Comparing the executive function ability of autistic and non-autistic adolescents with a manualised battery of neuropsychological tasks

Lorcan Kenny[1], Anna Remington[1], and Elizabeth Pellicano[1,2,]

[1]Centre for Research in Autism and Education (CRAE), UCL Institute of Education, University College London, UK; [3]Macquarie School of Education, Macquarie University, 29 Wally's Walk, Sydney, Australia.

*Corresponding author:

Lorcan Kenny

Centre for Research in Autism and Education (CRAE)

UCL Institute of Education

University College London

55–59 Gordon Square

London WC1H 0NU, UK.

Email: l.kenny@ucl.ac.uk

Abstract

Performance on a single executive function (EF) task (e.g., a card sorting task) is often taken to represent ability on the underlying subcomponent of EF (e.g., set shifting) without accounting for the non-specific and non-executive skills employed to complete the task. This study used a manualised battery of EF tasks to derive individual task scores and latent EF scores. Seventy-nine adolescents aged between 11 and 19 years, including 36 autistic and 42 non-autistic participants, matched on cognitive ability, completed the battery. Autistic adolescents had moderate global EF difficulties, but the samples did not differ on any of the specific individual subcomponents of EF (fluency, cognitive control, working memory). These findings prompt questions concerning EF and its measurement in autistic people.

There is an abundance of empirical research on executive function (EF) in autistic[1] people (see Demetriou et al., 2018, for meta-analytic review). Original theoretical accounts hypothesised that EF difficulties were universal in autism and causally related to the core features of the condition (Hughes, Russell, & Robbins, 1994; Russell, 1997). Many studies have indeed shown that autistic children, young people and adults have difficulties in those higher-order processes that underpin goal-directed activity and enable individuals to respond flexibly to change, including inhibiting prepotent but maladaptive responses, cognitive flexibility and future-oriented (or 'working') memory (Demetriou et al., 2018). Yet, there have also been many studies that have not demonstrated such difficulties (Bölte, Duketis, Poustka, & Holtmann, 2011; see Demetriou et al., 2018, for review; Pellicano, 2010; Schmitz et al., 2006), resulting in a literature that is mixed and challenging to interpret. One straightforward explanation for the discrepant findings relates to differences in the sampling characteristics of distinct studies. Another potential explanation, however, relates to inherent difficulties with measuring EF (Crawford & Channon, 2002; Jurado & Rosselli, 2007). This study sought to address some of these measurement difficulties (described below) by testing the executive abilities of cognitively-able autistic and non-autistic adolescents using an open-source, manualised EF battery (Kramer et al., 2014).

Current challenges to measuring EF in autism research

EF is an overarching term for a set of theoretical cognitive constructs that can be notoriously difficult to measure (Rabbitt, 2004). One methodological reason for why it has been difficult to make inferences from performance on EF tasks is what has been termed the 'task impurity problem' (Snyder, Miyake, & Hankin, 2015). The problem is such that performance on a single EF task is not simply composed of the systematic variance related to the EF construct of interest, but is also composed of the systematic variance related to EF ability in general, as well

---

[1] In the autistic community, identity-first language, e.g., "autistic person", is often preferred to, and considered less stigmatizing than, person-first language, e.g., "person with autism" (Gernsbacher, 2017; Kenny et al., 2016; Sinclair, 1999). We therefore use identity-first language throughout.

as the systematic variance related to non-executive factors (e.g., intellectual ability, language skills, motor coordination, reaction time) and non-systematic variance (i.e., error). The impurity problem can be addressed, at least partially, by assessing participants on multiple tasks tapping a given EF construct and creating a latent variable from the shared variance that reflects the EF construct one aims to measure. Yet, most EF research in the field of autism has used a single task to represent a discrete subdomain of EF ability – which means that autism research may suffer from a particularly pronounced task impurity problem, potentially leading to two possible (erroneous) assumptions. One, that an EF construct holds strong explanatory power when, in fact, the variance in non-EF or general EF abilities might be driving the effects in which one is interested. Two, that an EF construct holds no explanatory power when, in fact, it does, but we lack tasks to measure the construct with sufficient purity.

Another potential obstacle to progress in the EF literature is the diversity of executive tasks used by researchers, which can vary according to the type of assessment (psychometric tests vs. experimental tasks), modality (verbal vs. visuospatial), presentation format (computerised vs. tabletop) and participant response (verbal or motor) (Demetriou et al., 2018). Moreover, experimental tasks are often developed in-house by a single research group, are not formally manualised and are not reported upon in sufficient detail to be fully reproducible. Consequently, iterative studies apparently using the same paradigms are tweaked, variables are calculated in new ways, different dependent variables are reported upon, all of which results in sets of findings that cannot be meaningfully synthesised. Indeed, Demetriou et al. (2018) could not perform their meta-analysis on studies with participants below the age of 6 years "to account for qualitative differences in the types of assessment instruments used in younger aged groups" (p. 1200). Two ways to address this issue are to (i) use EF measures that are manualised so that administration, scoring and dependent measure selection are standardised across participants, and (ii) avoid using proprietary measures so that identical tasks can be readily accessible and used across different research groups and populations.

The NIH-EXAMINER battery

One such manualised, open-source battery of tasks is the Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (EXAMINER; Kramer et al., 2014), developed for a National Institutes of Health (NIH) funded project. The NIH-EXAMINER battery was designed to integrate the cognitive, experimental and clinical EF literatures to produce a battery of valid and reliable domain-specific EF tasks that also provides measures of common variance (i.e., factor scores) among multiple measures within a specific domain (Kramer et al., 2014). The NIH-EXAMINER was based on the Miyake et al. (2000) model of EF, which found support for three related, but separable, core EF constructs that contribute to complex executive tasks, such as planning. The battery therefore includes tasks tapping mental set shifting, information updating and monitoring, inhibition of pre-potent responses and planning. Tasks tapping fluency ability have also been included. A confirmatory factor analysis demonstrated that both a unitary factor (global executive composite) and a three-factor model (fluency, cognitive control and working memory) characterise the data well, at least in typical adults (Kramer et al., 2014).

Thus far, the studies that have employed the NIH-EXAMINER battery have predominantly focused on patients with frontal lobe pathophysiology. One such study, showed that patients with behavioural variant Frontotemporal Dementia had comparable difficulties to those with Alzheimer's Disease on the NIH-EXAMINER measures of working memory, semantic fluency and sustained attention but had more difficulties with letter fluency and anti-saccade accuracy, relative to a typical comparison group (Kramer et al., 2014). Another study with neurological patients of mixed aetiology found that damage to different neuroanatomical structures associated with EF difficulties was related to performance on different tasks within the battery (Robinson, Calamia, Gläscher, Bruss, & Tranel, 2014).

There have been two studies, to our knowledge, using the NIH-EXAMINER battery with neurodevelopmental populations. In one study, with a sample of young people with sickle

cell disease, performance on all individual measures within the NIH-EXAMINER battery decreased with increasing neurologic morbidity and two factor scores were related to reduced white matter area (Schatz, Stancil, Katz, & Sanchez, 2014). In a second study, children and adolescents with Attention Deficit/Hyperactivity Disorder (ADHD) have been shown to score lower on the working memory but not on the fluency or cognitive control factor scores than those without ADHD (Schreiber, Possin, Girard, & Rey-Casserly, 2014). To our knowledge, the battery has not been employed to assess EF in autistic participants – and the current study sought to do just that.

