Published in final edited form as:

Clin Neurophysiol. 2020 February ; 131(2): 406–413. doi:10.1016/j.clinph.2019.09.024.

Test-Retest Reliability of the N2 Event-Related Potential in School-Aged Children with Autism Spectrum Disorder (ASD)

Cremone-Caira Amanda^{a,b}, Vaidyanathan Akshita^c, Hyatt Danielle^d, Gilbert Rachel^e, Clarkson Tessa^f, Faja Susan^{a,b}

^aDivision of Developmental Medicine, Boston Children's Hospital, 02215

^bHarvard Medical School, 02215

^cThe Banyan, 600037

^dCollege of Medicine, University of Illinois at Chicago, 60612

^eCollege of Medicine, University of Florida, 32610

^fDepartment of Psychology, Temple University, 19122

Abstract

Objective: The N2 ERP component is used as a biomeasure of executive function in children with autism spectrum disorder (ASD). The aim of the current study was to evaluate the test-retest reliability of N2 amplitude in this population.

Methods: ERPs were recorded from 7 to 11-year-old children with ASD during Flanker ($n = 21$) and Go/Nogo tasks ($n = 14$) administered at two time points separated by approximately three months. Reliability of the N2 component was examined using intraclass correlation coefficients (ICCs).

Results: Reliability for mean N2 amplitude obtained during the Flanker task was moderate (congruent: ICC = 0.542, 95% CI [0.173,0.782]; incongruent: ICC = 0.629, 95% CI [0.276,0.831]). Similarly, reliability for the Go/Nogo task ranged from moderate to good ('go': ICC = 0.817, 95% CI [0.535,0.937]; 'nogo': ICC = 0.578, 95% CI [0.075,0.843]).

Conclusions: These findings support the use of N2 amplitude as a biomeasure of executive function in school-aged children with ASD.

Significance: This research addresses a critical gap in clinical neurophysiology, as an understanding of the stability and reliability of the N2 component is needed in order to

Corresponding Author: Susan Faja, PhD, Laboratories of Cognitive Neuroscience, Division of Developmental Medicine, Boston Children's Hospital, Harvard Medical School, 1 Autumn Street, Boston, MA 02215, U.S.A. Phone: (617) 919 – 4486, susan.faja@childrens.harvard.edu.

Disclosure: All authors have contributed to, and approved, the final article.

Conflict of Interest Statement: None of the authors have potential conflicts of interest to be disclosed.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

differentiate variance explained by repeated measurement versus targeted treatments and interventions.

Keywords

autism; executive function; N2; ERP; test-retest; reliability

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition that affects 1 in 59 children in the United States (Baio et al. 2018). Symptoms of ASD include social and communication difficulties, repetitive and restrictive behaviors, and sensory sensitivities (American Psychiatric Association 2013). Alongside hallmark symptoms of the disorder, many individuals with ASD have notable deficits in executive function (Kenworthy et al. 2008; Demetriou et al. 2017) – a set of higher order cognitive processes that rely upon top-down control over actions and behaviors (Diamond 2014; Friedman and Miyake 2017).

The presentation of symptoms and executive function deficits varies widely between individuals with ASD. Furthermore, problems with executive function may exacerbate ASD symptoms (Russell 1997; Geurts et al. 2014) which may contribute to variability of symptom presentation. Taken together, the widespread heterogeneity in ASD is a significant hindrance to the development of effective treatments. To address this issue, many researchers have begun to search for biomesasures associated with specific impairments in ASD. Reliable biomesasures can be assessed over time to evaluate experimental interventions and treatment response (Casanova et al. 2012; Dawson et al. 2012; Robb et al. 2016). Many groups have begun to assess event-related potentials (ERPs) as candidate biomesasures given the feasibility of acquiring them from individuals with ASD (Siper et al. 2016; Loth et al. 2017; Sokhadze et al. 2017; Foss-Feig et al. 2018).

ERPs are time-locked measures of brain activity used to characterize neural processes that underlie specific cognitive and behavioral responses. This methodology has significant utility in ASD research as ERPs can be obtained from individuals with sensory sensitivities, minimal language, and limited motor skill (Jeste and Nelson 2009; Webb et al. 2015). ERPs are associated with a range of higher order functions that mark impairments in cognitive outcomes (Banaschewski and Brandeis 2007; Jeste and Nelson 2009). The N2 ERP component is well characterized as a neural underpinning of executive function among typical developing children (Espineta et al. 2012, 2013; Brydges et al. 2014) and children with ASD (Pfefferbaum et al. 1985; Jodo and Kayama 1992; Faja et al. 2016). These studies and others (Larson et al. 2012; Kim et al. 2018) demonstrate increased interest in the N2 component in ASD research.

The N2 is a frontal, negative fluctuation observed approximately 300 ms after stimulus onset. Larger (more negative) N2 amplitudes are indicative of more effortful neural processing in the context of conflicting information. For example, N2 amplitudes were larger in response to incongruent trials in the Eriksen Flanker task, during which there is a mismatch between the target stimulus and the distractor stimuli on either side of the target (Abundis-Gutiérrez et al. 2014). In typically developing children, larger N2 amplitude on

incongruent trials was associated with less efficient executive function (Abundis-Gutiérrez, Checa, Castellanos, & Rosario Rueda, 2014; Buss, Dennis, Brooker, & Sippel, 2011). Similarly, larger N2 amplitude on 'nogo' trials of the Go/Nogo task was associated with reduced monitoring and inhibition (Pfefferbaum et al. 1985; Jodo and Kayama 1992). Among children with ASD, those who had larger differentiation in N2 amplitude between congruent and incongruent trials had greater difficulty on behavioral tasks assessing executive function (Faja et al. 2016).

