Title:

Towards an individualised neural assessment of receptive language in children

Authors, Affiliations

Selene Petit[*,a,b,c], Nicholas A. Badcock[a,b,c], Tijl Grootswagers[a,b,c,h], Anina N. Rich[a,b,c], Jon Brock[b,c], Lyndsey Nickels[b,c], Denise Moerel[a,b], Nadene Dermody[a,g], Shu Yau[c,d], Elaine Schmidt[e,f], Alexandra Woolgar[a,b,c,f]

a. Perception in Action Research Centre, Macquarie University, Australia

b. Department of Cognitive Science, Macquarie University, Australia

c. ARC Centre of Excellence in Cognition and its Disorders (CCD)

d. University of Bristol, Bristol, UK

e. Child Language Lab, Department of Linguistics, Macquarie University, Australia

f. Medical Research Council (UK), Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK

g. Max Planck Institute for Human Cognitive Brain Sciences, Leipzig, Germany

h. School of Psychology, University of Sydney, Australia

*. Corresponding author at: Department of Cognitive Science, Macquarie University, Australia. Selene.petit@mq.edu.au

Keywords

N400, EEG, language comprehension, sensitivity, Emotiv EPOC+, multivariate pattern analyses

Highlights

- New methods show hints that brain data can be used to infer language processing
- A gaming headset records EEG waveforms comparable to a research grade system
- 50% of individual children show significant N400 effects in two paradigms
- We can decode semantic context in up to 88% of children using multivariate analyses

Abstract:

Can we use electrophysiological data to assess language processing? Although we usually cannot infer cognition from brain data, there are some circumstances in which discriminative brain signals are enough to tell us that cognitive processing has occurred. For example, if we observe differential neural responses to identical stimuli that vary only in cognitive context, we can infer an influence of context on stimulus processing. Here, we used this logic to develop a test for receptive language ability in individual children. We were motivated by the suggestion that some non-verbal individuals, such as a subset of autistic children, may understand more language than they can demonstrate; for these people a neural test would be transformative. We developed two child-friendly paradigms in which typically developing children listened to identical spoken words in congruent and incongruent lexical-semantic contexts. In the congruent condition, the target word was strongly predicted by the context, while in the incongruent condition, the target word violated lexical-semantic predictions. In both paradigms, we simultaneously measured EEG with a research-grade

system, Neuroscan's SynAmps2, and an adapted gaming system, Emotiv's EPOC+. In both paradigms, at a group level, we found a statistically significant N400 effect to lexical-semantic violations, and we also detected significant effects in about half of our participants individually. The same effects were present, although with a numerically smaller amplitude, with EPOC+. Using multivariate analyses on individual children's EEG data increased our sensitivity to detect evidence of lexical-semantic processing, reaching an 88% detection rate in a paradigm using sentences. This provides a promising avenue to assess language comprehension in individuals, which could be used to identify language comprehension abilities in children who may otherwise struggle to demonstrate how much they understand.

Introduction

Language is a crucial part of everyday life, and something we often take for granted. In cases where people cannot speak or reliably communicate it can be difficult to assess whether the capacity to understand spoken language is present. Examples include cases of disorders of consciousness, minimally-verbal autism, or cerebral palsy (Giacino and Smart, 2007; Tager-Flusberg and Kasari, 2013). Electroencephalography (EEG) allows passive recording of the electrical response generated by brain activity, and offers the opportunity to measure language understanding in the absence of reliable behavioural responses. Here, our goal was to develop a measure of language comprehension that could be used at an individual level, using an accessible and inexpensive method for recording EEG signals.