The current study

The aims of this study were twofold. First, we sought to determine whether autistic adolescents (aged 11 to 19 years) would show, on average, greater difficulties on the latent variables generated to represent fluency, cognitive control, working memory and overall EF ability from the NIH-EXAMINER battery (Kramer et al., 2014) relative to non-autistic participants of similar age and intellectual ability. On the basis of existing research showing that autistic participants have difficulties on each EF subcomponent (Demetriou et al., 2018; Geurts, van den Bergh, & Ruzzano, 2014; Lai et al., 2017; Landry & Al-Taie, 2016; Wang et al., 2017), we had strong reasons to predict that our autistic participants would underperform relative to the non-autistic participants on each of the factor scores derived from the NIH-EXAMINER battery. Second, we examined the extent to which individual differences in non-executive background variables such as chronological age, verbal and non-verbal IQ and autistic features were related to task performance.

Method

Participants

Seventy-eight adolescents aged between 11 and 19 years (M = 14.75, SD = 2.15) were recruited through community contacts, including 36 autistic (12 female, 24 male) and 42 non-autistic (22 female, 20 male) participants. All autistic participants had received an independent

clinical diagnosis of autism, according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR; APA, 2000, or DSM-5; 2013), or The International Classification of Mental and Behavioural Disorders, tenth edition (ICD-10; WHO, 1992) criteria. Participants also scored above the threshold indicative of autism on either the parent-reported Social Responsiveness Scale, Second edition (SRS-2; Constantino & Gruber, 2012) or the Autism Diagnostic Observation Scale, Second edition (ADOS-2; Lord et al., 2012; data for one autistic participant were not available on either measure). None of the non-autistic participants obtained an SRS-2 t-score of 60 or above, suggesting that they did not show clinically-significant features of autism (Constantino & Gruber, 2012; see Table 1).

All participants were considered not to have an additional intellectual disability by virtue of having verbal, nonverbal and full-scale IQ scores of 70 or above, as measured by the Wechsler Abbreviated Scale of Intelligence – Second edition (WASI-II; Wechsler, 2011; see Table 1). Samples were matched group-wise by age, $t(76)=0.74$, $p=.465$, $d=0.17$, verbal IQ , $t(76)=0.28$, $p=.780$, $d=0.06$, nonverbal IQ, $t(76)=0.35$, $p=.730$, $d=0.08$, and gender distribution, $\chi^2(1, n=78)=2.14$, $p=0.14$, $\phi=0.19$.

Table 2 reports the background characteristics of the two groups. As expected, a greater number of our autistic participants were reported to have at least one co-occurring developmental or psychiatric disability and to be taking medication. Almost three quarters of autistic participants were from white backgrounds compared with less than half of the non-autistic participants; this difference was significant, Fisher's exact test, $p=.001$. Six non-autistic participants and one autistic participant did not speak English as their first language. Parents also reported the age at which they left full-time education, a widely used proxy for socioeconomic status (Liberatos, Link, & Kelsey, 1987), which was used to derive the number of years of education they received after they turned 16 years of age. There were no significant differences between the autistic ($M=5.11$, $SD=2.94$) and non-autistic ($M=5.39$, $SD=5.06$) participants in level of parental/caregiver education, $t(61) = 0.27$, $p = .785$, $d = 0.07$.

Measures

The NIH-EXAMINER (memory.ucsf.edu/examiner) tasks are a combination of computer and paper-and-pen measures that tap working memory, inhibition, set-shifting, fluency and planning ability that generate working memory, cognitive control, fluency and global executive composite scores. There are three forms for each task of comparable difficulty to allow for repeated testing. Form A was used in this study. The computerised tasks within the NIH-EXAMINER were presented with PsychoPy software (Peirce, 2007) and latent factor scores were calculated using item response theory with the *ltm* package (Rizopoulos, 2006) in R (version 3.4.3; R Core Team, 2017) using the script provided (Kramer et al., 2013). The variables that contribute to each factor score and the individual tasks from which these variables are derived are described below.

Fluency Factor. A fluency factor score was derived, according to manual instructions (Kramer et al., 2013), by combining the total number of correct responses in the semantic and phonemic fluency tasks. In the *Phonemic Fluency task*, participants were asked to name as many words as they could that began with a given letter of the alphabet in one minute. Two trials were administered, with the letters L and F. Names of places, people or numbers were not permitted. In the *Semantic Fluency task,* participants were asked to name as many words as they could within a given category in one minute. Two trials were administered, with the categories 'animals' and 'vegetables'. For each task, the number of correct responses, rule violations and repetitions were totalled across both trials. More correct responses were indicative of greater phonemic/semantic fluency.

Cognitive Control Factor. Following Kramer et al. (2013), a cognitive control factor score was derived by combining the total shift score from the set-shifting task, the total flanker score from the flanker task, the total score from the anti-saccade condition of the saccade task and the total number of dysexecutive errors recorded by the examiner during administration of the battery. The *set-shifting task* was designed to measure cognitive flexibility. On each trial,

participants were presented with a target image in the centre of the screen that was either red or blue and either a rectangle or triangle. Participants were also presented with a condition cue (the word shape or the word colour) at the bottom of the screen, which indicated the dimension participants should use to match the target to one of the two comparison images presented in the bottom left and right of the screen, see Figure S1 (Supplementary Materials) for example of the stimuli. If required to match by colour, the participant was asked to press the left arrow key for red and the right arrow key for blue and, if matching by shape, they were asked to press the left arrow key for triangle and the right arrow key for rectangle. The participant was asked to respond as quickly as possible while avoiding errors. The target image was displayed in the centre-screen for five seconds or until the participant responded. Trials were organised into blocks where a single cue (colour or shape) was repeatedly presented or in blocks where a mixture of these cues were used, and participants had to adapt flexibly to changes in cues. Up to three sets of practice blocks, consisting of eight sequential colour trials and eight sequential shape trials, were used to ensure participants understood the task instructions. Participants advanced to the test block of test trials if at least 12 of the 16 trials in one practice block were answered correctly. The accuracy score was the proportion of correct responses in the shifting block (out of 64 trials), multiplied by 5 to create a range from 0 to 5. RT data are often positively skewed and, to adjust for this skew, a $\log_{10}$ transformation was applied to the median RT score. To further reduce skewing, the minimum RT was set to 500ms and the maximum reaction time was set to 3000ms; scores that fall outside that range were truncated (e.g., an RT of 4000ms was set equal to 3000ms and 300ms was set to 500ms). Log values were algebraically rescaled from a $\log_{500} - \log_{3000}$ range to a 0 to 5 range. The accuracy score (ranging from 0 to 5) and the adjusted RT score (ranging from 0 to 5) were summed to create the shift score. Possible scores ranged from 0 to 10; higher scores were indicative of better performance.

