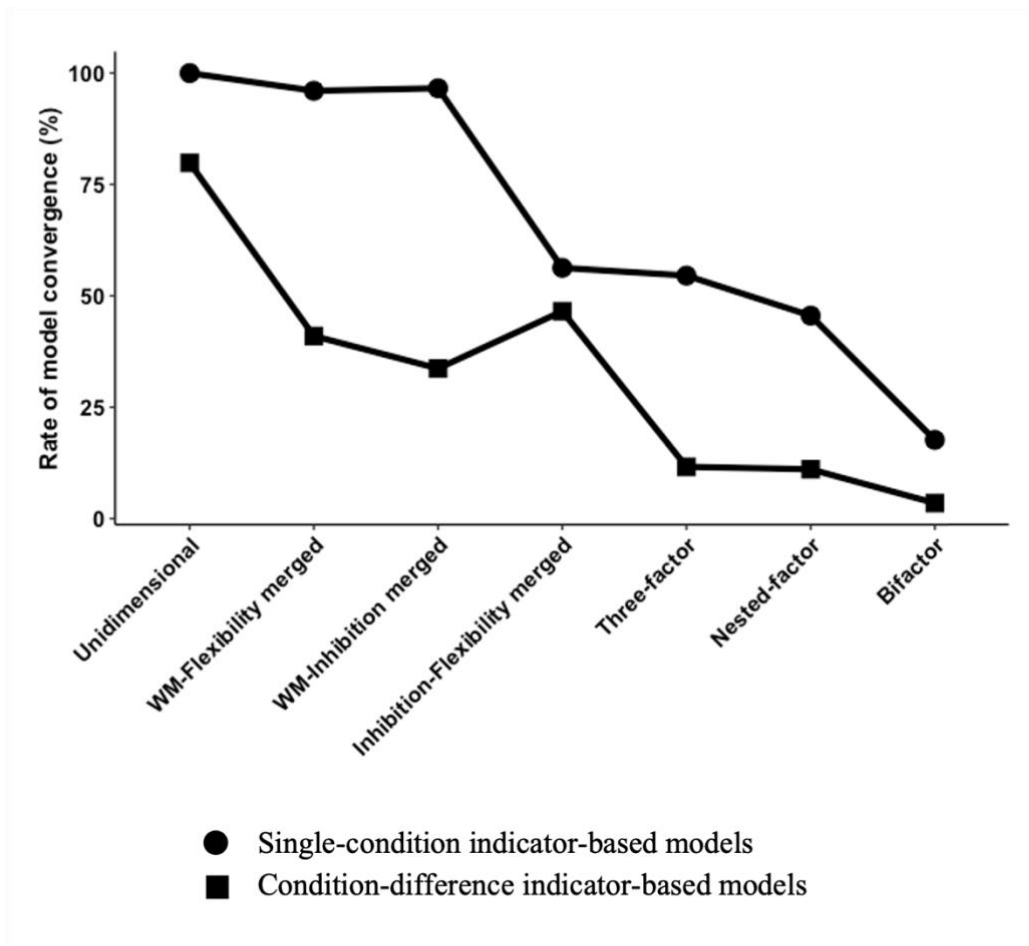
706 **Results**

707 **Model convergence**

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

709 First, we examined potential differences in model convergence, as a function of the
710 tasks' conditions that can be used to operationalize performance in executive functions tasks.
711 Thus, indicator-based models were grouped as (i) single-condition indicator-based models
712 versus (ii) condition-difference indicator-based models. Collapsed across all seven models,
713 single-condition indicator-based models showed a remarkably higher rate of convergence (\bar{x} =
714 66.63%, min = 2.2%, max = 100%), compared to condition-difference indicator-based models
715 (\bar{x} = 31.56%, min = 0.98%, max = 94.92%).

716 When we looked more in detail at differences between both groups across measurement
 717 models of executive functions, we observed that single-condition indicator-based models
 718 showed systematically higher rates of convergence compared to condition-difference indicator-
 719 based models (Figure 4). As a case example, the three-factor models showed a much higher
 720 rate of convergence on single-condition indicator-based models ($\bar{x} = 51.58\%$, min = 3.4%,
 721 max = 96.26%) compared to condition-difference indicator-based models ($\bar{x} = 11.01\%$, min =
 722 2.82%, max = 19.24%). The same pattern was observed across the seven measurement models
 723 of executive functions (Figure 4).



724 **Figure 4.** Average rate of convergence across single-condition vs. condition-difference
 725 indicator-based models, as a function of measurement models of executive functions.
 726

727 Given this striking difference, the remainder of our analyses will focus exclusively on
 728 single-condition indicator-based models, which showed the best psychometric properties when

729 it comes to convergence, that is, they showed a much lower percentage of improper solutions.
730 Full results of model convergence and acceptance of condition-difference indicator-based
731 models are provided in Table S3 (online supplementary material) for the interested reader.
732 Importantly, Table S3 confirms that the poor performance of condition-difference indicator-
733 based models does not hide a tradeoff between convergence and acceptance. Single-condition
734 indicator-based models perform better based on their enhanced rates of acceptance when both
735 CFI and RMSEA fit indices are taken into account. Furthermore, for the interested reader,
736 Table S4 (online supplementary material) reports the percentage of measurement models that
737 (i) converged (without warning), (ii) did not converge, and (iii) that converged with a warning
738 message (e.g., negative variance; variance-covariance matrix not positive definite, etc.)

739 *Impact of the indicator used*

740 Table 3 lists the mean percent of models that converged across the 5'000 data sets, as a
741 function of the indicator-based model tested for each of the seven measurement models of
742 executive functions models. Within each measurement model, there were remarkable
743 differences in model convergence depending on the indicator used. While unidimensional
744 models all converged regardless of the indicator used, two factor models also showed high rate
745 of convergence overall, but less so when RT was used as an indicator. As model complexity
746 increased convergence rates not only decreased as expected, but surprisingly this effect was
747 much more marked for RT-based and IES-based models than for accuracy-based and drift rate-
748 based models. This information is illustrated in Figure 5 (thick black line). Furthermore,
749 averaging across measurement models of executive functions, we observed remarkable
750 differences between single-condition indicator-based models in model convergence. On
751 average, accuracy-based models ($\bar{x} = 85.54\%$, min = 28.88%, max = 100%) showed the
752 greatest rate of model convergence followed closely by drift rate-based models ($\bar{x} =$
753 81.93%, min = 24.12%, max = 100%), whereas IES-based models ($\bar{x} = 53.36\%$, min = 12.44%,