We sought to elicit and record the N400 event-related potential (ERP) as a neural marker of language processing. The N400 ERP component is elicited by hearing or reading words, with the N400 peak being larger when the words violate the semantic or predictive context in which they are presented (Kutas, 1993; Kutas and A. Hillyard, 1980; Kutas and Federmeier, 2011). For instance, the word "door" elicits a larger N400 ERP when it is presented

in the sentence "The clouds are high up in the *door*", compared to the sentence "I had no key to open the *door*". This difference in the amplitude of the N400 due to context is called the N400 effect. This effect is well-documented, and has been recorded in groups of adults, children, and special populations (for a comprehensive review see Kutas and Federmeier, 2011), making it a strong candidate for assessing lexical-semantic processing. Here, we set out to examine two untested aspects of N400 studies which we considered critical for future clinical application. First, we tested whether N400 effects could be reliably recorded with a low-cost portable EEG system (Emotiv EPOC+). Second, despite consistent N400 results across studies, populations, and paradigms in groups of participants, few studies have quantified the reliability of the effect at the individual level, and even fewer have done so methodically in individual children (but see Cantiani et al. (2016)). We therefore examined the reliability of two auditory paradigms for detecting statistically significant N400 effects in individual children. We used both univariate analyses and multivariate pattern analyses to detect differential responses to identical spoken target words in the two semantic contexts.

In the first experiment, typically developing children listened to pairs of spoken words which were either normatively associated (e.g., "arm – *leg*") or unrelated (e.g., "boat – *leg*"). We then performed a second experiment, with the aim of replicating the findings from the first experiment using a different paradigm, and comparing the sensitivity of the two paradigms. In this experiment, children listened to spoken sentences with either a congruent completion (i.e., "she wore a necklace around her *neck*") or an incongruent/anomalous completion (i.e., "the princess may someday become a *neck*"). Both these paradigms are typically used in N400 studies, as they elicit strong and reliable N400 effects in adults and children (Borovsky et al., 2012; Friedrich and Friederici, 2005; Kiang et al., 2013; Rämä et al., 2013; Torkildsen et al., 2007). Based on this large body of literature, we predicted that the ERP evoked by the identical spoken word token (in our example, *neck*) would vary according to semantic context. To make

the paradigms suitable for children we created game-like tasks where children encountered friendly or evil aliens.

Standard EEG systems can be inconvenient to set up and rather uncomfortable. With a view to future clinical applications, we therefore sought to validate an accessible EEG system that avoids many of the typical setup inconveniences. A traditional 32-channel, gel-based EEG system takes around 35 minutes to setup and is somewhat uncomfortable, involving rubbing the participant's scalp with gel in order to bridge the EEG electrodes to the skin. In addition, typical EEG setups are not portable, with extensive wiring compelling participants to remain seated, and signals are best recorded in an electrically shielded room, which can be intimidating for some subjects. For these reasons, using a medical-grade EEG system is not always suitable for children or adults with cognitive disorders. Recently, more accessible and more portable EEG systems have become available, one of which has been adapted and validated for the measurement of auditory ERPs in adults (Badcock et al., 2013; de Lissa et al., 2015) and children (Badcock et al., 2015), and even autistic individuals (Yau et al., 2015). The Emotiv EPOC+ system, hereafter referred to as "EPOC+", was originally designed for gaming purposes, and consists of a wireless headset with 14 electrodes that connect to the scalp via saline solution-soaked cotton-rolls. The setup is fast (approx. 5-10 minutes), and it is not necessary to rub the scalp. This system is also low in cost compared with medical-grade systems, and is wireless and portable, allowing its use outside of the laboratory (e.g., in homes or schools).

Although the EPOC+ system has been validated against research grade systems for recording early ERPs, such as auditory ERPs (Badcock et al., 2015, 2013; Barham et al., 2017), and face-sensitive N170 (de Lissa et al., 2015), studies on later components such as the P300 have yielded less consistent results. Vos et al. (2014) report similar performance of the Emotiv amplifier compared to a research-grade amplifier, and Elsawy et al (2014) found acceptable

results when using a classifier on P300 EPOC+ data, while Duvinage et al. (2013) report that the EPOC+ recorded a significantly noisier signal compared to the research grade ANT system (Advanced Neuro Technology, ANT, Enschede, The Netherlands). Recording reliable N400 ERPs using the EPOC+ system would be a significant step towards our goal of easily identifying language comprehension in children who cannot speak. To our knowledge, no studies have tested the ability of EPOC+ to record N400 ERPs. Hence, the first aim of our study was to test the fidelity of the adapted Emotiv EPOC+ EEG system in measuring the N400 effect in children. To this end we recorded concurrently with the adapted EPOC+ system and a research-grade Neuroscan system during the experimental tasks.