On the *flanker task*, participants were shown on the computer screen a row of five black arrows on a white background and asked to indicate by pressing the left or right arrow button

whether the centre, target arrow was pointing to the right or to the left, see Figure S2 (Supplementary Materials) for example stimuli. Participants were asked to respond as quickly as possible while avoiding errors. In the congruent condition, all of the non-target arrows were pointing in the same direction as the target arrow. In the incongruent condition, all of the non-target arrows were pointing in the opposite direction to the target arrow. There were 24 trials for each condition, yielding 48 trials in total. Between trials, a fixation point appeared for a random time interval between 1000ms and 3000ms and stimuli were sometimes presented above and sometimes below the fixation point. Up to three sets of practice trials were presented prior to the test trials. Participants advanced to the test trials if they obtained at least six correct responses in a practice block. A score combining accuracy and reaction time (RT) was calculated for the incongruent trials. The accuracy score was the proportion of correct responses (out of 24 trials), multiplied by 5 to create a score that ranged from 0 to 5. To correct for skew in RT data, the median RT was transformed in the same way as for the set-shifting task (see above) to generate a score from 0 to 5. The accuracy score (ranging from 0 to 5) and the adjusted RT score (ranging from 0 to 5) were summed to create the flanker score. Possible scores ranged from 0 to 10; higher scores reflected better performance.

In the *anti-saccade task*, which measured distractor inhibition, there were two distinct conditions. In the pro-saccade condition, participants were asked to fixate on a white circle, 0.3cm in width in centre-screen, which appeared for 1000ms. This was followed by a blank screen for 200ms and then the white circle was displayed parallel with the fixation point on the left or right extremes of the screen for 900ms, see Figure S3 (Supplementary Materials) for example stimuli. Participants were required to shift their eye gaze to the white circle's new location. They were then required to move their eye gaze back to centre-screen and wait for the next trial to begin. In the anti-saccade condition, the trial procedure was similar but, critically, participants were required to shift their eye gaze in the *opposite* direction to where the circle had moved. Participants completed one block of three practice trials followed by two blocks of 10

trials. Participants received one point each time their saccade was in the opposite direction to the circle on screen, with maximum scores of 10 and 20 in the pro- and anti-saccade conditions, respectively. The pro-saccade condition was administered first to establish a pre-potent response but the score on this condition was not evaluated. Higher scores on the anti-saccade condition were indicative of greater difficulty with distractor inhibition.

Working Memory Factor. A working memory factor score was calculated by combining the total dot-counting score and the d-prime (d') from the spatial 1- and 2-back tasks described below. The *dot-counting task* assessed verbal working memory and was based on the counting span task (Case, Kurland, & Goldberg, 1982). Participants were presented with a mixed array of green circles, blue circles and blue squares on a computer screen and asked to count and remember the total number of blue circles, see Figure S4 (Supplementary Materials) for example stimuli. Once they had repeated the total aloud, the examiner switched to a new screen with a novel array and participants were asked to count aloud and remember the number of blue circles on this new screen. After a set number of screens had been presented, participants were asked to recall the total number of blue circles they counted on each of the screens presented in the order in which they were presented. Participants were given one point for each correct total that a participant recalled in the correct location of a trial sequence. The task began with a practice block of three trials, followed by six test trials where the number of display screens increased by one in each successive trial, from two to seven. The number of correct responses was recorded and totalled across all six trials (maximum score = 27 points). Higher scores reflected better verbal working memory.

In the *spatial 1-back task*, which assessed spatial working memory, the participant was asked to remember the location of a 2.5cm x 2.5 cm white square (the probe) that appeared in one of 15 possible locations on a computer screen for 1000ms. This was followed by a 500ms delay when the screen was entirely blank. Participants were asked to read aloud a number (e.g., 8, that was approximately 1.5cm tall) that was presented in the centre of the screen for 1000ms.

Following a 500ms delay, another white square appeared (the target). The participant was asked to indicate whether this target square was in the same location as the probe by pressing the left arrow key for 'yes' and the right arrow key for 'no'. After the participant responded, another white square appeared which served as both a target that needed to be held in memory for the following trial and as a probe that is compared to a target displayed on the earlier trial. Participants completed up to two sets of practice blocks, of 10 probes each, see Figure S5 (Supplementary Materials) for an example of the stimuli. The participant advanced to the testing block if at least seven of the 10 trials in one practice block were answered correctly. The test block consisted of 30 probes. Based on signal detection theory, the preferred measure in *n*-back tasks such as this one is a net score that takes into account the relative proportion of hits and false alarms, called a *d* prime (*d'*; Swets, Tanner, & Birdsall, 1988). *d'* for the spatial 1-back task was calculated using the following formula: $d' = Z_{[hit\ rate]} - Z_{[false\ alarm\ rate]}$. Possible scores ranged from -1.11 to 3.67, where higher scores are indicative of better performance.

For a similar, albeit more difficult, *spatial 2-back task*, the probe was compared to the target from the trial before the preceding trial (i.e., the trial 2 back from the current trial). During the 2-back task, the stimulus displayed on the preceding trial was the "intermediate" target and so no number was displayed because there was not a need to draw participants fixation away from the probe. The timings were identical to the 1-back task. Similar to the 1-back task, participants completed up to three sets of practice blocks of 10 probes followed by a test block consisting of 90 probes, see Figure S6 (Supplementary Materials) for sample stimuli. A *d'* was also calculated for the 2-back task. Possible scores ranged from -1.94 to 3.88; higher scores were indicative of better spatial working memory performance.

Two additional tasks were administered that did not contribute to the factor scores generated, because task performance did not load clearly onto either the one- or three-factor models reported in Kramer et al. (2014).

In the *Continuous Performance Task (CPT),* participants were asked to press the left arrow key, as quickly as they could, every time the target image (a white five-pointed star, 3cm in diameter) was presented but was asked not to press any key if any of the other five non-target shapes were presented. Each trial began with the display of an image for a duration of 750ms. Once the image was displayed, participants responded by clicking the left arrow key. Participants completed up to three sets of practice blocks, consisting of 15 displays of the target image and one display of each non-target image (20 trials), see Figure S7 (Supplementary Materials) for example. Participants needed to respond correctly on at least 16 trials in one practice block to proceed. The testing block consisted of four sets of 25 trials (100 in total) with each set having 20 displays of the target image and one display of each non-target image. The total number of false alarm errors was recorded (maximum score = 20), which was reverse coded so that higher scores were indicative of greater ability with respect to sustained attention.