754 max = 100%), and RT-based models (\bar{x} = 45.70%, min = 2.2%, max = 99.98%) showed much
 755 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 $\geq .90$ (%)	CFI $\geq .95$ (%)	RMSEA $\leq .08$ (%)	RMSEA $\leq .05$ (%)
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

756 **Model acceptance**

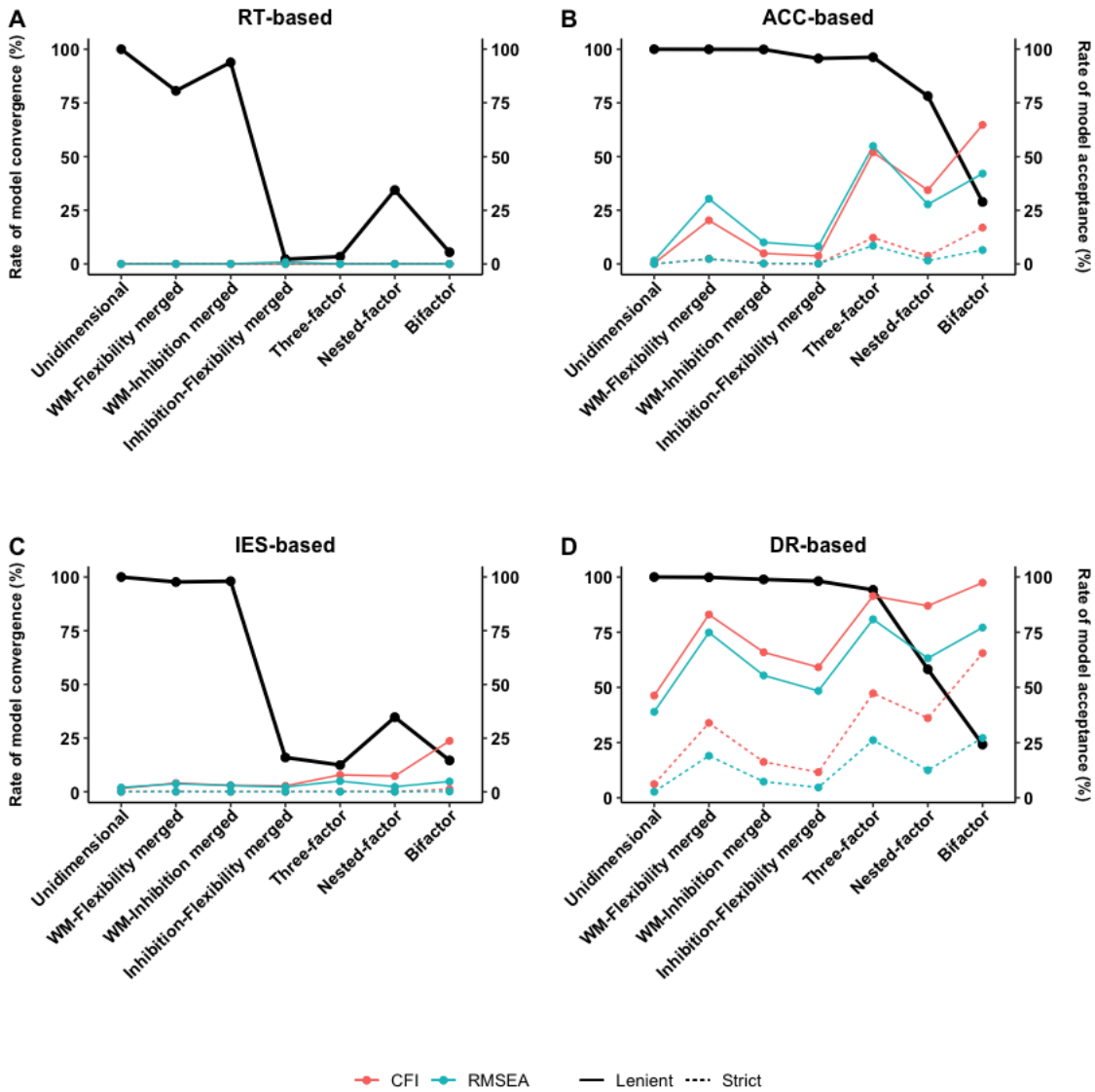
757 The second step of the analysis aimed to examine potential differences in model
 758 acceptance between single-condition indicator-based models. Thus, we looked at the percent
 759 of models meeting either lenient ($CFI \geq .90$ or $RMSEA \leq .08$) or strict fit thresholds ($CFI \geq .95$
 760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

767 *Impact of the indicator used*

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

772 In sum, averaging across the seven measurement models of executive functions, drift
 773 rate-based models (Figure 5D) showed the highest rate of model acceptance based on both
 774 lenient (CFI: $\bar{x} = 75.74\%$; RMSEA: $\bar{x} = 62.69\%$) and strict thresholds (CFI: $\bar{x} = 30.98\%$;
 775 RMSEA: $\bar{x} = 14.20\%$). These rates are remarkably higher than those of any other indicator-
 776 based model. More precisely, RT-based models (Figure 5A) showed extremely low rates of
 777 acceptance, for both lenient (CFI: $\bar{x} = 0.01\%$; RMSEA: $\bar{x} = 0.14\%$) and strict thresholds (CFI:
 778 $\bar{x} = 0\%$; RMSEA: $\bar{x} = 0\%$). The same pattern was observed on IES-based models (Figure 5C),
 779 for both lenient (CFI: $\bar{x} = 7.18\%$; RMSEA: $\bar{x} = 3.29\%$) and strict thresholds (CFI: $\bar{x} = 0.22\%$;
 780 RMSEA: $\bar{x} = 0.02\%$). Furthermore, although accuracy-based models (Figure 5B) showed a

781 slightly higher rate of convergence than drift rate-based models (as discussed in the previous
 782 section), they showed only modest to low rates of acceptance for both lenient (CFI: \bar{x} =
 783 25.75%; RMSEA: \bar{x} = 24.95%) and strict thresholds (CFI: \bar{x} = 5.09%; RMSEA: \bar{x} = 2.73%).
 784