Differences in N2 amplitude have also been reported between children with ASD and comparison groups: N2 amplitude during the Flanker task was larger in children with ASD, relative to typically developing children across trial types (Faja et al. 2016; but see Samyn et al. 2014). In contrast to findings with the Flanker task, N2 amplitudes of children with ASD were similar to (Hoyniak 2017) or smaller than (Tye et al. 2014) children without ASD across trial types of a continuous performance version of the Go/Nogo task.

1.1 Current Study

Provided evidence of group differences in N2 amplitude (Tye et al. 2014; Faja et al. 2016), and associations between the N2 and executive function (Buss et al. 2011; Faja et al. 2016), this ERP component represents a potential biomeasure in ASD research. However, a critical component of a biomeasure is its reliability over time. N2 amplitude has moderate to excellent test-retest reliability in typically developing adults (ICCs between 0.48 and 0.55; Clayson and Larson 2013) and 10 to 15-year-old children with attention-deficit/hyperactivity disorder (ADHD; ICCs between 0.7 and 0.9; Kompatsiari et al., 2016). Yet, to our knowledge, test-retest reliability of N2 amplitude has not been evaluated among individuals with ASD.

As such, the aim of this study was to evaluate the test-retest reliability of the N2 ERP component in school-aged children with ASD. As part of a larger clinical trial, ERP responses were collected as children, 7 to 11 years of age, completed a Flanker task to elicit the N2 component. ERP responses to a Go/Nogo task were also available from a subsample of these children and were evaluated separately to support findings derived from the Flanker task.

Retest reliability was measured via intraclass correlation coefficients (ICCs) for the mean N2 amplitude obtained from each of these tasks during two time points, separated by approximately three months. N2 amplitudes are characteristically larger for incongruent versus congruent trials. As such, collapsing across trial types may reduce the robustness of condition specific effects (i.e., N2 to incongruent trials; Thigpen et al. 2018). N2 amplitudes for congruent and incongruent stimuli are also typically highly correlated, presenting both conceptual and practical issues with subtraction-based difference scores (Meyer et al. 2017). Moreover, previous research suggests that both children with ASD (Faja, Clarkson, & Webb, 2016) and typically developing children (Abundis-Gutiérrez et al., 2014; Hoyniak, 2017; Ladouceur, Dahl, & Carter, 2007; Rueda, Posner, Rothbart, & Davis-Stober, 2004) do not consistently show differentiation of the N2 component across trial types. To avoid these potential confounds, we evaluated retest reliability of condition-averaged N2 amplitude for congruent and incongruent trials (as well as 'go' and 'nogo' trials), separately.

As discussed, the N2 component has moderate to excellent retest reliability in typically developing adults and children with ADHD (Clayson and Larson 2013; Kompatsiari et al. 2016). Provided the heterogeneity of neural responses in children with ASD (Salmond et al. 2007), we hypothesized that the N2 component would have moderate test-retest reliability in our sample.

Provided the growing interest in using ERPs to characterize clinical, developmental populations, we also explored individual differences by assessing ICCs from subgroups split by age. We posited that ICCs would be greatest among older children (Hämmerer et al. 2013). In all cases, moderate (to excellent) test-retest reliability of N2 amplitude would lend support to the use of this component as a potential target biomeasure of executive dysfunction in ASD.

2. Methods

2.1 Participants

As part of a larger clinical trial examining the effects of a three-month executive function intervention, children with an existing diagnosis of ASD were recruited from a participant registry in a hospital setting and local community sources. The three-month study period is comparable to a similar study targeting executive function in typically developing children (Rueda et al. 2005, 2012). Thus, the retest reliability of the N2 component across this period of time is of particular interest to clinical researchers.

Caregivers of interested families participated in a phone-screening interview to assess their child's eligibility. Exclusionary criteria included a history of head injury, seizures, significant sensory or motor impairment, or medical condition that affected the central nervous system. Children were also excluded if they had ever taken any anticonvulsant medications, as these may impact EEG recordings (Konishi et al. 1995). In accordance with guidelines for EEG/ERP studies in ASD (Webb et al. 2015), other medications were not exclusionary. Nonetheless, as change(s) in medication or therapy may impact test-retest reliability of neural components over time (Webb et al. 2015), post-hoc retest reliability estimates were also evaluated among a subsample of children without changes in medication or therapy during the study period.

At the initial time point, diagnosis of ASD was confirmed using the Autism Diagnostic Interview-Revised (ADI-R; Rutter et al. 2015) and the Autism Diagnostic Observation Schedule (ADOS-2; Gotham et al. 2007). All participants also had a Wechsler Abbreviated Scale of Intelligence (WASI-2; Wechsler 2013) full scale IQ estimated at 80 or above.

Thirty-five children (31 males, $M_{\text{age}} = 8.69$, $SD = 1.41$ years at Time 1) were randomly assigned to the waitlist condition (i.e., no intervention) of the larger clinical trial. Thus, participants provided ERP data at two time points, approximately three months apart, without receiving the study intervention in the intervening time. Two children were lost to follow-up between time points. Of the 33 children that participated in both time points, 21 children (18 males, $M_{\text{age}} = 9.00$, $SD = 1.52$ years at Time 1) provided usable ERP data (10 trials per condition; Lamm et al. 2006; Todd et al. 2008) during the Flanker task at both time

points. Fourteen children (11 males, $M_{\text{age}} = 9.14$, $SD = 1.35$ years at Time 1) provided usable ERP data (10 trials per condition) during the Go/Nogo Task. As the Go/Nogo task was administered *after* the Flanker task (see Procedure), this attrition was expected.

There were no differences in age, IQ, or diagnostic severity for the children who provided usable data at both time points compared with those who did not ($ps = 0.107$ and 0.119 for the Flanker and Go/Nogo tasks, respectively). The average duration between time points was 10.72 weeks ($SD = 2.03$ weeks) for the 21 children with usable Flanker ERP data and 10.52 weeks ($SD = 1.83$) for the 14 children with usable Go/Nogo data. Caregiver-report on the Behavior Rating Inventory of Executive Function (BRIEF; Gioia, Isquith, Guy, & Kenworthy, 2000) indicated that 80% of the children included in the final sample had clinically significant deficits in executive function (T-Scores on the Global Executive Composite ≥ 65). Additional information regarding participant demographics for the sample included in the final dataset ($ns = 21$ and 14 for the Flanker and Go/Nogo tasks, respectively) is included in Table 1.