The *unstructured task* is conceptually similar to the Six Elements Task (Burgess, Alderman, Evans, Emslie, & Wilson, 1998) and designed to measure strategy generation, planning, and strategy execution. In a practice condition, participants were presented with one sheet of paper containing six puzzles (grey-scale images; see Figure S9 of Supplementary Materials for examples of puzzles). In the main task, participants were provided with three test booklets, each booklet contained six pages, with four puzzles on each page. In each booklet, the number of points that could be earned was identical but the number of points available on each successive page was reduced and so advancing through a booklet resulted in diminishing returns. The puzzles varied with respect to the points earned for completing them and the time they took to complete. Participants were asked to earn as many points as possible in six minutes. They were informed that they did not have to complete an entire page or an entire booklet before moving to a different puzzle and that they would only receive points if a puzzle were fully completed. A timer remained visible to participants throughout. Each puzzle was designated as high or low value based on its point value relative to the average time taken to complete it. The total number of

points earned for completing each puzzle, and the number of high value and low value puzzles completed in six minutes, were recorded. Following Kramer et al. (2013), a weighted composite was generated that combined the total points earned with the number of high and low value puzzles completed using the formula below. Higher scores were indicative of better planning ability.

$$\left(\left(\frac{completed\ high\ value\ items}{completed\ high\ value\ items\ +\ completed\ low\ value\ items}\right) \times 100\right) \times \log_{10}(total\ points\ earned + 1)$$

Scoring. In addition to individual scores on each task, an NIH-EXAMINER *global executive factor* score was calculated by combining each of the variables that contributed to the fluency factor, cognitive control factor and working memory factor scores. The NIH-EXMAINER also generates a composite measure of dysexecutive errors committed across the administration of the whole battery. Such errors captured executive-related difficulties, which can manifest as impulsive errors, failure to shift set, perseverative behaviour, and stimulus-boundedness, even when overall descriptive achievement scores on tests are unremarkable. Using the NIH-EXAMINER Behaviour Rating Scale, (Kramer et al., 2013), the experimenter rated their subjective perception of participants' behaviour according to manual instructions with respect to stimulus boundedness, social inappropriateness, perseverative responding and motivation. These ratings were combined with the number of repetitions and rule violations in each of the semantic and phonemic fluency tasks, the false alarm responses on the CPT task, the number of errors made on the incongruent trials relative to congruent trials of the flanker task, and the number of errors made on shift trials relative to non-shift trials in the set-shifting task.

General Procedure

Participants were seen individually in a quiet room at the university (*n*=71) or at their school (*n*=7). To begin, participants completed the NIH-EXAMINER battery (approximately one hour). In accordance with manual instructions, viewing distance was approximately 75 cm. Participants were then administered the WASI-II (Wechsler, 2011). Where necessary,

participants were seen on multiple occasions to complete testing (one visit: *n*=50, two visits: *n*=25, three visits: *n*=3). Parents of all participants completed a demographic questionnaire and the SRS-2 and returned them to researchers during their children's in-person sessions or via regular post shortly thereafter. The ADOS-2 was not administered specifically for this study but a subset (n=19) of our autistic participants had been assessed on this instrument within a year of their in-person sessions; these data are provided for descriptive purposes only (see Table 1), rather than used in analyses.

Ethics approval was granted through UCL Institute of Education. Parents of all participants, and young people themselves, gave written, informed consent prior to their participation in this study.

Results

Data screening

To begin, dependent variables from the individual tasks and the composite scores were checked for normality. Performance scores on the spatial 1-back, dot counting, and unstructured tasks were normally distributed, and all of the factor scores were normally distributed. Performance scores on the following tasks deviated from normality: phonemic fluency, semantic fluency, set-shifting, flanker, anti-saccade, dysexecutive errors, CPT the spatial 2-back. Tukey's Ladder of Powers was performed on each of the variables that were not normally distributed using the *rcompanion* package in R (Mangiafico, 2016) and were successfully transformed to normal, with the following exceptions, where untransformed values were used in analyses: spatial 2-back, the anti-saccade and the CPT. Given that analysis of variance is robust to violations of assumptions of normality, we proceeded with conducting ANCOVAs to analyse these data (see below). As stated previously, to enable comparison across tasks, scores from the CPT were recoded so that a high score reflected good performance. We then converted scores on each EF task to z scores. All subsequent analyses were performed using the z scores for each variable,

although for ease of interpretation the untransformed means and standard deviations for each variable can be found in Table 3.

Three participants failed to progress beyond the practice condition in the spatial 2-back (n=1; non-autistic) and in the flanker tasks (n=2; both autistic).

Analysis plan

To test our hypothesis that autistic participants would differ significantly from non-autistic participants on each of the factor scores from the NIH-EXAMINER (global executive, fluency, cognitive control, and working memory factors), separate analyses of covariance (ANCOVA) were conducted, with 'group' (autistic, non-autistic) as the between-participants factor and, following Schreiber et al. (2014), chronological age as the covariate. We ran additional separate ANCOVAs on the unstructured and CPT task scores as these were not used in the creation of factor scores.

To test further the relationships between performance on the NIH-EXAMINER variables and other, non-executive variables, we conducted a series of correlational analyses. All analyses were conducted in R (version 3.4.3; R Core Team, 2017) and RStudio (version 1.1.4; RStudio Team, 2015).

Group differences

Tables 3 and 4 show descriptive statistics for the dependent variables from individual tasks and generated factor scores. Figure 1 shows performance on factor scores (z-scores of raw scores are used here, to facilitate plots on a single figure).

After adjustment for age, there was a significant group difference on the NIH-EXAMINER global executive composite score, $F(1,75)=4.90$, $p=.030$, with a generalised eta squared (η^2_G) of .061; autistic adolescents scored lower, on average, than the non-autistic adolescents. There were no significant group differences on any other factor, once variation in age had been adjusted for (fluency factor: $F(1,75)=1.56$, p=.216, $\eta^2_G=.02$; working memory factor: $F(1,75)=1.95$, p=.166, $\eta^2_G=.03$; cognitive control factor: $F(1,75)=2.80$, $p=.098$, $\eta^2_G=.04$).

With respect to the two tasks that did not contribute toward factor scores, we found a significant group difference on unstructured task scores, $F(1,75)=6.41$, $p=.013$, $\eta^2_G=.08$, but not on CPT scores, $F(1,75)=0.18$, $p=.676$, $\eta^2_G=.002$, after adjustment for age. Autistic adolescents performed worse, on average, on the unstructured task compared to non-autistic adolescents.

Individual differences

The second aim of this study was to test whether there were relationships between performance on the NIH-EXAMINER and on other, non-executive variables. Figure 2 shows the relationships between the global executive factor and verbal IQ, performance IQ, chronological age and autistic features (as indexed by the SRS-2). Age was associated with the global executive factor for non-autistic participants, $r(40)=.48$, $p=.001$. An r-to-z transformation showed that the magnitude of the relationship between the autistic and non-autistic groups was not significantly different, $z=-0.05$, $p=.960$. For the autistic group only, there was also a significant association between the global executive factor and verbal IQ, $r(34)=.49$, $p=.003$, and non-verbal IQ, $r(34)=.54$, $p<.001$. None of the other tested correlations were significant (all ps$>.440$).

Discussion

To our knowledge, this was the first study to compare autistic and non-autistic participants on the NIH-EXAMINER battery. We showed that autistic adolescents have global EF difficulties of moderate effect compared to their non-autistic counterparts, suggesting that the NIH-EXAMINER battery was sensitive to the global EF difficulties previously reported among autistic participants (Demetriou et al., 2018; Hill, 2004a; Lai et al., 2017). Future investigations can therefore use this open-source and completely standardised battery to scale up the assessment of the EF of autistic people to include large, diverse samples, to conduct like-for-like replications and to re-assess the accuracy of the conclusions previously drawn in the literature in an iterative manner, as more data are collected, shared and compiled.