785

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

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

792

793

794 **Relationship between model convergence, acceptance, and complexity**

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

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

811 **Model selection**

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

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

824

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

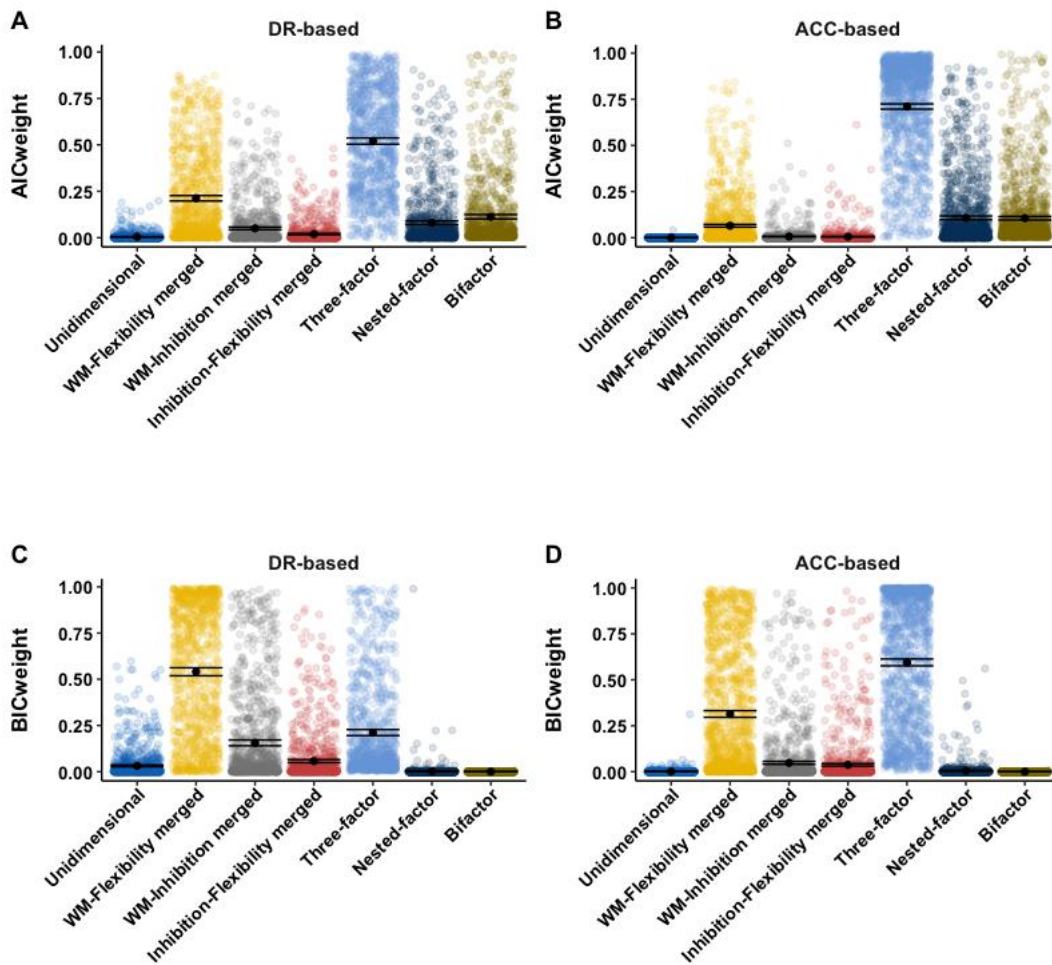
830

831

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

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

843 In sum, despite the AIC and BIC fit indices yielding a slightly different pattern of
 844 results, the three-factor model appears, overall, as the best model among the seven
 845 measurement models tested, which is also the best fitting model based on CFI and RMSEA
 846 indices (see table S6 online supplementary material, *Mean CFI and RMSEA, and percent of*
 847 *measurement models that meet lenient thresholds*).
 848



849
 850 **Figure 6.** AIC weights and BIC weights of measurement models of executive functions, as a
 851 function of drift rate-based and accuracy-based models.
 852

853 *Note.* Figure 6A: AIC weights drift rate-based models; Figure 6B: AIC weights accuracy-based
 854 models; Figure 6C: BIC weights drift rate-based models; Figure 6D: BIC weights accuracy-based
 855 models. Black dots represent mean weights; top and bottom horizontal black bars represent
 856 bootstrapped 95% confidence intervals. Note that the relative model probabilities are normalized by
 857 dividing by the sum of the probabilities of all the models.
 858

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

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

871 The drift rate-based three-factor model on average showed good fit to the data (CFI:
 872 $\bar{x} = .95$; RMSEA: $\bar{x} = .06$). The observed correlations between the three latent constructs on
 873 average were high, inhibition-cognitive flexibility ($\bar{r} = .75$), inhibition-working memory ($\bar{r} =$
 874 $.76$), working memory-cognitive flexibility ($\bar{r} = .79$), and were stronger compared to those
 875 observed in three-factor accuracy-based model. Importantly, all the tasks loaded significantly
 876 into their corresponding latent construct, showing moderate-to-strong factor loadings ($\bar{x} = .61$;
 877 $\text{min} = .41$; $\text{max} = .75$). The accuracy-based three-factor model on average showed acceptable
 878 fit to the data (CFI: $\bar{x} = .90$; RMSEA: $\bar{x} = .08$). The three constructs on average were strongly
 879 correlated, inhibition-cognitive flexibility ($\bar{r} = .59$), inhibition-working memory ($\bar{r} = .56$),
 880 working memory-cognitive flexibility ($\bar{r} = .62$). All the tasks loaded significantly into their
 881 corresponding latent construct and more importantly, they showed moderate-to-strong factor
 882 loadings ($\bar{x} = .60$; $\text{min} = .43$; $\text{max} = .76$).

883 In sum, both indicator-based models achieved satisfactory factor loadings, except for
 884 the Backwards digit-span task and the Letter-Memory task (working memory latent construct).
 885 In addition, both indicator-based models included latent factors that achieved levels of
 886 reliability remarkably higher compared to any other indicator-based model (see Table S7).
 887

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}	<i>SD</i>	\bar{x}	<i>SD</i>	<i>r</i>	<i>SD</i>	<i>r</i>	<i>SD</i>	<i>r</i>	<i>SD</i>
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

888 *Note.* This information corresponds to three-factor models that converged without warnings/errors out of 5000
 889 data sets. DR-based (n) = 4710 data sets; ACC-based (n) = 4813 data sets.
 890

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)
Inhibition	Flanker	0.71 (± .06)	0.71 (± .07)
	Simon	0.65 (± .06)	0.56 (± .07)
	Go/No-Go	0.74 (± .06)	0.74 (± .07)
<i>Omega H index</i>		0.75 (± .03)	0.71 (± .04)
		0.76 (± .04)	0.75 (± .05)
Cognitive flexibility	Color-Shape switch	0.50 (± .08)	0.50 (± .09)
	Gender-Smile switch	0.67 (± .07)	0.60 (± .10)
	Trail Making Test	0.58 (± .07)	0.61 (± .09)
<i>Omega H index</i>		0.61 (± .05)	0.60 (± .05)
		0.63 (± .05)	0.63 (± .06)
Working memory	Backwards Digit-Span	0.43 (± .08)	0.46 (± .09)
	Spatial n-Back	0.75 (± .08)	0.76 (± .10)
	Letter-Memory	0.41 (± .07)	0.43 (± .08)
<i>Omega H index</i>		0.55 (± .06)	0.57 (± .06)
		0.65 (± .08)	0.67 (± .11)
	Mean factor loadings	0.61	0.60
	Median factor loadings	0.63	0.60

891 *Note.* * Models that converged with no warning and errors. EF: executive functions; DR: drift rate; ACC:
 892 accuracy. In brackets (): standard deviations.
 893

894

Discussion

895

896

897

898

899

900

901

902

903

904

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.