2.2 Procedure

ERP responses were recorded during computerized Flanker and Go/Nogo tasks. The tasks were administered in a fixed order (Flanker first, Go/Nogo second) at the beginning of two visits scheduled approximately three months apart. The Human Subjects Division approved all study procedures. Caregiver consent and child assent was obtained prior to testing.

2.3 Flanker Task

A Flanker task was used to elicit an N2 ERP component (Faja et al. 2016; Rueda et al. 2004; Figure 1A). The task included 12 practice trials and 108 test trials presented over three blocks (36 test trials/block). For each trial, children were instructed to indicate the orientation of a central target stimulus via button press on a keypad (left versus right) as quickly and accurately as possible. Trials were “congruent” if the target stimulus was oriented in the same direction as the flanking stimuli on either side of the target. Trials were “incongruent” if the target stimulus was oriented in the direction opposite of the flanking stimuli. An equal number of congruent (50%) and incongruent (50%) trials were presented during each block. Each trial was preceded by a tone played for 150 ms and a fixation cross presented on the screen for 450 ms. Test stimuli were then presented for 2000 ms or until a response was registered. Thus, responses made more than 2000 ms after the stimulus was first presented on the screen were marked as incorrect. Visual and auditory feedback for correct and incorrect responses was presented after each response. Only ERPs for correct trials were analyzed.

2.4 Go/Nogo Task

A cued Go/Nogo task (Figure 1B) was also used to elicit an N2 component in children with ASD. The task included 17 training trials, up to three blocks of 20 practice trials (practice trials repeated until children reached 80% accuracy), and 200 test trials presented over four blocks (50 test trials/block). During ‘go’ trials (70% of trials), children were instructed to press a single button on a keypad each time a letter appeared on the screen. During ‘nogo’ trials (30% of trials), children were instructed to withhold their response when a specific

letter ('Y' or 'D'; order counterbalanced across participants) appeared on the screen. To account for the confound of frequency with condition, two letters were used for 'go' trials – an infrequent 'go' (30% of trials) and a frequent 'go' (40% of trials). 'Go' responses were analyzed only for the infrequent 'go' trials. Each trial was preceded by a fixation cross presented on the screen for 500 ms followed by a 750 ms delay. Test trials were then presented for 700 ms and there was a random intertrial interval (between 500 and 1000 ms) between trials. Responses were recorded while the stimulus was on the screen and during the variable intertrial interval. Thus, response windows varied between 1200 and 1700 ms after stimulus onset. Visual feedback was presented only during practice trials. Only ERPs for correct trials that followed a correct 'go' response were analyzed in order to ensure consistent motor response on the previous trial.

2.5 Electrophysiological Measurement

During the Flanker and Go/Nogo tasks, ERP responses were recorded using Net Amps 400 (Electrical Geodesics, Inc.) from a 128-channel Hydrocel Geodesic Sensor Net (GSN). The net was soaked in a potassium chloride electrolyte solution then placed on the participant's head following manufacturer specifications. Impedances were below 50 k Ω at the start of each recording. Data were recorded using a vertex reference and a 4 KHz antialiasing hardware filter at a sampling rate of 500 Hz.

2.6 Data Editing and Extraction

Using Netstation 5, data were re-filtered off-line using a high-pass (0.1 Hz) and low-pass (30 Hz) filter (Kaiser-type FIR filter with 2 Hz rolloff). Data were then segmented with a 200 ms baseline period that immediately preceded stimulus onset and a duration of 800 ms after stimulus onset. Trials with incorrect behavioral responses or artifacts were excluded.

Following segmentation, the data were baseline corrected using the average of their respective baseline periods. Automated artifact detection excluded trials with the following parameters: (1) presence of an eye blink using the Netstation Eye Blink algorithm set at 220 μ V with an 80 ms moving average and (2) more than 10 channels with a fluctuation exceeding 140 μ V or less than 1 μ V with an 80 ms moving average. Data were then visually inspected to confirm the results of automated artifact detection. During visual inspection, additional segments were excluded if they contained significant drift, movement artifacts, eye movements or blinks, or mechanical artifacts.

Next, channels that were marked as containing an artifact for more than 20% of trials were replaced using spherical spline interpolation which estimates voltages by imputing across channels. Data were then averaged individually by condition and re-referenced offline to the average of all electrodes (excluding the four eye channels) using the polar average reference effect (PARE) correction. Data were baseline corrected once more and individually inspected for quality.

Consistent with previous studies (Faja et al. 2016), the N2 component was derived from an electrode cluster that included Fz and adjacent electrodes (Hydrocel GSN electrodes 4, 11, and 19). The N2 mean amplitude was computed with a window of 300 to 400 ms post-stimulus.

2.7 Data Analysis

ICC is an index of test-retest reliability that accounts for the degree of correlation and agreement between measurements (Koo and Li 2016). ICC estimates and the corresponding 95% confidence intervals were calculated for mean N2 amplitude obtained during congruent and incongruent trials of the Flanker task as well as 'go' and 'nogo' trials of the Go/Nogo task between time points.

To explore age-related differences in retest reliability, we used a median split to group participants into younger and older cohorts. ICCs were then computed for each cohort and trial type, separately. Additionally, we ran a set of post-hoc ICC analyses excluding children whose medication or therapy changed between the first and second time points (based on caregiver-report). Of the 21 children with usable ERP data from the Flanker task, eleven children had notable changes to medication or therapy (e.g., hours of ABA) between time points. To evaluate ICCs in the subsample of children without notable changes to medication or therapy, these eleven children were excluded from analyses, and post-hoc ICCs were re-run on the subsample of 10 children without changes in these domains. For the Go/Nogo task, five of the 14 children with usable data had changes to medication or therapy between time points. Here too, ICCs were re-run on the subsample of nine children with usable Go/Nogo data and no changes to medication or therapy.