Notwithstanding, there are other aspects of our findings that warrant significant caution in the use of this measure and call into question some of the implicit arguments underpinning it. Although, as expected, autistic participants showed more difficulties than non-autistic participants on the global composite EF score, we found no significant group differences on any of the factor scores representing individual subcomponents of EF (fluency, cognitive control and working memory). As such, the current findings provide support neither for the claim that there is an autism-specific profile of executive difficulties, with challenges with flexibility and planning in particular (Hill, 2004b; Ozonoff & Jensen, 1999), nor the suggestion that autistic people experience executive difficulties that apply evenly across individual domains and have an overall, rather than a fractionated, profile of EF difficulties (Demetriou et al., 2018; Lai et al., 2017).

There are several possible explanations for our failure to support these hypotheses. First, the tasks grouped by EF subcomponent might not be reliably tapping the same construct. For example, one relevant claim is that *n*-back tasks are not a reliable metric of spatial working memory (Kane, Conway, Miura, & Colflesh, 2007), which, if true, means that the shared variance between the dot-counting task and the *n*-back task might not reflect working memory ability *per se* and might also reflect non-executive sources of shared variance, or simply measurement error. The non-executive features that differ between tasks, such as cognitive load, language demands, and task difficulty do not only impact EF performance but can also impact a participant's motivation and affective response to that task, exacerbating this issue (Poldrack & Yarkoni, 2016). Deriving factor scores across tasks where task performance is differentially sensitive to the underlying construct of interest might introduce measurement error into the factor scores rather than counteracting the task impurity problem they were derived to address.

Second, it is possible that the executive difficulties reported upon in the literature vary at different stages of development, especially during adolescence. Previous reports have showed that executive difficulties abate with age (Geurts et al., 2014) or – consistent with the current

results – disappear altogether during adolescence for measures of cognitive control and working memory (Demetriou et al., 2018; see also, Uddin, 2021, for discussion). This might be because adolescents experience a relative dip in EF ability due to the neural re-organisation that takes place with the onset of puberty (Blakemore & Choudhury, 2006). This means that there might not be a specific abatement of EF difficulties in young autistic people but that it might be more difficult to detect differences between autistic and non-autistic participants during this period. It is also possible that there is a genuine abatement such that young autistic people learn the executive skills with which they struggled during childhood by the time they reach adolescence, possibly by harnessing their more general cognitive abilities (Livingston & Happé, 2017).

Third, autistic people might struggle with unstructured tasks rather than EF in general. According to White et al. (2009), an EF task is unstructured when it has been explicitly designed such that there are a number of ways of doing the task or is characterised explicitly to be open-ended as acknowledged clearly by the task's creator. Three tasks from the NIH-EXAMINER battery fit this definition, including the phonemic and semantic fluency tasks and the unstructured task. While performance did not differ between our autistic and non-autistic groups on the phonemic fluency task, our autistic participants did have greater difficulty than non-autistic participants on the semantic fluency and the unstructured tasks. The fact that young autistic people struggled on two of the three less-structured tasks within the NIH-EXAMINER battery provides some support for the notion that tasks involving open-ended responses (White, 2013; White et al., 2009) and multifactorial tasks that tap planning and multitasking abilities (Hill & Bird, 2006; Mackinlay, Charman, & Karmiloff-Smith, 2006) are especially difficult for autistic participants. The autistic participants in the current study had particular difficulties on the tasks that required them to generate a strategy, to continuously appraise and modify their behaviour and to sustain their attention without external support from the experimenter. It remains possible that the majority of the tasks within the NIH-EXAMINER battery (and similar neuropsychological assessments), all conducted within a necessarily-controlled laboratory

environment, provide excessive scaffolding – that is, too much cueing, attention orienting, and task monitoring – and therefore do not map on to the kinds of executive demands placed on autistic people beyond the laboratory, which appear to be captured better by questionnaire measures, like the Behavior Rating Inventory of Executive Function – second edition (BRIEF-2; Gioia, Isquith, Guy, & Kenworthy, 2015).

This third potential explanation could further lead to a more significant challenge. Some authors have recently suggested that there is a fundamental discrepancy between EF as conventionally understood and measured by performance on objective, lab-based measures and the everyday executive difficulties experienced subjectively in everyday life (e.g., Dang, King, & Inzlicht, 2020). It may be, that is, that they are tapping different constructs. As discussed by Toplak et al. (2013), performance-based measures of EF are administered within 'optimal or maximal performance situations' (similar to intelligence tests), in which task instructions and interpretations are highly constrained, and participants are supervised by an external examiner and instructed 'to do their best'. Instead, questionnaire-based measures ask participants to reflect on their usual, day-to-day performance in situations which are unsupervised and less constrained. Accordingly, each type of measure provides useful, yet conceptually distinct information: performance-based measures might be beneficial for understanding underlying cognitive mechanisms, while questionnaire-based measures might have more predictive utility in clinical and educational settings (see also Dang et al., 2020). It is equally possible, though, that performance-based EF measures are unlikely to be informative (at a mechanistic level of explanation) because they lack representativeness (the correspondence between the task and real-life settings) and generalisability (the degree to which task performance predicts problems in real-life settings) (see Burgess et al., 2006).

One way to address this issue in future is to focus our efforts on designing carefully-controlled measures to elucidate the mechanisms underlying the sorts of difficulties people have

in their day-to-day lives (e.g., Cribb, Kenny, & Pellicano, 2019) and to contrast those with pre-existing theoretical approaches to EF and to the measures usually deployed to measure it.

Limitations

This study is not without its limitations. First, one of the advantages of using a fully manualised EF battery is that the decisions for calculating the key dependent variables representing for individual tasks and overall EF ability were predetermined and, as such, were impervious to the large number of researcher degrees of freedom that could have been employed to impact the likelihood of producing group differences. That said, it remains unclear whether such decisions outlined within the manual (e.g., summing RTs and accuracy scores) are optimal for examining group differences when comparing autistic and non-autistic participants. Future exploratory research should therefore assess the sensitivity of each of the possible variables that could be generated from these tasks when comparing autistic and non-autistic participants. Second, it is possible that group differences exist on the various tasks and factors, but the effect sizes are smaller than many studies have previously reported. Indeed, while the latent factors were not significantly different, the effect size estimates were small, suggesting that our study may have lacked sufficient statistical power to detect significant group differences of small statistical magnitude that may still have clinical significance. We encourage researchers to replicate the use of the NIH-EXAMINER Battery and to deposit the data collected into a public repository, which would allow other researchers to build a cumulative dataset, that would over time allow the assessment of samples that are increasingly large and diverse. This would allow more definitive analyses regarding whether group differences of the magnitude found in this paper are reproducible and validation against other measures would allow us to determine if effect sizes of this magnitude are clinically meaningful.

Conclusion

This study sought to assess autistic and non-autistic adolescents on a wide-ranging battery of EF tasks that are typical of systematically designed laboratory investigations of EF in

autism. Critically, the NIH-EXAMINER battery involves assessing participants on multiple tasks from each executive subdomain so that latent variables could be generated to represent EF ability, reducing the measurement error from non-executive factors often implicated in the literature. Overall, we found that autistic adolescents have moderate difficulties with EF ability in general. Future work will need to compare autistic and non-autistic adolescents on more open-ended, less-structured tasks – ones that are executively demanding but more closely resemble the demands of everyday life.

References

APA. (2000). *Diagnostic and statistical manual, text revision* (4th ed.). Washington: American Psychiatric Association.

APA. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington: American Psychiatric Association.