905

906

907

908

909

910

911

912

913

914

915

916

917

918

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

919 the present study is that measurement models that included RT-based measures fared quite
920 poorly as compared to measurement models that included accuracy-based measures, which
921 showed much greater acceptance rates. A third finding is that drift rate, modeled by the EZ-
922 diffusion model, was the indicator that showed the best psychometric properties to model
923 executive functions, in terms of both model convergence and acceptance. A fourth main finding
924 is that when considering models that converge, the three-factor model remains the most
925 preferred model based on the direct comparison of the fit indices of the seven measurement
926 models of executive functions tested. Finally, measurement models that included drift rate as
927 the main indicator, showed comparatively moderate to high contribution (i.e., factor loadings)
928 from most of the tasks into their corresponding latent construct, providing a path forward for
929 re-analysis of existing data set or future ones, as the tasks used in the present study tend to be
930 commonly used in the field of executive functions.

931 **The advantage of using single scores over difference scores in measurement models of** 932 **executive functions**

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

941 The use of the difference score method has produced robust experimental effects, such
942 as the well-known Stroop effect (Stroop, 1935), Simon effect (Simon & Rudell, 1967), Flanker
943 effect (Eriksen & Eriksen, 1974), or task-switch costs (Monsell, 2003). However, robust

944 experimental effects often fail to produce reliable individual differences in cognition (Hedge
945 et al., 2018). This phenomenon likely occurs due to distinct nature of two historical approaches
946 in psychological research, experimental research versus correlational research. The former
947 aims to characterize the cognitive mechanisms underlying responses to different experimental
948 manipulations (e.g., within-subject variance), whereas the latter aims to characterize the
949 cognitive mechanisms underlying inter-individual differences in cognitive processing (e.g.,
950 between-subjects variance). Recently, Hedge et al. (2018) assessed in three studies the
951 reliability of seven classical cognitive tasks using different scoring methods (i.e., mean RT, RT
952 difference, accuracy rate, and accuracy difference), with most tasks showing test-retest
953 reliabilities below .70 when performance was operationalized through the difference score
954 method. Thus, the pattern of low convergence of condition-difference indicator-based models
955 is likely due to the psychometric issues of difference scores reported in the literature by Hedge
956 et al. (2018), in line with other claims (Caruso, 2004; Draheim et al., 2016, 2019). Accordingly,
957 several studies have consistently shown that RT difference measures tend to show weaker
958 association with other variables of interest, compared to the single RT measures from which
959 RT differences are computed, as discussed in the introduction (e.g., Hughes et al., 2014; Paap
960 & Sawi, 2016; Salthouse et al., 1998; Siegrist, 1997). The main reason is that the subtraction
961 method removes part of the common variance between the two variables from which the score
962 difference is computed increasing the proportion of error variance of this sort of measures
963 (Cronbach & Furby, 1970; Hedge et al., 2018).

964 To shed further insight into the psychometric issues of condition-difference indicator-
965 based models, we looked at the factor loadings of the tasks' indicators in the three-factor model,
966 for each indicator-based model (see online supplementary material, Table S7). Factor loadings
967 of condition-difference indicator-based models were, on average rather low (RT difference:
968 $\bar{x} = .38$); Accuracy difference: $\bar{x} = .40$; IES difference: $\bar{x} = .38$; Drift rate difference: $\bar{x} =$

969 .34), compared to most single-condition indicator-based models (RT: $\bar{x} = .47$; Accuracy: $\bar{x} =$
 970 $.60$; IES: $\bar{x} = .55$; Drift rate: $\bar{x} = .61$). Lower factor loadings have been associated with
 971 convergence issues regardless of sample size (Gagné & Hancock, 2006; Marsh et al., 1998).
 972 Thus, the systematic lower factor loadings observed in condition-difference indicator-based
 973 models, possibly account for their convergence issues. It is also in line with a recent
 974 investigation that reported a pattern of factor loadings that differed between difference scores
 975 and single scores, when modeling tasks such as the Stroop, Simon, Flanker, Global-Local ones
 976 with CFA (Rey-Mermet et al., 2021). In addition, the latent factors from condition-difference
 977 models showed consistently lower reliability (i.e., omega ω and H index) compared to those
 978 from single-condition indicator-based models. This is a matter of concern because many latent
 979 variable studies still use difference scores, such as RT difference, to operationalize executive
 980 functions in their RT-based tasks (e.g., Agostino, Johnson, & Pascual-Leone, 2010; Brydges et
 981 al., 2014; Duan, Wei, Wang, & Shi, 2010; Xu et al., 2013). Although less commonly used,
 982 accuracy difference has the same psychometric issues than RT difference, but the former is
 983 more prone to reduced variance issues due to ceiling effects (Wang et al., 2008).

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

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

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

992 RT-based models showed moderate rates of convergence, and very low rates of
 993 acceptance for both fit index and thresholds. Accuracy-based models, on the other hand, show

994 greater convergence rates and low-to-moderate rates of acceptance, suggesting better
995 psychometric properties than RT-based measures. Note that the relative low rate of acceptance
996 of RT-based and accuracy-based models was to some extent expected given the low rates of
997 acceptance across executive functions models reported by Karr et al. (2018), as most of the
998 studies included in their re-analysis used RT or accuracy-based measures to operationalize
999 executive functions (e.g., Carlson et al., 2014; Lee et al., 2013; Lerner & Lonigan, 2014; Miller
1000 et al., 2012; Monette et al., 2015; Rose et al., 2012; Wiebe et al., 2011).