Finally, to sensitively assess language comprehension in non-speaking children, it is crucial that the paradigm yields reliable responses at an individual level. Despite the large body of literature on the N400 effect in groups of participants, few studies report the reliability of the N400 effect in individual subjects, and even fewer of these studies used child participants (but see Cantiani et al., (2016)). Cruse et al. (2014) investigated the sensitivity of the N400 in adults at an individual level using semantic congruency paradigms with spoken sentences and semantically or normatively associated word pairs. They reported that even though clear N400 effects were always obtained for the group as a whole, statistically reliable effects at the individual subject level were seen for only 0% - 75% of subjects, depending on the paradigm. This indicates large inter-individual variability, which is likely to be at least as large in children due to increased movement, internal noise, and poorer focus on the task. Although Cantiani et al. (2016) reported that N400 effects were present in 80% of their child participants (aged 4 to 7 years), this was determined by a visual inspection of the ERP and was not quantitatively analysed. Thus, our second objective was to quantify the detection rate of statistically significant N400 effects in children.

We assessed the ERP differences between semantic conditions using both typical univariate N400 analyses and multivariate pattern analyses (MVPA). Traditional univariate analyses of ERPs usually require an *a priori* choice of electrodes and time points of interest. When testing individual participants, especially children and special populations, this *a priori* knowledge may not be available. Thus, in addition to univariate analyses of the N400 effect, we tested the sensitivity of MVPA on our EEG data. To this end we trained a linear classifier to discriminate between the two semantic conditions (congruent and incongruent) on individual EEG data. The advantage of MVPA, relative to univariate analysis, is that it targets the information contained in the pattern of activation across sensors, making it robust to individual differences in signal direction and topology (Grootswagers et al., 2017; Haynes, 2015; Hebart and Baker, 2018).

We found that the EPOC+ system recorded similar N400 effects to the research-grade EEG system. We found robust group-level level univariate N400 effects, in both paradigms and systems, and our individual subject detection rate was around 50%. MVPA increased the detection rate to 88% but only in the sentences paradigm for the research-grade EEG system. These data may help direct future development of paradigms for assessing language comprehension in non-speaking children.

Experiment one: normatively associated word pairs

Methods

Participants

16 children were recruited using the Neuronauts database of the Australian Research Council Centre of Excellence in Cognition and its Disorders (CCD). All participants were native English speakers and had non-verbal reasoning and verbal abilities within the normal

range as measured by the matrices section of the Kaufman Brief Intelligence Test, Second Edition (K-BIT 2, Kaufman and Kaufman, 2004) and the Peabody Picture Vocabulary Test—4th Edition (PPVT – 4, Dunn and Dunn, 2007). Participants received $25 for their participation, as well as a sticker and a certificate. The data from one participant were excluded due to technical issues during recording. The final set of data thus came from 15 participants (age range: 6 to 12 years old, *M=9.2, SD=2.6,* 4 male and 11 female). This study was approved by the Macquarie University Human Research Ethics Committee (Reference number: 5201200658). Participants' parents or guardians provided written consent and the children provided verbal consent.

Stimuli

Stimuli comprised 63 pairs of normatively associated words. Following Cruse et al. (2014), we began with word pairs taken from the Nelson et al. (1998) free association norms database. These norms comprise a large number of cue-target pairings, developed by asking participants to produce the first meaningfully or strongly associated word that comes to mind when presented with a particular cue. We initially chose pairs from the normative database with a forward associative strength (cue to target) greater than 0.5. That is, for each cue included in the list, more than 50% of the participants in the Nelson et al. norm-development study produced the target word in response to that cue. We included only pairs where the target was a noun, and where the cue and target were one syllable in length. We also included only words that had an Age of Acquisition (AoA) rating of 8-years or less (Kuperman et al., 2012), meaning that these words were typically known by children of 8-years and above. We excluded any pairs where either the cue or target had a homophone (according to the N-watch database; Davis, 2005), where the AoA was less than or equal to 10 years, and we excluded pairs where the cue or target was not applicable to the Australian context (e.g., FUEL-GAS).