Blakemore, S.-J., & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry*, *47*(3-4), 296–312.

Bölte, S., Duketis, E., Poustka, F., & Holtmann, M. (2011). Sex differences in cognitive domains and their clinical correlates in higher-functioning autism spectrum disorders. *Autism*, *15*(4), 497–511.

Burgess, P. W., Alderman, N., Evans, J. J., Emslie, H., & Wilson, B. A. (1998). The ecological validity of tests of executive function. *Journal of the International Neuropsychological Society*, *4*(6), 547–558.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386–404.

Constantino, J. N., & Gruber, C. P. (2012). *Social responsiveness scale, second edition (SRS-2)*. Torrance, CA: Western Psychological Services.

Crawford, S., & Channon, S. (2002). Dissociation between performance on abstract tests of executive function and problem solving in real-life-type situations in normal aging. *Aging and Mental Health*, *6*(1), 12–21.

Cribb, S., Kenny, L., & Pellicano, E. (2019). 'I definitely feel more in control of my life': the perspectives of young autistic people and their parents on emerging adulthood. *Autism*, 1362361319830029.

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269.

Demetriou, E. A., Lampit, A., Quintana, D. S., Naismith, S. L., Song, Y. J. C., Pye, J. E., … Guastella, A. J. (2018). Autism spectrum disorders: a meta-analysis of executive function. *Molecular Psychiatry*, *23*(5), 1198–1204.

Gernsbacher, M. A. (2017). Editorial perspective: the use of person-first language in scholarly writing may accentuate stigma. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *58*(7), 859–861.

Geurts, H. M., van den Bergh, S. F. W. M., & Ruzzano, L. (2014). Prepotent response inhibition and interference control in autism spectrum disorders: two meta-analyses. *Autism Research*, *7*(4), 407–420.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. E. (2015). *Behavior Rating Inventory of Executive Function; Professional Manual, Second edition* (2nd ed.). Florida: PAR.

Hill, E. L. (2004a). Evaluating the theory of executive dysfunction in autism. *Developmental Review*, *24*(2), 189–233.

Hill, E. L. (2004b). Executive dysfunction in autism. *Trends in Cognitive Sciences*, *8*(1), 26–32.

Hill, E. L., & Bird, C. M. (2006). Executive processes in asperger syndrome: patterns of performance in a multiple case series. *Neuropsychologia*, *44*(14), 2822–2835.

Hughes, C., Russell, J., & Robbins, T. W. (1994). Evidence for executive dysfunction in autism. *Neuropsychologia*, *32*(4), 477–492.

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: a review of our current understanding. *Neuropsychology Review*.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622.

Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms

should be used to describe autism? perspectives from the uk autism community. *Autism*, *20*(4), 442–462.

Kramer, J. H., Mungas, D. M., Possin, K. L., Rankin, K. P., Boxer, A. L., Rosen, H. J., … Widmeyer, M. (2013). *Executive Abilities: Measures and Instruments for Neurobehavioral Evaluation and Research (EXAMINER), User Manual, version 3.6*. NIH-EXAMINER team.

Kramer, J. H., Mungas, D. M., Possin, K. L., Rankin, K. P., Boxer, A. L., Rosen, H. J., … Widmeyer, M. (2014). NIH examiner: conceptualization and development of an executive function battery. *Journal of the International Neuropsychological Society*, *20*(1), 11–9.

Lai, E. C. L., Lau, Z., Lui, S. S. Y., Lok, E., Tam, V., Chan, Q., … Cheung, E. F. C. (2017). Meta-analysis of neuropsychological measures of executive functioning in children and adolescents with high-functioning autism spectrum disorder. *Autism Research*, *10*(5), 911–939.

Landry, O., & Al-Taie, S. (2016). A meta-analysis of the wisconsin card sort task in autism. *Journal of Autism and Developmental Disorders*, *46*(4), 1220–1235.

Liberatos, P., Link, B. G., & Kelsey, J. L. (1987). The measurement of social class in epidemiology. *Epidemiologic Reviews*, *10*(1), 87–121.

Livingston, L. A., & Happé, F. G. E. (2017). Conceptualising compensation in neurodevelopmental disorders: reflections from autism spectrum disorder. *Neuroscience and Biobehavioral Reviews Biobehavioral Reviews*.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule – 2nd edition (ADOS-2)*. Los Angeles, CA: Western Psychological Corporation.

Mackinlay, R., Charman, T., & Karmiloff-Smith, A. (2006). High functioning children with autism spectrum disorder: a novel test of multitasking. *Brain and Cognition*, *61*(1), 14–24.

Mangiafico, S. (2016). *Summary and analysis of extension program evaluation in R* (Vol. 442). New Brunswick, NJ: Rutgers.

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.

Ozonoff, S. J., & Jensen, J. (1999). Brief report: specific executive function profiles in three neurodevelopmental disorders. *Journal of Autism and Developmental Disorders*, *29*(2), 171–177.

Peirce, J. W. (2007). PsychoPy - psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Pellicano, E. (2010). Individual differences in executive function and central coherence predict developmental changes in theory of mind in autism. *Developmental Psychology*, *46*(2), 530–544.

Poldrack, R. A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annual Review of Psychology*, *67*(1), 587–612.

R Core Team. (2017). R: a language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rabbitt, P. (2004). Do "frontal tests" measure executive function? issues of assessment and evidence from fluency tests. In P. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 193–215). London: Routledge.

Rizopoulos, D. (2006). Ltm: an r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*.

Robinson, H., Calamia, M., Gläscher, J., Bruss, J., & Tranel, D. (2014). Neuroanatomical correlates of executive functions: a neuropsychological approach using the examiner battery. *Journal of the International Neuropsychological Society*, *20*(01), 52–63.

RStudio Team. (2015). RStudio: integrated development for r. Boston, MA: RStudio, Inc.

Russell, J. (1997). *Autism as an executive disorder*. Oxford University Press.

Ruxton, G. D., & Neuha, M. (2010). When should we use one-tailed hypothesis testing ?, 114–117.

Schatz, J., Stancil, M., Katz, T., & Sanchez, C. E. (2014). EXAMINER executive function battery and neurologic morbidity in pediatric sickle cell disease. *Journal of the International Neuropsychological Society*, *20*(1), 29–40.

Schmitz, N., Rubia, K., Daly, E. M., Smith, A., Williams, S. C. R., & Murphy, D. G. M. (2006). Neural correlates of executive function in autistic spectrum disorders. *Biological Psychiatry*, *59*(1), 7–16.

Schreiber, J. E., Possin, K. L., Girard, J. M., & Rey-Casserly, C. (2014). Executive function in children with attention deficit/hyperactivity disorder: the nih examiner battery. *Journal of the International Neuropsychological Society*, *20*(1), 41–51.

Sinclair, J. (1999). Why i dislike 'person-first' language. Retrieved October 30, 2014, from http://web.archive.org/ web/20090210190652/http://web.syr.edu/~jisincla/person_first.htm (accessed

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*, 328.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1988). Decision processes in perception. *Adolescent Gynecology*, *84*(1967), 1932–1935.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *54*(2), 131–143.

Uddin, L. Q. (2021). Brain mechanisms supporting flexible cognition and behavior in adolescents with autism spectrum disorder. *Biological Psychiatry*, *89*(2), 172–183.