1001 The psychometric issues of RT-based measures are likely due to a combination of
1002 factors, such as the low reliability of RT-based measures and their sensitivity to speed-accuracy
1003 trade-offs. This is not the first work to highlight such weaknesses. For example, Miller and
1004 Ulrich (2013) proposed a model to investigate the psychometric properties of mean RT and RT
1005 difference to predict individual differences on these measures. Their model estimated three
1006 different parameters underlying RTs, that is, (i) common processing speed across tasks, (ii)
1007 processing speed for individual tasks, and (iii) residual differences in RT related to neither
1008 general nor task-specific processing speed. Their model showed that mean RT can be reliable
1009 across tasks, provided that the model parameters show a reasonable amount of variability. As
1010 a consequence, mean RT reliability can come to depend on a parameter of no interest for
1011 researchers such as the residual differences in RT, a state of affair which is problematic.
1012 Interestingly, their model showed that the mechanisms underlying RT during decision making
1013 are far more complex than what is commonly assumed in the literature, in line with recent views
1014 (Frischkorn & Schubert, 2018; Ratcliff et al., 2016). A well-known source of that complexity
1015 arises from the sensitivity of RT-based measures to speed-accuracy trade-off, whereby
1016 individuals adopt different response strategies emphasizing speed over accuracy and vice-versa
1017 (Starns & Ratcliff, 2012). Research exploring individual or developmental differences in
1018 cognition based on RT analyses might be particularly impacted by such trade-offs (Draheim et

1019 al., 2016; Hertzog et al., 1993; Yang et al., 2015), with the potential for misleading conclusions
 1020 due to the complex interplay between speed and accuracy.

1021 Accuracy-based models although showing higher rates of convergence, displayed rather
 1022 low-to-moderate rates of acceptance. Accuracy-based measures suffer from two weaknesses.
 1023 First, they are often subject to speed-accuracy trade-offs, and in doing so confound information
 1024 processing efficiency with strategic issues of boundary setting (Drugowitsch et al., 2015; Maris
 1025 & van der Maas, 2012; Ratcliff & Rouder, 1998; Starns & Ratcliff, 2012). Second, accuracy-
 1026 based models often suffer from a lack of sensitivity with most individuals operating around a
 1027 narrow range of high accuracies. It has been noted before that accuracy-based measures are
 1028 reliable for individual differences research, only if individuals make enough errors during the
 1029 task and thus, there exist sufficient between subjects' variability in the rate of errors (Wang et
 1030 al., 2008). Note this is the case of the present study. Although the level of accuracy (%) on our
 1031 RT-based tasks was relatively high in the most demanding tasks' conditions (e.g., incongruent,
 1032 switch, 2-back trials), the variability was quite large even after removing univariate outlier
 1033 values (Flanker: $\bar{x} = 87\%$, range = 56-100 %; Simon: $\bar{x} = 88\%$, range = 53-100 %; Go/No-
 1034 Go: $\bar{x} = 93\%$, range = 74-100 %; Color-Shape switch: $\bar{x} = 82\%$, range = 50-99 %; Gender-
 1035 Smile switch: $\bar{x} = 87\%$, range = 55-99 %; N-back: $\bar{x} = 76\%$, range = 40-100 %. Nevertheless,
 1036 despite their low-to-moderate acceptance rates, it is important to point out that the most likely
 1037 accuracy-based model (i.e., three-factor), on average showed acceptable fitting to the data.
 1038 Furthermore, all the tasks showed moderate-to-high factor loadings, except for the Backwards
 1039 digit-span task and Letter-Memory updating task, and more importantly, most of the tasks
 1040 showed similar factor loadings, which indicates that the three latent factors of the model
 1041 reflected to a greater or lesser extend common variance across the tasks. This is in line with
 1042 accuracy-based measure having been useful to investigate CFA measurement models of
 1043 executive functions (e.g., Agostino et al., 2010; Alfonso & Lonigan, 2021; Brocki & Tillman,

1044 2014; Carlson et al., 2014; Lerner & Lonigan, 2014; Masten et al., 2012; Monette et al., 2015;
1045 Willoughby et al., 2012), working memory (Engle, 2018; Luck & Vogel, 2013; Waris et al.,
1046 2017; Wilhelm et al., 2013), and of intelligence (Conway et al., 2002; Rey-Mermet et al., 2019).
1047 Moreover, accuracy-based and capacity measures have been the measures of excellence in the
1048 broad field of working memory (Engle, 2018; Luck & Vogel, 2013; Waris et al., 2017; Wilhelm
1049 et al., 2013). This is the case during clinical and educational evaluations with the forward and
1050 backward digit span task, as part of standard batteries such as the WAIS (Wechsler, 1981) or
1051 the WISC (Wechsler, 1991), as well as in most laboratory experiments, albeit using more
1052 sophisticated forms of span tasks, such as the operation span (Kane et al., 2004). The
1053 introduction about 20 years ago of change detection tasks to measure working memory capacity
1054 point to possible changes. In particular, recently several authors have proposed to go beyond
1055 capacity, as measured in terms of memory slots (Rouder et al., 2011; Zhang & Luck, 2008), to
1056 rather characterize working memory in terms of processing efficiency (Lew & Vul, 2015; Ma
1057 et al., 2014).

1058 In sum, our results confirm that the use of accuracy-based models, despite the two
1059 weaknesses described above, is a sound choice to model executive functions based on (i) their
1060 high convergence rates, and (ii) the fact that the most likely model of executive functions,
1061 which also tends to be the most accepted model in the literature, showed acceptable fitting to
1062 the data, and moderate-to-high factor loadings.

1063 ***The promising psychometric properties of Drift Rate to model executive functions with***
1064 ***latent variable methods***

1065 An issue when considering just RTs or accuracy is that neither fully capture behavioral
1066 performance as illustrated by the speed-accuracy trade-off discussed earlier. To address this
1067 issue, other indicators have been developed, such as the IES, which is often used in the

1068 developmental or aging literature (Vandierendonck, 2017), or drift rate, a measure rarely
1069 considered in the executive functions' literature.

1070 The present work tested the IES, a measure that combines speed and accuracy in a single
1071 metric in an effort to mitigate the issue of speed-accuracy trade-offs (Townsend & Ashby,
1072 1978). IES-based models on average showed moderate rates of convergence and very low rates
1073 of acceptance. These results were very similar to the ones observed on RT-based models, which
1074 was not surprising since an individual's IES can be considered as the RT corrected by the
1075 proportion of errors committed.

1076 A unique contribution of the present study is to show that drift rate has excellent
1077 psychometric properties to model executive functions with latent variable methods, such as
1078 CFA. Drift rate-based models were the most stable in terms of model convergence and
1079 acceptance, showing moderate-to-high rates of convergence and acceptance, across all
1080 measurement models of executive functions, regardless of fit index (CFI or RMSEA) and
1081 thresholds (lenient or strict).