Although these secondary, exploratory analyses have small sample sizes and, consequently, low statistical power, these ICCs were intended to explore the retest reliability of the N2 component in specific age groups and independent of changes to medication and therapy in the intervening time between ERP measurements. Provided the small sample sizes, these results should be interpreted with caution.

All calculations were performed in SPSS Version 23.0 (IBM Corp) using a two-way mixed effects model with absolute agreement (Kompatsiari et al. 2016; Koo and Li 2016). In accordance with standard reliability classification rates (Portney and Watkins 2000), ICC values < 0.5 indicated poor reliability, values between 0.5 and 0.75 indicated moderate reliability, values between 0.75 and 0.9 indicated good reliability, and values > 0.9 indicated excellent reliability.

3. Results

Table 2 outlines the average number of trials included in analyses, mean N2 amplitude for Times 1 and 2, and ICC estimates for each task and trial type. Although highly correlated ($r = 0.721$, $p = 0.001$), N2 amplitudes did not differ between congruent and incongruent trials ($t(20) = 1.584$, $p = 0.129$), consistent with the results in children with (Faja et al., 2016) and without ASD (Abundis-Gutiérrez et al., 2014; Hoyniak, 2017; Ladouceur et al., 2007; Rueda et al., 2004). The same was true for 'go' and 'nogo' trials of the Go/Nogo task ($r = 0.615$, $p = 0.019$; $t(13) = 0.451$, $p = 0.660$). Behavioral outcomes for each task are outlined in Table 3. As expected, accuracy was greater (main effect of trial type: $F(1,20) = 5.25$, $p = 0.033$) and cRT faster (main effect of trial type: $F(1,20) = 36.86$, $p = 0.001$) for congruent compared to incongruent trials of the Flanker task. Similarly, accuracy was greater for 'go' relative to 'nogo' trials (main effect of trial type: $F(1,13) = 71.71$, $p = 0.001$).

During the Flanker task, mean N2 amplitude had moderate reliability for both congruent (single measure ICC = 0.542, 95% CI [0.173, 0.782]; Figures 2 and 3) and incongruent trials (single measure ICC = 0.629, 95% CI [0.276, 0.831]; Figures 2 and 3). Similar results were obtained with the Go/Nogo task, such that mean N2 amplitude had good reliability for 'go' trials (single measure ICC = 0.817, 95% CI [0.535, 0.937]; Figures 4 and 5) and moderate reliability for 'nogo' trials (single measure ICC = 0.578, 95% CI [0.075, 0.843]; Figures 4 and 5).

3.1 Age-Related Differences in ICCs

We next explored age-related differences in retest reliability. To do so, we used a median split at 9.33 years of age (at Time 1), to differentiate our sample into 'younger' and 'older' cohorts. Using this criterion, 11 children (9 males, $M_{\text{age}} = 8.31$, $SD = 0.78$ years) were grouped into the younger cohort (7.17 to 9.33 years) and 10 children (9 males, $M_{\text{age}} = 10.66$, $SD = 0.70$ years) were grouped into the older cohort (9.83 to 11.83 years) for the Flanker task. Retest reliability for congruent and incongruent trials was poor among younger children (single measure ICCs = 0.130, 95% CI [-0.499, 0.660] and 0.169, 95% CI [-0.528, 0.692] for congruent and incongruent trials, respectively) but moderate to good among older children (single measure ICCs = 0.695, 95% CI [0.197, 0.913] and 0.861, 95% CI [0.560, 0.963] for congruent and incongruent trials, respectively).

Of the 14 children with usable data for the Go/Nogo task, seven children (5 males, $M_{\text{age}} = 8.54$, $SD = 0.77$ years) were grouped into the younger cohort (7.17 to 9.33 years) and seven children (6 males, $M_{\text{age}} = 10.61$, $SD = 0.78$ years) into the older cohort (9.83 to 11.83 years). Here too, retest reliability was poor for 'go' and 'nogo' trials among younger children (single measure ICCs = 0.311, 95% CI [-0.664, 0.844] and 0.283, 95% CI [-0.710, 0.836] for 'go' and 'nogo' trials, respectively) but moderate to good among older children (single measure ICCs = 0.877, 95% CI [0.499, 0.977] and 0.599, 95% CI [-0.263, 0.919] for 'go' and 'nogo' trials, respectively). The number of usable trials included in the computations of ERPs did not differ between cohorts for any trial type, at either time point ($p = 0.303$).

3.2 ICCs among Children without Changes to Medication or Therapy

Lastly, we evaluated ICCs in subsamples of children without notable changes to medication or therapy during the three-month period between time points. Among the 10 children with usable Flanker ERP data who did not have changes to medication or therapy between Times 1 and 2, mean N2 amplitude had good reliability (single measure ICCs = 0.810, 95% CI [0.393, 0.950] and 0.836, 95% CI [0.469, 0.957] for congruent and incongruent trials, respectively).

Similarly, of the nine children with usable Go/Nogo data who did not have changes in medication/therapy, mean N2 amplitude had moderate to good reliability (single measure ICCs = 0.816, 95% CI [0.411, 0.954] and 0.720, 95% CI [0.209, 0.928] for 'go' and 'nogo' trials, respectively).

4. Discussion

The aim of the current study was to evaluate test-retest reliability of the N2 ERP component in children, 7 to 11 years of age, with ASD. Our results indicate that mean N2 amplitude has moderate to good retest reliability when sampled three months apart. This finding was consistent across both a Flanker and Go/Nogo task. A median split analysis indicated that retest values were greatest among older children (> 9.33 years of age). Results were modestly improved when a subgroup of children with changes to medication and therapy were excluded from analyses. Collectively, these findings suggest that the N2 ERP component is a reliable measure in school-aged children with ASD.