Wang, Y., Zhang, Y., Liu, L., Cui, J., Wang, J., Shum, D. H. K., … Chan, R. C. K. (2017). A meta-analysis of working memory impairments in autism spectrum disorders. *Neuropsychology Review*, *27*(1), 46–61.

Wechsler, D. (2011). *WASI-II: Wechsler abbreviated scale of intelligence--*. Psychological Corporation.

White, S. J. (2013). The triple i hypothesis: taking another('s) perspective on executive dysfunction in autism. *Journal of Autism and Developmental Disorders*, *43*(1), 114–121.

White, S. J., Burgess, P. W., & Hill, E. L. (2009). Impairments on "open-ended" executive function tests in autism. *Autism Research*, *2*(3), 138–147.

WHO. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines* (Vol. 1). Geneva: World Health Organization.

Table 1.

Descriptive statistics for the developmental and background variables of participants assessed on the NIH-EXAMINER battery for autistic (n=36) and non-autistic (n=42) groups

	Non-autistic M (SD) Range	Autistic M (SD) Range	p-value
Developmental variables			
Gender (M:F)	20:22	24:12	.144
Age (in years)	14.91 (2.25)	14.55 (2.03)	.465
	12 – 18	11 – 19	
Full-scale IQ	103.74 (12.39)	103.56 14.72	.953
	77 – 130	76 – 132	
Verbal IQ	102.98 (13.58)	102.14 (12.58)	.780
	78 – 132	73 – 123	
Nonverbal IQ	103.14 (13.31)	104.44 (19.68)	.730
	71 – 132	75 – 154	
SRS-2	47.96 (5.93)[a]	77.61 (9.45)[b]	< .001
	38 – 59	56 – 90[c]	
ADOS-2 severity score[d]	- -	5.83 (2.64)[e]	-
	-	2 – 10[f]	

Note. Full-scale IQ = Full-scale IQ, 4 subtest version, derived from the Wechsler Abbreviated Scale of Intelligence, second edition (WASI-II; Wechsler, 2011); Verbal IQ = Verbal comprehension index derived from the WASI-II (Wechsler, 2011); Nonverbal IQ = Perceptual reasoning index derived from the WASI-II (Wechsler, 2011), all M=100, SD=15; SRS-2 = Social Responsiveness Scale, Second edition, *t*-scores were calculated separately by gender (Constantino & Gruber, 2012); ADOS-2 = the Autism Diagnostic Observation Schedule, Second edition (Lord et al., 2012). [a]*n*=32. [b] *n*=35. [c]One participant fell below the threshold indicative of an autism diagnosis (*t*-score of 60) but were retained in the analysis because they received a score above the threshold for autism on the ADOS-2. [d]Calibrated severity scores are reported here (maximum score = 10). [e] The ADOS-2 was completed on a subset of participants, *n*=19. [f]Two autistic participants obtained ADOS-2 severity scores below the threshold indicative of an autism diagnosis (a severity score of 2 or below) but were retained in the analysis because they received a score above the threshold indicative of an autism diagnosis on the SRS-2.

Table 1.

Demographic information about the participants assessed on the NIH-EXAMINER battery.

	Non-autistic (*n*=42)	Autistic (*n*=36)
	n	*n*
Parent-reported co-occurring conditions		
Attention deficit hyperactivity disorder	0	9
Dyslexia	0	10
Developmental coordination disorder	0	7
Sensory processing disorder	0	2
Obsessive compulsive disorder	0	1
Complex language disorder	0	0
Depression	1	0
Medication		
ADHD	0	5
Antipsychotic	0	1
Sleep inducing	0	4
Antiepileptic	0	1
Antidepressants	1	3
Non-psychoactive	2	2
Ethnicity		
Any White background	13	26
Any Asian background	3	3
Any Black background	9	1
Any mixed background	4	4
Other ethnic group	1	1
Missing or prefer not to say	11	1

Note. Any White background = White British, White Irish or any other White background; Any Black background = Black British, Black African, Black Caribbean or any other Black background; any Asian background = Chinese, Indian, Pakistani, Bangladeshi or any other Asian background; Mixed/multiple ethnic groups = Mixed White and Asian, Mixed White and Black African, Mixed White & Black Caribbean, Any other Mixed background.

Table 3.

Descriptive statistics for the dependent variables specified for each task and factor in the NIH-EXAMINER battery

	Non-autistic (n=42)		Autistic (n=36)	
	M (SD)		M (SD)	
	Range		Range	
Fluency				
Phonemic fluency task	22.31 (7.75)		20.17 (6.33)	
	8 – 48		9 – 36	
Semantic fluency task	30.40 (8.98)		27.19 (7.69)	
	16 – 61		8 – 46	
Fluency Factor	0.20 (0.69)		-0.01 0.60	
	-1.31 – 2.34		-0.95 – 1.37	
Cognitive control				
Set-shifting task	8.18 (0.85)		7.76 (0.92)	
	5.42 – 9.39		5.83 – 9.62	
Flanker task	8.96 (0.53)		8.78 (0.59)[a]	
	7.13 – 9.78		7.17 – 9.85	
Anti-saccade task	31.43 (6.27)		30.47 (5.22)	
	15 – 40		19 – 40	
Cognitive Control Factor	0.73 (0.63)		0.48 (0.61)	
	-0.99 – 2.05		-0.54 – 1.89	
Working memory				
Dot counting task	18 (3.97)		14.75 (4.45)	
	10 – 26		2 – 22	
Spatial 1-back task	2.28 (0.56)		2.13 (0.83)	
	1.02 – 3.67		0.14 – 3.67	
Spatial 2-back task	1.03 (0.74)[b]		0.95 (0.69)	
	-0.21 – 3.00		-0.23 – 2.79	
Working Memory Factor	0.30 (0.61)		0.08 (0.68)	
	-0.73 – 1.91		-1.59 – 1.82	
Variables not contributing to factor scores				
Continuous performance task[c]	16.19 (3.89)		16.44 (3.83)	
	3 – 20		5 – 20	
Unstructured task	162.55 (29.80)		142.75 (37.70)	
	91.63 – 234.47		56.44 – 210.98	
Dysexecutive errors	10.39 (6.43)		11.89 (6.84)	
	1 – 30		3 – 30	
Global Executive Factor	0.43 (0.54)		0.15 (0.50)	
	-0.66 – 2.09		-0.76 – 1.08	