To minimise repetition across the stimulus set, we allowed each target word to appear in a maximum of two word-pairs. The cue words were only used once as cues, but they could also appear up to twice as targets. Thus, the maximum number of repetitions of particular words across the entire list of related items was three (i.e., once as a cue, and twice as a target – this was the case for 13 words). For word pairs with singular and plural forms (e.g., GIRL-BOY and GIRLS-BOYS), only the pair with the strongest association was included. The final set of 63 pairs had a mean forward associative strength of 0.676 (see supplementary Table S1 for the list of stimuli). These word pairs formed the *related* condition.

We created a list of 63 *unrelated* word pairs by recombining the cue and target words from the related condition. Constructing the unrelated list in this way ensured a fully balanced design in which the cues and targets in the related and unrelated lists were identical (and therefore matched for word frequency, familiarity, phoneme length, etc.). We ensured that target words in the unrelated condition did not start with the same sound, rhyme, or have any semantic or associative connections with the cue or related target. In addition, we respected the grammatical number structure of the related word pair when choosing an unrelated target. For example, in creating an unrelated combination for a plural-singular pair (e.g., SUDS-SOAP), another singular target word was chosen (e.g., SUDS-ART).

Stimuli were digitally recorded by a female native Australian-English speaker, and the best auditory tokens, where the voice had a natural intonation and was not raspy, were selected using Praat software (Boersma, 2001). We used the same target tokens in the related and unrelated conditions so that there were no auditory differences to drive a differential EEG response to the two conditions. For each target, the related and the unrelated cue were recorded close together in time, and were chosen to have approximately the same length, intensity, and voice quality as the target.

EEG Equipment

We recorded simultaneously from two EEG systems in an electrically-shielded room. The research EEG system Neuroscan SynAmps2 (Scan version 4.3) Ag-AgCl electrodes were fitted to an elastic cap (Easy Cap, Herrsching, Germany) at 33 locations (Fig. 1), according to the international 10-20 system, including M1 (online reference), AFz (ground electrode), and M2. We measured vertical and horizontal eye movements with electrodes placed above and below the left eye and next to the outer canthus of each eye. Neuroscan was sampled at 1000Hz (downsampled to 500Hz during processing) with an online bandpass filter from 1 to 100Hz. We marked the onset of each sentence and target word using parallel port events generated using the Psychophysics Toolbox 3 extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in MATLAB.

The EPOC+ is a wireless headset with flexible plastic arms holding 16 gold-plated sensors. In order to accommodate the concurrent setup with the Neuroscan system, we placed the EPOC+ sensors at the following scalp locations of the international 10-20 system (See Fig. 1)[1]. M1 acted as the online reference, and M2 was a feed-forward reference that reduced external electrical interference. The signals from the other 14 channels were high-pass filtered online with a 0.16 Hz cut-off, pre-amplified and low-pass filtered at an 83 Hz cut-off. The analogue signals were then digitised at 2048 Hz. The digitised signal was filtered using a 5th-order sinc filter to notch out 50Hz and 60 Hz, low-pass filtered and down-sampled to 128 Hz (specifications taken from the EPOC+ system web forum). The effective bandwidth was 0.16–43 Hz.

[1] Note that in the Emotiv software, TestBench 3.1.21, the electrodes at FC3/4 are labelled F3/4, F3/4 are labelled AF3/4 and FT7/8 are labelled FC5/6; this is because we adjusted the electrode placement to accommodate the concurrent setup. In this paper we refer to the electrodes according to their placement on the scalp when worn concurrent with the Neuroscan EasyCap, not the labels used in the Emotiv software.

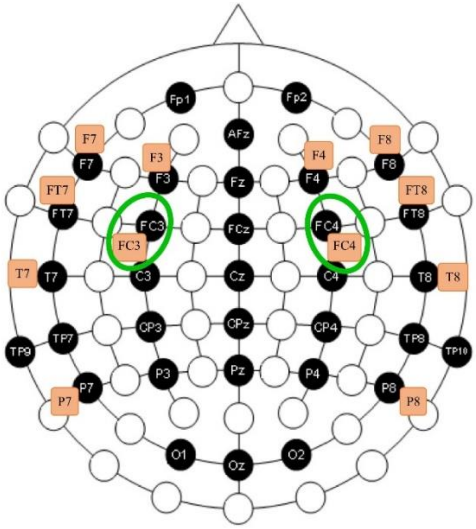

Fig. 1. Electrode position on the scalp for the Neuroscan (black circles) and EPOC+ (orange rectangles) systems. The adjacent electrodes used to calculate correlations between the two systems are circled in green.