1082 Drift rate is an appealing measure as it can be relatively easily computed from the EZ-
1083 diffusion model, provided each participant's response accuracy, mean RT and RT standard
1084 deviation is available (Wagenmakers et al., 2007). By acknowledging that speed and accuracy
1085 arise from the same underlying generative process of integration to a bound, diffusion models
1086 allow to assess the rate at which information processing accumulates, or the quality of
1087 information processing, separately from the height of the bound to be reached for a decision to
1088 be triggered. In doing so, drift rate provides a better estimation of information processing
1089 quality than RT or accuracy separately (Ratcliff et al., 2016). Accordingly, drift rate, by
1090 capturing sensitivity to information processing, appears as a more valid representation of the
1091 executive processes evaluated than using RT-based or accuracy-based measures, which are
1092 conflated with other processes such as response conservativeness or non-decision time.

1093 Although seldom used in executive functions latent variable research, the reliability and
1094 stability of DDM parameters has been previously documented. The within-session and
1095 between-session reliability of DDM parameters in a lexical decision task was investigated by
1096 Yap et al. (2012 - $n = 819$) and confirmed by Lerche and Voss (2017 - $n = 105$). More precisely,
1097 Yap et al. (2012) reported that the three parameters of the DDM of greatest psychological
1098 interest showed excellent within-session reliability (drift rate: .81; boundary separation: .91;
1099 non-decision time: .93), and acceptable between-session reliability (drift rate: .69; boundary
1100 separation: .71; non-decision time: .72). Moreover, some works point to the criterion validity
1101 of diffusion model parameters, in particular drift rate, in the context of individual differences
1102 in intelligence. These studies have shown that a drift rate latent factor (derived from non-
1103 executive RT-based tasks) is associated with intelligence, thus suggesting that individuals with
1104 larger drift rate show greater scores in tests of intelligence (Lerche et al., 2020; Ratcliff et al.,
1105 2010; Schmiedek et al., 2007; Schmitz & Wilhelm, 2016). Furthermore, Schmiedek et al.
1106 (2007) and Schmitz and Wilhelm (2016) included measures of working memory capacity, and
1107 both studies modeled latent constructs for each of the three main parameters of the DDM (i.e.,
1108 drift rate, boundary separation, and non-decision time). Interestingly, the drift rate latent
1109 construct was the main predictor of working memory capacity and intelligence latent
1110 constructs, whereas boundary separation and non-decision time showed very low associations
1111 with intelligence. Therefore, the better psychometric properties shown by drift rate in the
1112 present study are in line with the results from previous studies, which have highlighted the
1113 reliability and validity of drift rate for research in cognitive psychology.

1114 Finally, post-hoc analyses, not included in the present study, were performed to
1115 examine the convergence and acceptance rates of measurement models of executive functions
1116 when drift rate was replaced by, either boundary separation or non-execution time, the other
1117 two parameters from the EZ-diffusion model. These models showed extremely poor

1118 convergence and acceptance rates, which was not surprising given that drift rate is the main
1119 parameter of the DDM capturing the quality of information processing. Accordingly, drift rate
1120 has been consistently associated with cognitive ability, such as fluid intelligence or working
1121 memory (Lerche et al., 2020; Ratcliff et al., 2010; Schmiedek et al., 2007; Schmitz & Wilhelm,
1122 2016).

1123 **The interplay between model complexity and the rate of convergence and acceptance**

1124 Our results indicate that the simplest models (i.e., unidimensional, two-factor) despite
1125 converging more often, on average showed low rates of model acceptance, whereas the most
1126 complex models showed an opposite pattern. These results are expected and in line with the
1127 re-analysis of published models from Karr et al. (2018), who observed a similar trade-off
1128 between model convergence/acceptance and model complexity in both children/adolescent and
1129 adult samples.

1130 The likelihood that a model converges and fits well the data depends on multiple
1131 factors, but mainly on a complex interplay between sample size, model parameters, and the
1132 level of commonality between the variables that form the model (Gagné & Hancock, 2006;
1133 Kyriazos, 2018; MacCallum et al., 1999; Wolf et al., 2013). These are important aspects of the
1134 study design, which along with statistical power, must be considered by researchers prior to
1135 conducting their study. More complex models, which have more parameters, often need larger
1136 samples to converge than simpler models (Green & Yang, 2018; Kline, 2016; Nicolaou &
1137 Masoner, 2013). In our case, the two most complex models, the nested-factor and bifactor
1138 models, had an important difference compared to the simpler models; they both included a
1139 common factor and specific sub-factors. Thus, the indicators had to load at the same time in
1140 the common factor and their corresponding sub-specific factor (note that in the nested-factor
1141 model, indicators from inhibition tasks loaded only in the common factor). There are several
1142 reasons that might explain why nested-factor and bifactor models showed a remarkably lower

1143 rate of convergence. A first reason concerns sample size; our study may be underpowered to
1144 identify these models with high frequency. Unlike in other domains of psychology, the field of
1145 latent variable research has not yet fully tackled how to best determine the sample size prior to
1146 conducting a study. There exist recommendations about how to estimate the sample size based
1147 on Monte Carlo simulation studies. For instance, Tanaka (1987) suggested a ratio 5:1 between
1148 N and parameters, whereas Bentler & Chou (1987) suggested a ratio of 10:1. More recently,
1149 Wolf et al. (2013), conducted a series of Monte Carlo simulations to understand sample size
1150 requirements, as a function of model type, number of factors and indicators, strength of the
1151 factor loadings and the amount of missing data. Their results show that the sample size
1152 requirements are far more complex than the recommendations from Tanaka (1987) and Bentler
1153 & Chou (1987). They observed that sample size requirements ranged from 30 cases (one-factor
1154 model with four indicators loading at .80), to 460 cases (two-factor model with three indicators
1155 per factor loading at .50). Importantly, they pointed out that inter-indicator correlations and the
1156 factor loadings also play an important role on statistical power and model identification,
1157 beyond sample size. Thus, our two most complex models, nested-factor and bifactor models,
1158 may be difficult to identify due to the complex interplay between sample size, the number of
1159 model parameters, the inter-indicator correlations, and the factor loadings from each indicator
1160 into both the common factor and their corresponding specific sub-factor.