Notes. Phonemic fluency = total correct responses across both trials of the task, Semantic fluency = total correct responses across both trials of the task, Set-shifting = total weighted accuracy and reaction time score, Anti-saccade = total number of trials in the anti-saccade condition, Dot counting = total correctly recalled screens, Spatial 1-back = d' on the 1-back condition of the *n*-back task. Spatial 2-back = d' on the 2-back condition of the *n*-back task. Continuous performance = the total number of false alarm errors recorded. Unstructured task = Weighted composite score from the unstructured planning task which combined points earned with the relative number of the high-value and low-value puzzles completed. Dysexecutive errors = combined score of errors committed across the administration of the NIH-EXAMINER battery, [a] *n*=35. [b] *n*=41. [c] Scores on the CPT were reverse coded, such that higher scores reflect better performance.
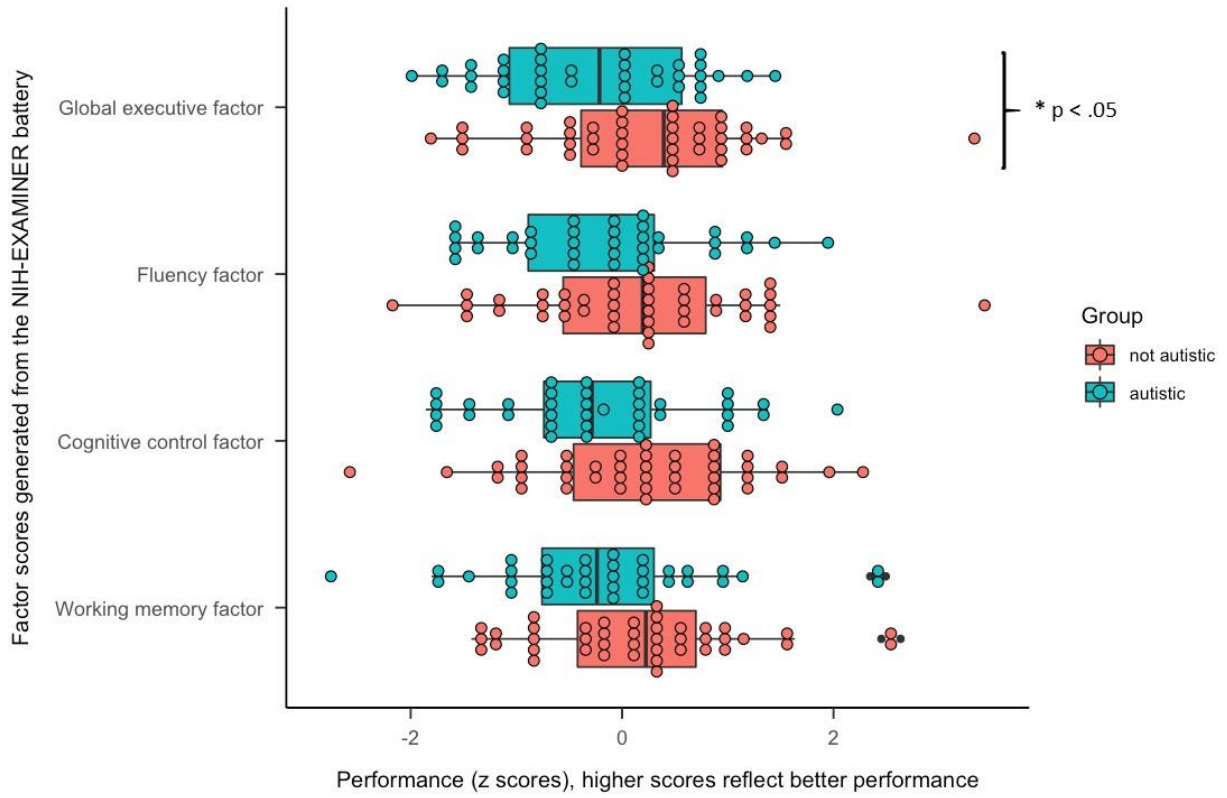
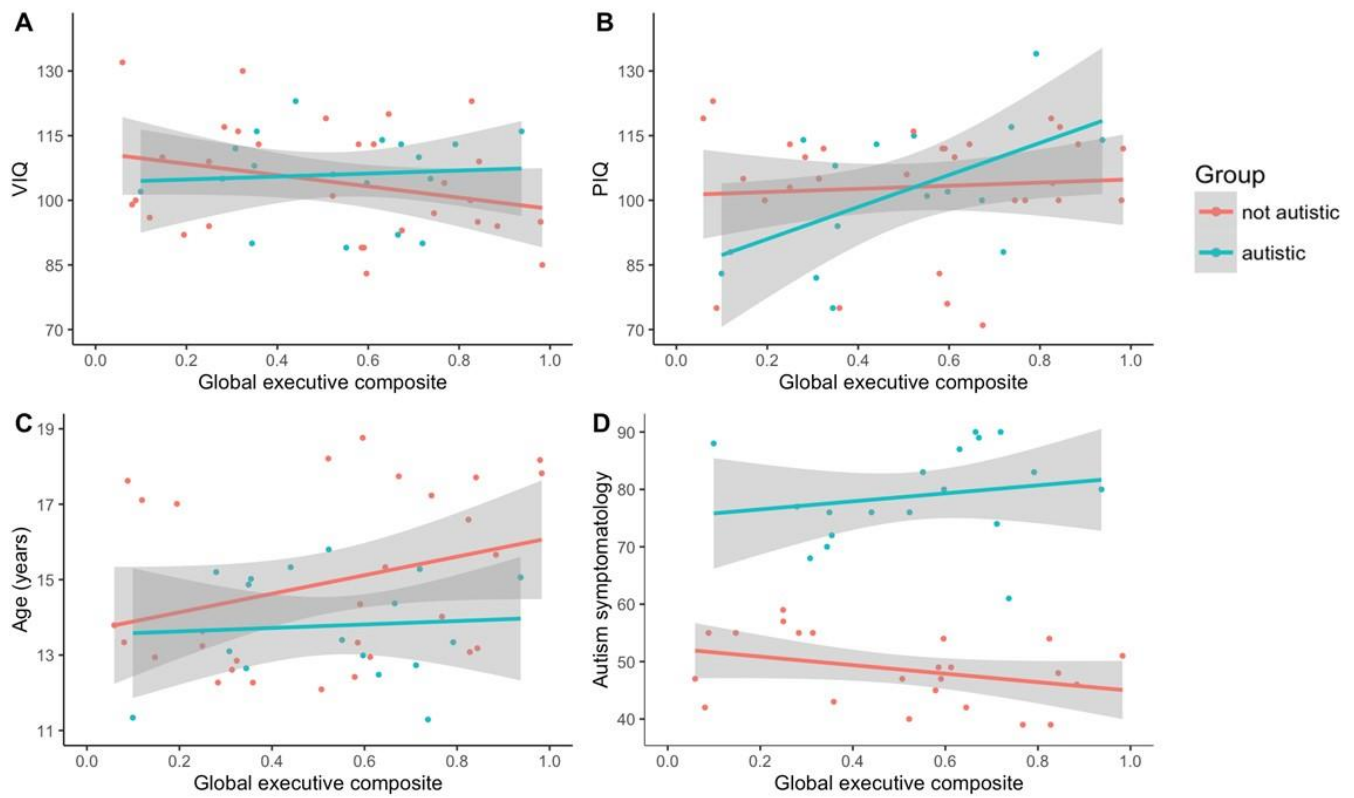*Figure 1*. Performance on the factor scores generated by the NIH-EXAMINER battery, by group.

Figure 2. Relationship between performance on the global executive composite score from the NIH-EXAMINER battery and (A) VIQ = verbal comprehension index on the Wechsler Abbreviated Scale of Intelligence, Second edition (WASI-II; Wechsler, 2011), (B) PIQ = perceptual reasoning index on the WASI-II, (C) age (in years) and (D) autistic features, indexed by t-scores from the Social Responsiveness Scale, Second edition (Constantino & Gruber, 2012).