To accurately time-lock the ERPs to the onset of the target word, we modified the EPOC+ system to incorporate event markers, following Badcock et al. (2015). We did this using a custom-made transmitter unit communicating with a custom receiver unit through infrared light (Thie, 2013). The transmitter box was connected to the audio output of the presentation computer. At the onset of each sentence and each target word, a tone of particular frequency (2400Hz for the sentence onset, 600Hz for a related target onset, and 1600Hz for an unrelated target onset) was sent to the transmitter unit. The tone could not be heard by participants (the left (tone) and right (speech) audio channels were split using a DJ splitter). This in turn activated the receiver unit, which generated an electrical pulse in the O1 and O2 channels (from which we did not acquire neural data). The length of pulse corresponded to tone frequency, allowing us to recover the condition labels from the pulse-related response in the EEG signal.

Experimental Procedure

For each participant, we set up the Neuroscan system first, and adjusted the impedances to under 5 kOhms. We then placed the EPOC+ system over the top, with cotton wool bridging scalp to sensor through custom slits in the EasyCap. EPOC+ impedances were adjusted to be below 220 kOhms in the TestBench software. Setup took up to 50 minutes, during which participants watched a DVD of their choice. Following setup, participants were seated in front of a 17-inch monitor, with speakers on both sides of the screen, at a viewing distance of about 1 metre. Before the main experiment, participants completed the PPVT 4, and after the EEG session, they completed the matrices section of the K-BIT 2.

Participants completed two EEG acquisition sessions of 20 minutes, separated by a 5-minute break. Each session included all 126 cue-target word pairs (63 related, 63 unrelated). The stimuli were presented using Psychophysics Toolbox 3 extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in MATLAB. The word pairs were presented in a pseudo-random order which was reversed in the second session. The order was optimised to minimise order bias in the sequence of related and unrelated trials, with the additional constraint that no more than 4 trials in a row were of either condition. The first session was divided into 17 blocks consisting of 5 to 15 trials, and the second section was divided into 11 blocks of 8 to 15 trials, as the participants were more familiar with the task during the second session.

To make the task more engaging for children, it was introduced in the context of a story. The child was told that they were listening to different aliens trying to learn English, and they had to rate each alien (1-5 stars) depending on its skill at producing pairs of words that went together. We asked participants to listen carefully to each pair and to decide whether the two words were related or unrelated, then give an overall judgment of the alien at the end of each block. This encouraged the children to pay attention to, and make a covert decision about, the relatedness of the words in each pair.

At the beginning of each block, an alien appeared on the screen, then moved behind a black "recording booth" in the middle of the screen. A light bulb was depicted on top of the box, and lit up during each trial, to encourage the children to pay attention and to reduce eye movements during the trial. The light bulb lit up 500 ms before the cue word, remained lit until 1500 ms after the target word onset, then turned off. The interval between the cue and target was 1000 ms. After another 1500 ms, the next trial began. At the end of each block, the alien moved out of the box and participants were prompted to grade the alien on a five-point scale regarding its overall performance. At the end of the experiment, participants were shown the "winners" of the alien contest. The total time, including the EEG setup, was about 1hr 40 minutes.