1161 The rate of acceptance followed an opposite pattern. That is, more complex models,
1162 when they converged, tended to have higher rates of acceptance than simpler models. These
1163 results are consistent across indicator-based models. The observed higher acceptance of two-
1164 factor and three-factor models over unidimensional models was expected given that an
1165 important number of studies have shown that executive functions tend to organize as a two-
1166 factor or three-factor structure during middle childhood (Brydges et al., 2014; Lehto et al.,
1167 2003; Rose et al., 2012). The high acceptance of nested-factor and bifactor models was less

1168 expected, as the use of these models is not an extended practice in executive functions research.
1169 The nested-factor model has been proposed in adolescents and adults studies (e.g., Friedman,
1170 Miyake, Robinson, & Hewitt, 2011; Friedman et al., 2008), and one study in children 9-to-12
1171 years old (Sluis et al., 2007). The bifactor model reported in the present study and in Karr et
1172 al. (2018) is uncommon in children's executive functions research; indeed, a bifactor model
1173 based on a battery of tasks tapping inhibition, working memory and cognitive flexibility is
1174 rarely considered (for an exception see Yangüez et al., 2021). The higher acceptance of these
1175 two models must be taken with caution since there are several concerns about the tendency of
1176 these models to overfit the data (Bonifay et al., 2017; Karr et al., 2018; Murray & Johnson,
1177 2013; Sellbom & Tellegen, 2019). Indeed, as noted by Hancock and Mueller (2008), a model
1178 with better fitting does not necessarily represent the true model for the population, but perhaps
1179 it is just a model that captures the data well, thanks in part to their ability to overfit the data
1180 (Preacher et al., 2013). Re-analyses with models that have a better balance between
1181 convergence and acceptance seems to be a healthy practice to adopt for the field. Unfortunately,
1182 rarely papers report on the necessary simulations to estimate the rate of convergence, when
1183 applying CFA to more practical ends. The present work calls for a more systematic assessment
1184 and report of convergence rates as a given model structure is chosen.

1185 **Impact of tasks' operationalization on model selection**

1186 A concern raised by the present study is that of the dependence of the most likely model
1187 on how executive functions are operationalized. Given the expected differences between
1188 indicator-based models in model convergence and acceptance, one could have expected that
1189 the preferred model, based on direct comparison of fit indices, would differ across indicator-
1190 based models. Only drift rate-based and accuracy-based models provided sufficient data sets
1191 to conduct model selection based on the seven measurement models of executive functions
1192 tested. This state of affair limits our understanding of the impact of the choice of indicators on

1193 model selection, since RT-based or IES-based models, as well as condition-difference
1194 indicator-based too often failed to converge and showed poor fitting to the data.

1195 Nevertheless, the results of model selection indicate the preferred model to be the three-
1196 factor model, and the next preferred one being the two-factor model (working memory-
1197 cognitive flexibility merged). These results are in line with the bulk of the literature on
1198 executive functions reporting a three-factor structure on children of a similar age range
1199 (Agostino et al., 2010; Arán Filippetti & Richaud, 2017; Duan et al., 2010; Lehto et al., 2003;
1200 Rose et al., 2012), although for a sample of 8-to-12 years old, a two-factor structure has been
1201 also documented (Brydges et al., 2014; Huizinga et al., 2006; Lee et al., 2013; Monette et al.,
1202 2015; Scionti & Marzocchi, 2021; Usai et al., 2014; Van der Ven et al., 2013). That is, when it
1203 was directly compared against alternative models, the three-factor model was the best fitting
1204 model regardless of which indicator was used (i.e., accuracy or drift rate), except for drift rate-
1205 based models when using BIC_w ; here, the two-factor model merging working memory and
1206 cognitive flexibility was preferred. The discrepancy between AIC_w and BIC_w most likely
1207 reflects how both fit indices penalize model complexity with the BIC favoring more parsimony
1208 than the AIC (Vrieze, 2012). The finding that the second most preferred model might be a more
1209 parsimonious two-factor model, with working memory merged with cognitive flexibility is in
1210 line with the view that cognitive flexibility develops later (Diamond, 2013; Karr et al., 2018).
1211 It is also well aligned with previous latent variable studies that report a non-differentiated
1212 cognitive flexibility (also termed shifting in some works) factor from working memory in pre-
1213 school and school-aged children (Monette et al., 2015; Scionti & Marzocchi, 2021; Usai et al.,
1214 2014). Indeed, our sample included children with a varied age range extending from 8 years of
1215 age all the way to 12 years of age, a period of development during which executive functions
1216 undergo rapid developmental changes (Zelazo et al., 2016). Interestingly, not only
1217 improvements in efficiency whereby children become faster and more precise have been

1218 documented during that age range (e.g., Anderson et al., 2001; Davidson et al., 2006; Huizinga
1219 et al., 2006; Lee et al., 2013; Zelazo & Carlson, 2012), but also qualitative changes in the
1220 structure and organization of executive functions with greater likelihood of two-factor models
1221 in younger samples and of the three-factor model in older samples (e.g., Brydges et al., 2014;
1222 Huizinga et al., 2006; Lee et al., 2013; Lehto et al., 2003; Rose et al., 2012; Wiebe et al., 2011).

1223 Last but not least, to confirm the selected model fits well to the data, we looked at the
1224 average CFI and RMSEA values shown by accuracy-based and drift rate-based models, which,
1225 unlike AIC or BIC, can be interpreted based on their absolute value (Hu & Bentler, 1999). Both
1226 fit indices show not only the three-factor model on average fits well the data, but also that it
1227 shows better fit to the data than the second model that received further support based on AIC
1228 and BIC weights (i.e., two-factor model with cognitive flexibility and working memory
1229 merged). Indeed, among the seven measurement models tested, the three-factor model showed
1230 the best trade-off between convergence, acceptance, and parsimony. Importantly, this was the
1231 case whether all samples were considered (see table 3) or only the subset used for model
1232 selection (for the interested reader see table S6). It would seem good practice in future works
1233 that apply AIC/BIC model selection as we did here, to also check that the selected models fit
1234 well data by looking at indices that can be interpreted based on their absolute values, such as
1235 the CFI and RMSEA as used in the present study, or root mean squared residual (SRMR), or
1236 the goodness-of-fit index (GFI) among others (for a review see, Hu & Bentler, 1999). Finally,
1237 it would also seem important when evaluating best models to check for model convergence
1238 more systematically, as one should avoid drawing strong conclusions based on well-fitting
1239 models that often show convergence or estimation issues.

1240 **Limitations**

1241 This study offers a comprehensive empirical evaluation of the sensitivity of cognitive
1242 functions latent variable models to different methods proposed in the literature to

1243 operationalize such functions. We have provided a list of robust findings based on thousands
1244 of models and bootstrapped samples. However, these findings should also be interpreted
1245 considering the limitations of this work.