A qualifying feature of a biomeasure is the ability to show sensitivity to change with experimental manipulation or intervention (Dawson et al. 2012; Robb et al. 2016). However, in order to evaluate sensitivity to change, the measure must have retest reliability across the intervening time during which the change will occur (e.g., pre- and post-measurements). The results of the present study demonstrate moderate to good retest reliability of the N2 in children with ASD across three months, suggesting that this ERP component may be a useful biomeasure in clinical research.

A second qualifying feature of a biomeasure is the ability to predict a clinical outcome or endpoint (Robb et al. 2016). The N2 is strongly associated with executive function in both typically developing children (Espinet et al. 2012, 2013; Brydges et al. 2014) and those with ASD (Pfefferbaum et al. 1985; Jodo and Kayama 1992; Faja et al. 2016). Recent studies by Rueda and colleagues (2012; 2005) have utilized the N2 as a biomeasure of executive function in typically developing children between 4 and 6 years of age. Their results indicate that the greatest difference in N2 amplitude between congruent and incongruent trials was observed among children who completed training on executive function exercises, relative to untrained children. This finding suggests that training improved the efficiency of neural correlates of executive function in typically developing children. Whether a similar intervention would alter the N2, and ultimately improve executive function, in children with ASD warrants further research.

Although retest reliability of the N2 amplitude was previously unexplored in ASD, our results align with studies demonstrating reliability of this component in typically developing adults (ICCs between 0.48 and 0.55; Clayson and Larson 2013) and children with ADHD (ICCs between 0.7 and 0.9; Kompatsiari et al. 2016). The methods used in these papers were comparable to the current study, as both Clayson & Larson (2013) and Kompatsiari and colleagues (2016) used similar executive function tasks (i.e., Flanker and Go/Nogo tasks) and reliability measurements (i.e., ICCs) of distinct trial types (rather than difference scores).

The observation that ICCs were consistent across two executive function tasks suggests that children who were able to provide more data during ERP tasks may have more efficient neural processes underlying executive function. This is particularly apparent when examining ICCs of 'nogo' trials of the Go/Nogo task: because this task was administered second, and is typically more difficult for school-aged children, reliability of the N2

component in this task suggests that children who were able to complete our experimental protocol, and respond to difficult trials accurately, may have more advanced executive function. Furthermore, ICCs from a subsample of 10 children who provided usable data for *both* the Flanker and Go/Nogo tasks were moderate to good (single measure ICCs between 0.588 and 0.860), providing additional support for this hypothesis.

The subgroup analysis was also consistent with our hypothesis, as ICCs were greatest among older children. Specifically, ICCs were greater among 9 to 11 year old children, relative to 7 to 9 year old children, consistent with evidence that the neural structures underlying executive function mature with development (Abundis-Gutiérrez et al., 2014; Buss et al., 2011; Espinet, Anderson, & Zelazo, 2012; Hoyniak, 2017). Notably, however, the older age group also had greater variability in N2 amplitude (see Figures 2 and 4). This suggests that, although there are individual differences in the development of executive functions (i.e., variability in N2 amplitude between individuals), amplitudes are consistent over time within individuals.

Given the utility of ERP in ASD populations (Jeste and Nelson 2009), and associations between these measures and executive function in typical and atypical development, EEG/ERP may be a particularly useful biomeasure in samples with wide ranging cognitive and intellectual abilities. Additional studies are needed to characterize retest reliability in more diverse samples, as participants in the current sample were required to have an IQ ≥ 80 and primarily male.

4.1 Limitations and Future Directions

In summary, the results of the current study demonstrate retest reliability of the N2 ERP component among school-aged children with ASD. However, there were limitations and unexpected findings that should be addressed in future research. First, the ICC value for congruent trials was lower than the value obtained for incongruent trials in the Flanker task (Table 2). We posit that state-dependent traits may have contributed to variability in N2 amplitude, and reduced ICC values, particularly on the less effortful, congruent trials. That is, because congruent trials are less directly related to the task demand of conflict monitoring, performance on these trials may be more state-dependent and susceptible to alternative factors (e.g., anxiety, see Booth & Paker, 2017). Additional studies are needed to empirically determine whether state-dependent traits influence variability of ERP components and ICCs in this population.

Second, there was notable variance in ICC values, as evidenced by the large 95% confidence intervals (Table 2). This variability is likely a consequence of individual differences within and across children with ASD. Future studies should evaluate the reliability of the N2 component in a larger sample, and include individuals with comorbidities and other developmental disabilities, to support the establishment of the N2 as a valid biomeasure of executive functioning in ASD.

Third, although our sample size was comparable to other test-retest studies in clinical populations (e.g., Kompatsiari et al. 2016), our analyses were limited, and may have been underpowered. According to Bujang & Baharum (2017), a sample of 15 participants is

needed for an ICC of 0.6, whereas a sample of 22 participants is needed for an ICC of 0.5 (assuming a null $R = 0$ and $\alpha = 0.05$) with 80% power. As such, we believe the sample from our primary analysis (Flanker task, $n = 21$) was sufficient to detect ICCs between 0.5 and 0.6 with 80% power. However, our secondary, exploratory results pertaining to age and therapy/medication use should be interpreted with caution, as subgrouping reduced sample sizes further and may have increased the likelihood of Type II error.

Our small sample size also limited the exploration of individual differences. Previous studies have demonstrated relations between N2 ERP component and executive function in ASD (Pfefferbaum et al. 1985; Jodo and Kayama 1992; Faja et al. 2016). Relations between N2 and executive function were not explored in the current study due to the small number of children who were able to provide usable ERP data at both time points. The small samples size also prevented a systematic exploration of the effect of trial quantity (which impacts the ERP signal) on ICCs in children with ASD. Additionally, 80% of our sample had clinically significant deficits in executive function. As not all children with ASD exhibit executive dysfunction, replication with larger and more diverse samples is needed to adequately assess retest reliability in this population. Future studies should also include an age-, IQ-, and gender-matched control group(s) to facilitate interpretation of ICC reliabilities relative to typically developing children.