Offline EEG processing

We processed all EEG signals in EEGLAB (v13.4.4b, Delorme and Makeig, 2004) in MATLAB (R2014b). We first applied a band-pass filter between 0.1 and 40Hz, then cut the data into epochs from -100 ms to 1000 ms around target word onset. We ran an Independent Component Analysis (ICA) on all the epochs. Components with scalp distribution, frequency and timing that corresponded to eye movements and eye blinks were removed from the Neuroscan data. In line with a previous EPOC+ study with children (Badcock et al., 2015), we could not identify any eyeblink artefacts in the EPOC+ data. This may be because eye blinks were not consistent or strong enough to affect EPOC+ data. Alternatively, the ICA for the Neuroscan data could have benefited from the signal recorded by the Neuroscan electrodes recording eye movements. The EPOC+ did not have such electrodes so the ICA for the EPOC+ used only the scalp electrodes. The epochs were baseline corrected against the averaged signal from the 100 ms preceding the target onset. These data were then used for the decoding analyses. For univariate analyses, we further removed, from all electrodes, epochs with extreme

values (±150μV) in any of the electrodes of interest (see below). An average of 6 epochs were rejected from the Neuroscan data, and of 15 epochs from the EPOC+ data. For visualisation purposes, the data from the remaining trials were averaged together for each condition (related and unrelated) separately.

Group ERP analyses

The N400 is typically recorded over the centroparietal regions of the brain. We therefore focused our univariate analyses on three electrodes sites of interest: Cz, as the N400 effect is reported to be the strongest in centro-parietal sites (Kutas and Federmeier, 2011), and FC3 and FC4, which are the closest channels to Cz that we can compare between the Neuroscan and EPOC+ systems.

We first verified that our paradigm evoked a classic N400 effect at a group level, and tested whether this was detectable in both Neuroscan and EPOC+ systems. For trial-averaged waveforms from each system separately, we ran group analyses with paired t-tests comparing the two conditions at each time point from 150 ms after stimulus onset for each of the five sensors of interest. We corrected for multiple-comparisons using a statistical temporal cluster extent threshold. We calculated the temporal cluster extent threshold following Guthrie and Buchwald's (1991) method for each channel independently. Briefly, this method is based on the fact that the values at consecutive time points are correlated rather than independent, and this autocorrelation score can be calculated for each channel. We can then determine the minimum number of consecutive time points that need to show a statistical difference in a paired one-tailed t-test between the related and unrelated conditions to be considered a significant cluster at $p<0.05$ corrected (for details, see Guthrie and Buchwald, 1991). We used a one-tailed test because the direction of the N400 effect (a more negative response in the unrelated condition) was pre-specified. The first 150 ms were not tested for statistical

differences as the N400 is typically reported to occur later than 150 ms (Kutas and Federmeier, 2011), and this *a priori* constraint on the time window of interest decreased the number of -comparisons that we made.

We additionally illustrated the topographic distribution of the N400 effect, based on the Neuroscan grand average data, by subtracting activation in the related condition from the unrelated condition, and averaging over sequential 200 ms time windows (200 to 400 ms, 400 to 600 ms, 600 to 800 ms and 800 to 1000 ms).

Comparison of EEG systems

We next sought to validate the EPOC+ system for recording N400 ERPs. To compare the shape of Neuroscan and EPOC+ waveforms, we ran intraclass correlations (ICC) (Bishop et al., 2007), a global index of waveforms similarities and amplitude; and Spearman's correlations, which measure the rank correlation between two waveforms and is therefore less sensitive to amplitude. For this analysis, in order to have a fair comparison between the two systems, we re-did the pre-processing so that the Neuroscan and EPOC+ data were as comparable as possible and treated in the same way. For this the processing proceeded as describe above except that we down sampled Neuroscan data to match EPOC+'s sampling rate of 128Hz, and we did not remove eye-blink components from either system. We then calculated the correlations for each condition, using the entire epoch, at our two locations of interest where the electrodes from the two systems lie in close proximity: the left and right frontocentral sites (FC3 and FC4 - see locations on Fig. 2). We calculated the correlation for each condition, in each individual at these two locations. We examined whether correlations were significant by computing the 95% confidence interval of the group mean and checking if they overlapped with 0 (which would correspond to no correlation). Finally, we asked whether the amplitude of the N400 effect differed between the two systems. To this end, we compared the area under

the difference curve (related – unrelated ERP), using a trapezoidal integration from 300 ms to 800 ms, between the two systems using a two-tailed, paired t-test across individuals. These time points correspond to the expected N400 effect timecourse (Kutas and Federmeier, 2011).