1246 First, we could not derive the same exact indicators for each task because for some
1247 tasks, either RT (i.e., Backwards digit-span task, Letter-Memory task) or accuracy (Trail
1248 Making Test) were not recorded. Ideally, studies should aim for homogeneous models with the
1249 same type of indicators. Indeed, mixed-indicators models that use the most common indicator
1250 for each task appear not only weak on theoretical grounds, but also arbitrary in its application.
1251 For instance, on RT-based tasks (i.e., Flanker, Simon, Color-Shape switch, and Gender-Smile
1252 switch), the most common measure is unclear as both RT and accuracy have been used as
1253 relevant indicators of performance in such tasks, as discussed in the introduction. Models that
1254 search for the combination of indicators that result in the greatest rates of convergence and
1255 acceptance, would result in an explosion of combinations and not be necessarily valid.
1256 Therefore, due to the wide variety of possible combinations of indicators across tasks, the
1257 present work was systematic in the choice of indicators, by creating the most homogeneous
1258 model for each dominant indicator. It is possible that a principled combination of indicators,
1259 for instance, combining drift rate as indicator of performance on RT-based tasks (e.g.,
1260 inhibition and cognitive flexibility), with accuracy-based or capacity-based measures when
1261 modeling performance in capacity tasks (e.g., working memory) may be a possible avenue for
1262 future research.

1263 Second, the sample size of the present study was modest ($n = 182$). As discussed above,
1264 there is little consensus about how to estimate the required sample size for latent variable
1265 models, even if there exist some recommendations (or rules of thumb) from several studies
1266 based on the results from Monte Carlo simulations (Bentler & Chou, 1987; MacCallum &
1267 Austin, 2000; Tanaka, 1987; Wolf et al., 2013). Thus, given the sensitivity of latent variable

1268 models to sample size, model complexity, and the quality of the data, it would be welcome to
1269 see the present results confirmed in an even bigger sample. For now, we note that the sample
1270 size of this study ($n = 182$) is comparable to the one reported by most of the studies from the
1271 field (e.g., Brydges et al., 2014; A Miyake et al., 2000; Rose et al., 2012; Sluis et al., 2007;
1272 Usai et al., 2014; Van der Ven et al., 2013).

1273 Third, the present work was not designed to address whether the different
1274 operationalizations of executive functions investigated lead to valid representations of the
1275 constructs. To the extent that we limited ourselves to well-known tasks that are accepted in the
1276 executive functions literature, this work is in line with the view that a key to validity is the
1277 choice of tasks, that is, whether the task measures what is intended to measure (Borsboom et
1278 al., 2009). To the extent that factor loadings in the preferred single-condition indicator-based
1279 models were to a greater or lesser satisfactory (e.g., three-factor model), this work is in line
1280 with the existing literature interested in identifying which operationalizations of performance
1281 in executive functions tasks, allow to reliably capture individual differences in the underlying
1282 cognitive constructs evaluated.

1283 Fourth, the seven measurement models of executive functions tested were identified in
1284 a literature review that shows a strong influence by Miyake & Friedman's work, adopting a
1285 CFA measurement approach, that is a *reflective* approach, to examine the structure of executive
1286 functions (Friedman et al., 2008; Miyake, Friedman, et al., 2000; Shah & Miyake, 1996). Yet,
1287 other statistical methods exist to assess cognitive and psychological processes, such as
1288 *formative* models (e.g., principal component analysis), which are less affected by the low
1289 shared variance between executive functions tasks (Willoughby et al., 2014; Willoughby &
1290 Blair, 2016). Exploratory SEM (ESEM) could also be an interesting approach in future works
1291 as it overcomes some limitations of CFA, such as model misspecification and misfit (Perry et
1292 al., 2015), and the excessive flexibility of exploratory factor analysis (Marsh et al., 2014).

1293 Traditional ESEM, nevertheless, has been challenged since it often lacks parsimony, and might
1294 cluster together constructs that are supposed to be separated in relation to theory, specially,
1295 when ESEM is applied to complex models and small samples (Marsh et al., 2014). Another
1296 interesting extension could be to complement CFA-based model selection with exploratory
1297 approaches, such as exploratory factor analysis (EFA), as Waris et al. (2017) did to investigate
1298 the structure of working memory. Finally, network modeling, which proposes that cognitive
1299 processes are conceptualized as networks of directly related manifested variables, has been
1300 proposed to overcome the limitations of *reflective* and *formative* models (Schmittmann et al.,
1301 2013). Interestingly, network modeling has proven to be an effective tool to study the
1302 differentiation process of executive functions during development (Hartung et al., 2020; Karr
1303 et al., 2022).

1304 Finally, we recognize there also exists alternative modeling specification approaches to
1305 the seven considered here. For instance, second-order (hierarchical) models, which are more
1306 common in intelligence research (Canivez et al., 2019; Reynolds & Keith, 2017; Schneider &
1307 Newman, 2015) could be evaluated as recently done by Hartung et al. (2020) and Wolff et al.
1308 (2016) in the case of executive functions. Similarly, other bifactor structure could be considered
1309 (e.g., inhibition and cognitive flexibility tasks loading into the same specific factor and working
1310 memory tasks loading into another specific factor). In short, future works will certainly benefit
1311 from combining the strengths of different statistical techniques (e.g., CFA and networking
1312 modeling) to further investigate the underlying structure of executive functions; in doing so,
1313 the present work highlights the importance of not just the task selection but also the choice of
1314 indicators.

1315 **Conclusions**

1316 The present study confirms the sensitivity of measurement models of executive
1317 functions to the different methods proposed in the literature to operationalize executive
1318 functions. Below we summarize the highlights from this work:

- 1319 • Difference scores should be avoided when modeling executive functions with latent
1320 variable methods. Measurement models that included difference scores often failed to
1321 converge and showed poor fit to the data, but in addition, they showed systematically
1322 lower factor loadings compared to measurement models that included single scores.
- 1323 • RT-based models showed poor fit to the data compared to measurement models
1324 including accuracy-based measures.
- 1325 • Drift rate showed the best psychometric properties for CFA models of executive
1326 functions among the four indicators tested (RTs, Accuracy, IES and drift rate). Of note,
1327 drift rate can be easily computed from individual's response accuracy, mean RT, and
1328 RT deviation through the EZ-diffusion model (Wagenmakers et al., 2007).
- 1329 • This work highlights the benefit of homogenizing indicators through the use of either
1330 accuracy or drift rate to reach acceptable levels of convergence and acceptance, as well
1331 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](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., Touloupoulou, 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](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 Ed.). 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](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., Picchiatti, 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/10.31234/osf.io/8pwd6>

- Roelofs, A. (2018). Acta Psychologica One hundred fifty 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 predictable 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., Roebbers, 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/PsychologyScience/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](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., & Greenberg, 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). *Grundzüge 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-](https://doi.org/10.3758/s13428-018-1055-2)

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>