The intervening time between the test and retest assessments in our study was notably longer than those reported in similar studies (Clayson and Larson 2013; Hämmerer et al. 2013; Kompatsiari et al. 2016), suggesting that N2 amplitude has strong stability over a longer period of time than is typically reported. However, the longer retest period may have contributed to variability in the data and lowered ICC values, particularly in younger children whose executive function is still rapidly developing. Thus, replication with a larger sample and longer delay period (e.g., six months to one year) is needed.

5. Conclusions

To better understand the widespread heterogeneity in ASD, many researchers have begun to search for biomeasures that underlie specific impairments characteristic of the disorder. In recent years, the N2 ERP component has gained traction as a candidate biomeasure of executive dysfunction in children with ASD. However, information regarding test-retest reliability of this component in this population is lacking. Here, we report that the N2 component has moderate to good test-retest reliability when sampled three months apart. These findings support the use of the N2 component as a biomeasure of executive function and treatment response in school-aged children with ASD.

Acknowledgements:

The authors would like to thank the staff and students who assisted with collecting and scoring these measures, Sara Jane Webb for her contributions to the development of the Go/Nogo task, and, especially, the children and families who participated in this study.

Funding: This work was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number K99/R00HD071966 (S.F.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Abundis-Gutiérrez A, Checa P, Castellanos C, Rosario Rueda M. Electrophysiological correlates of attention networks in childhood and early adulthood. *Neuropsychologia*. 2014;57(1):78–92. [PubMed: 24593898]
- American Psychiatric Association. DSM-V. American Journal of Psychiatry. 2013.
- Baio J, Wiggins L, Christensen D, Maenner M, Daniels J, Wareen Z, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *Surveill Summ*. 2018;67(6):1–23.
- Banaschewski T, Brandeis D. Annotation: What electrical brain activity tells us about brain function that other techniques cannot tell us – A child psychiatric perspective. *J Child Psychol Psychiatry*. 2007;48(5):415–35. [PubMed: 17501723]
- Booth RW, Paker M. State anxiety impairs attentional control when other sources of control are minimal. *Cogn Emot*. 2017;31(5):1004–11. [PubMed: 27073974]
- Brydges CR, Fox AM, Reid CL, Anderson M. Predictive validity of the N2 and P3 ERP components to executive functioning in children: A latent-variable analysis. *Front Hum Neurosci*. 2014;8:1–10. Available from: <http://journal.frontiersin.org/article/10.3389/fnhum.2014.00080/abstract> [PubMed: 24474914]
- Bujang MA, Baharum N. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Arch Orofac Sci*. 2017;12(1):1–11.
- Buss KA, Dennis TA, Brooker RJ, Sippel LM. An ERP study of conflict monitoring in 4-8-year old children: Associations with temperament. *Dev Cogn Neurosci*. 2011;1(2):131–40. Available from: <http://dx.doi.org/10.1016/j.dcn.2010.12.003> [PubMed: 21666879]
- Casanova MF, Baruth JM, El-Baz A, Tasman A, Sears L, Sokhadze E. Repetitive transcranial magnetic stimulation (rTMS) modulates event-related potential (ERP) indices of attention in autism. *Transl Neurosci*. 2012;3(2):170–80. [PubMed: 24683490]
- Clayson PE, Larson MJ. Psychometric properties of conflict monitoring and conflict adaptation indices: Response time and conflict N2 event-related potentials. *Psychophysiology*. 2013;50(12):1209–19. [PubMed: 23992600]
- Dawson G, Jones EJH, Merkle K, Venema K, Lowy R, Faja S, et al. Early behavioral intervention is associated with normalized brain activity in young children with autism. *J Am Acad Child Adolesc Psychiatry*. 2012;51(11):1150–9. Available from: 10.1016/j.jaac.2012.08.018 [PubMed: 23101741]
- Demetriou EA, Lampit A, Quintana DS, Naismith SL, Song YJC, Pye JE, et al. Autism spectrum disorders: A meta-analysis of executive function. *Nat Publ Gr*. 2017;23(5):1198–204. Available from: 10.1038/mp.2017.75
- Diamond A Executive Functions. *Annu Rev Clin Psychol*. 2014;64:135–68.
- Espinete SD, Anderson JE, Zelazo PD. N2 amplitude as a neural marker of executive function in young children: An ERP study of children who switch versus perseverate on the Dimensional Change Card Sort. *Dev Cogn Neurosci*. 2012;2(SUPPL. 1):S49–58. Available from: 10.1016/j.dcn.2011.12.002 [PubMed: 22682910]
- Espinete SD, Anderson JE, Zelazo PD. Reflection training improves executive function in preschool-age children: Behavioral and neural effects. *Dev Cogn Neurosci*. 2013;4:3–15. Available from: 10.1016/j.dcn.2012.11.009 [PubMed: 23280362]
- Faja S, Clarkson T, Webb SJ. Neural and behavioral suppression of interfering flankers by children with and without autism spectrum disorder. *Neuropsychologia*. 2016;93(May):251–61. Available from: 10.1016/j.neuropsychologia.2016.10.017 [PubMed: 27825750]
- Foss-Feig JH, Stavropoulos KKM, Mcpartland JC, Wallace MT, Stone WL, Key AP. Electrophysiological response during auditory gap detection: Biomarker for sensory and communication alterations in autism spectrum disorder? *Dev Neuropsychol*. 2018;43(2):109–22. [PubMed: 29436853]
- Friedman NP, Miyake A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*. 2017.

- Geurts HM, De Vries M, Van den Bergh SFWM. Executive functioning theory and autism In Goldstein S & Naglieri JA, Eds., *Handbook of Executive Functioning*. 2014 New York: Springer.
- Gioia G, Isquith P, Guy SC, Kenworthy L. Test review: Behavior Rating Inventory of Executive Function. *Child Neuropsychol*. 2000; 6(3): 235–238. [PubMed: 11419452]
- Gotham K, Risi S, Pickles A, Lord C. The autism diagnostic observation schedule: Revised algorithms for improved diagnostic validity. *J Autism Dev Disord*. 2007;37(4):613–27. [PubMed: 17180459]
- Hämmerer D, Li SC, Völkle M, Müller V, Lindenberger U. A lifespan comparison of the reliability, test-retest stability, and signal-to-noise ratio of event-related potentials assessed during performance monitoring. *Psychophysiology*. 2013;50(1):111–23. [PubMed: 23110313]
- Hoyniak C Changes in the NoGo N2 event-related potential component across childhood: A systematic review and meta-analysis. *Dev Neuropsychol*. 2017;42(1):1–24. [PubMed: 28059561]
- Jeste SS, Nelson CA. Event related potentials in the understanding of autism spectrum disorders: An analytical review. *J Autism Dev Disord*. 2009;39(3):495–510. [PubMed: 18850262]
- Jodo E, Kayama Y. Relation of a negative ERP component to response inhibition in a Go/No-go task. *Electroencephalogr Clin Neurophysiol*. 1992;82:477–82. [PubMed: 1375556]
- Kenworthy L, Yerys BE, Anthony LG, Wallace L. Understanding executive control in autism spectrum disorders in the lab and in the real world. *Neuropsychol Rev*. 2008;18(4):320–38. [PubMed: 18956239]
- Kim SH, Grammer J, Benrey N, Morrison F, Lord C. Stimulus processing and error monitoring in more-able kindergarteners with autism spectrum disorder: A short review and a preliminary Event-Related Potentials study. *Eur J Neurosci*. 2018;47(6):556–67. [PubMed: 28394438]
- Kompatsiari K, Candrian G, Mueller A. Test-retest reliability of ERP components: A short-term replication of a visual Go/NoGo task in ADHD subjects. *Neurosci Lett*. 2016;617:166–72. Available from: 10.1016/j.neulet.2016.02.012 [PubMed: 26861197]
- Konishi T, Naganuma Y, Hongou K, Murakami M, Yamatani M, Okada T. Effects of antiepileptic drugs on EEG background activity in children with epilepsy: Initial phase of therapy. *Clin Electroencephalogr*. 1995;26(2):113–99. [PubMed: 7781190]
- Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–63. Available from: 10.1016/j.jcm.2016.02.012 [PubMed: 27330520]
- Ladouceur CD, Dahl RE, Carter CS. Development of action monitoring through adolescence into adulthood: ERP and source localization. *Dev Sci*. 2007;10(6):874–891. [PubMed: 17973802]
- Lamm C, Zelazo PD, Lewis MD. Neural correlates of cognitive control in childhood and adolescence: Disentangling the contributions of age and executive function. *Neuropsychologia*. 2006;44(11):2139–48. [PubMed: 16310813]
- Larson MJ, South M, Clayson PE, Clawson A. Cognitive control and conflict adaptation in youth with high-functioning autism. *J Child Psychol Psychiatry*. 2012;4:440–8.
- Loth E, Charman T, Mason L, Tillmann J, Jones E, Wooldridge C, et al. The EU-AIMS Longitudinal European Autism Project (LEAP): Design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Mol Autism*. 2017;8(24):1–19. [PubMed: 28070266]
- Meyer A, Lerner MD, De Los Reyes A, Laird RD, Hajcak G. Considering ERP difference scores as individual difference measures: Issues with subtraction and alternative approaches. *Psychophysiology*. 2017;54(1):114–22. [PubMed: 28000251]
- Pfefferbaum A, Ford JM, Weller BJ, Kopell BS. ERPs to response production and inhibition. *Electroencephalogr Clin Neurophysiol*. 1985;60(5):423–34. [PubMed: 2580694]
- Portney L, Watkins M. *Foundations of clinical research: Applications to practice* (Vol. 2). Survey of Ophthalmology. 2000.
- Robb MA, McInnes PM, Califf RM. Biomarkers and surrogate endpoints: Developing common terminology and definitions. *JAMA - J Am Med Assoc*. 2016;315(11):1107–8.
- Rueda MR, Checa P, Cómbita LM. Enhanced efficiency of the executive attention network after training in preschool children: Immediate changes and effects after two months. *Dev Cogn Neurosci*. 2012;2(SUPPL. 1):S192–204. Available from: 10.1016/j.dcn.2011.09.004 [PubMed: 22682908]

- Rueda MR, Posner MI, Rothbart MK, Davis-Stober CP. Development of the time course for processing conflict: An event-related potentials study with 4 year olds and adults. *BMC Neurosci.* 2004;5(39):1–13. [PubMed: 14720305]
- Rueda MR, Rothbart MK, McCandliss BD, Saccomanno L, Posner MI. Training, maturation, and genetic influences on the development of executive attention. *Proc Natl Acad Sci U S A.* 2005;102(41):14931–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16192352> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1253585> [PubMed: 16192352]
- Russell J Autism as an executive disorder. Vol. 313, New York NY US Oxford University Press 1997.
- Rutter M, LeCouteur A, Lord C. Autism Diagnostic Interview - Revised (ADI-R). Statew Agric L Use Baseline 2015. 2015;1.
- Salmond CH, Vargha-Khadem F, Gadian DG, de Haan M, Baldeweg T. Heterogeneity in the patterns of neural abnormality in autistic spectrum disorders: Evidence from ERP and MRI. *Cortex.* 2007;43(6):686–99. [PubMed: 17710821]
- Samyn V, Wiersema JR, Bijttebier P, Roeyers H. Effortful control and executive attention in typical and atypical development: An event-related potential study. *Biol Psychol.* 2014;99:160–71. [PubMed: 24686073]
- Siper PM, Zemon V, Gordon J, George-jones J, Lurie S, Zweifach J, et al. Rapid and objective assessment of neural function in autism spectrum disorder using transient Visual Evoked Potentials. *PLoS One.* 2016;11(20):e0164422. [PubMed: 27716799]
- Sokhadze EM, Tasman A, Sokhadze GE, El-baz AS, Casanova MF. Behavioral, cognitive, and motor preparation deficits in a visual cued spatial attention task in autism spectrum disorder. *Appl Psychophysiol Biofeedback.* 2017;41(1):81–92.
- Thigpen N, Kappenman E, Keil A. Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology.* 2018;54(1):123–38.
- Todd RM, Lewis MD, Meusel LA, Zelazo PD. The time course of social-emotional processing in early childhood: ERP responses to facial affect and personal familiarity in a Go-Nogo task. *Neuropsychologia.* 2008;46(2):595–613. [PubMed: 18061633]
- Tye C, Asherson P, Ashwood KL, Azadi B, Bolton P, Mcloughlin G. Attention and inhibition in children with ASD, ADHD and co-morbid ASD + ADHD: An event-related potential study. *Psychol Med.* 2014;44:1101–16. [PubMed: 23673307]
- Webb SJ, Bernier R, Henderson HA, Johnson MH, Jones EJH, Lerner MD, et al. Guidelines and best practices for electrophysiological data collection, analysis and reporting in autism. *J Autism Dev Disord.* 2015;45(2):425–43. [PubMed: 23975145]
- Wechsler D WASI -II: Wechsler abbreviated scale of intelligence - second edition. *J Psychoeduc Assess.* 2013;31(3):337–41.

Highlights

- The N2 ERP component is commonly used as a biomeasure in children with autism.
- The test-retest reliability of N2 has not been explored in this population.
- Here, we report moderate to good retest reliability of N2 amplitude in children with autism.

Figure 1.
Schematics of the Flanker (A) and Go/Nogo tasks (B).

Figure 2. Averaged N2 ERP responses to congruent and incongruent trials of the Flanker task during Time 1 (top) and Time 2 (bottom).

Figure 3. Individual participant N2 amplitudes at Time 1 and 2 for congruent (top) and incongruent (bottom) trials of the Flanker task for younger (blue) and older (red) children.
*Notes: ns = not significant; * $p < 0.01$; ** $p < 0.001$*

Figure 4. Averaged N2 ERP responses to 'go' and 'nogo' trials of the Go/Nogo task during Time 1 (top) and Time 2 (bottom).

Figure 5. Individual participant N2 amplitudes at Time 1 and 2 for 'go' (top) and 'nogo' (bottom) trials of the Go/Nogo task for younger (blue) and older (red) children.
*Notes: ns = not significant; * $p < 0.01$*

Table 1.

Participant demographics.

	Flanker Task (<i>n</i> = 21) Mean (SD)	Go/Nogo Task (<i>n</i> = 14) Mean (SD)
Age (years)	9.00 (1.52)	9.14 (1.35)
Sex (males: females)	18:3	11:3
ADOS-2 Calibrated Severity Score	8.67 (1.16)	8.50 (1.09)
WASI-2 Full Scale	103.43 (10.79)	101.93 (11.87)
WASI-2 Verbal	103.10 (9.82)	102.36 (10.17)
WASI-2 Perceptual	103.10 (13.49)	101.21 (15.62)
BRIEF GEC Scale	69.00 (8.51)	72.07 (6.39)

Notes: SD = standard deviation; ADOS = Autism Diagnostic Observation Schedule; WASI-2 = Wechsler Abbreviated Scale of Intelligence; BRIEF = Behavior Rating Inventory of Executive Function; GEC = Global Executive Composite

Table 2.

Mean N2 amplitude and ICC estimates during the Flanker and Go/Nogo tasks.

	Time 1		Time 2		Single measure ICC [95% CI]
	Number of trials Mean (SD)	Mean N2 amplitude Mean (SD)	Number of trials Mean (SD)	Mean N2 amplitude Mean (SD)	
Flanker Task					
Congruent	28.67 (9.90)	-2.32 (3.23)	29.52 (11.39)	-3.68 (4.18)	0.542 [0.173, 0.782] *
Incongruent	28.95 (9.28)	-3.46 (4.72)	30.71 (11.43)	-3.11 (4.53)	0.629 [0.276, 0.831] **
Go/Nogo Task					
'Go'	30.43 (11.88)	-2.69 (4.14)	35.14 (9.50)	-2.11 (3.00)	0.817 [0.535, 0.937] **
'Nogo'	26.93 (10.04)	-3.15 (4.43)	25.79 (5.69)	-2.75 (4.59)	0.578 [0.075, 0.843] *

Notes: SD = standard deviation; ICC = intraclass correlation; 95% CI = 95% confidence interval

* p 0.01

** p 0.001

Table 3.

Behavioral outcomes for the Flanker and Go/Nogo tasks.

	Time 1		Time 2	
	Accuracy (%) Mean (SD)	cRT (ms) Mean (SD)	Accuracy (%) Mean (SD)	cRT (ms) Mean (SD)
Flanker Task				
Congruent	94.27 (6.90)	738.03 (153.88)	95.29 (8.76)	646.63 (107.32)
Incongruent	90.74 (9.69)	774.26 (156.35)	94.12 (8.85)	682.07 (113.93)
Go/Nogo Task				
'Go'	95.61 (4.65)	511.01 (106.16)	96.89 (3.25)	490.04 (92.18)
'Nogo'	75.36 (10.86)	-	72.38 (12.05)	-

Notes: SD = standard deviation; cRT = correct reaction time