Single subject ERP analyses

Our final goal was to assess the sensitivity of our paradigm and EEG systems to detect N400 effects in individual children. For each individual and each system, we conducted first-level (single subject) analyses using independent samples t-tests between the two conditions at the electrodes Cz (for Neuroscan only), and FC3 and FC4 (for both systems), for each time point starting at +150 ms after the target onset. Again, we used the autocorrelation score to determine the temporal cluster extent threshold for each electrode independently, to correct for multiple comparisons.

We illustrated the topographic distribution of the N400 effect in individuals based on the Neuroscan data by subtracting activation in the related from the unrelated condition, and averaging over sequential 200 ms time windows (200 to 400 ms, 400 to 600 ms, 600 to 800 ms and 800 to 1000 ms).

Group and single subject MVPA

In order to be sensitive to individual variation in the topology of N400 effects, without increasing multiple comparisons (i.e., without analysing every electrode separately), we used multivariate pattern analyses (MVPA). We analysed all the data using the CosmoMVPA toolbox (Oosterhof et al., 2016) in MATLAB. We analysed the group and individual data obtained by Neuroscan and EPOC+ separately, using a Linear Discriminant Analysis classifier. At each time point, we divided our data into a training set and a testing set, using a leave-one-target-out cross-validation approach. The training set consisted of the activation pattern across

all the electrodes for trials corresponding to all the targets but one, and the classifier was trained to find the decision boundary that best distinguished between the two categories (related vs. unrelated). Since each pair was repeated once, the training set consisted of 248 (62 stimuli * 2 conditions * 2 repetitions) trials. We then tested the classifier's ability to classify the category of the remaining four trials (two related, two unrelated) corresponding to the remaining target. We repeated this procedure 63 times, each time leaving a different target out. Finally, we averaged the accuracy of the classifier for these 63 tests to yield an accuracy value at each time point, and for each individual. The group average classifier accuracy, at each time point, was obtained by averaging each individual's accuracy. If the classifier performs significantly above chance (50%), we infer that there was information in the brain signals that differed between the two conditions (related vs unrelated words) at that time point.

For statistical inference we implemented a sign-permutation test (at the group level), or a label-permutation test (in individuals) at each time point (Maris and Oostenveld, 2007). The sign-permutation test consists of randomly swapping the sign (positive or negative after subtracting chance, 50%) of the decoding results of each of the participants. The label-permutation test consists of randomly permuting the condition label of the targets before classification to obtain classifier accuracies under the null-hypothesis. We performed 1000 permutations to obtain a null distribution at each time point. The observed (i.e., correctly labelled) decoding accuracies and permutation results were then transformed using Threshold Free Cluster Enhancement (TFCE; Smith and Nichols, 2009) which yields a statistic of cluster level support at each time point. To account for multiple comparisons across time, the maximum TFCE statistic of each permutation from across all time points was used to form a single corrected null-distribution. The observed (correctly labelled) TFCE statistic at each timepoint was considered significantly above chance if it was larger than 95% of the TFCE values in the corrected null-distribution.

Results

Behavioural results

All children had a standard score within or above the normal range (90-110) for non-verbal reasoning (K-BIT M= 123, 95% CI [114,131]) and receptive vocabulary (PPVT M= 120, 95% CI [115,126]). We asked children for a subjective rating (1-5 stars) of the performance of the aliens in each block, but as there was no 'correct' answer, we do not report accuracy. Children seemed to understand the instructions well and informally reported the task to be engaging.

Group ERP analyses

At the group level, we replicated the typical N400 effect using the Neuroscan system. We found significant N400 effects at all three of our regions of interest: Cz, FC3, and FC4 (Fig. 2, left panels). For the central location (Cz), the N400 effect was significant for a cluster of time points from 272 – 1000 ms, post-stimulus onset (Fig. 2, top panel). For FC3, the N400 effect was significant for a cluster from 292 – 1000 ms, and for FC4 the N400 effect was significant in a cluster from 302 – 1000 ms. The group-level topographic distribution of the effect (Fig. 2, bottom panel) was initially centro-parietal, spreading frontally at later time points. We were also able to record N400 effects for the group using the EPOC+ system in FC3 (from 350 to 747 ms) and FC4 (from 469 to 596, and from 684 to 739 ms), our two locations of interest (Fig. 2, right panels).

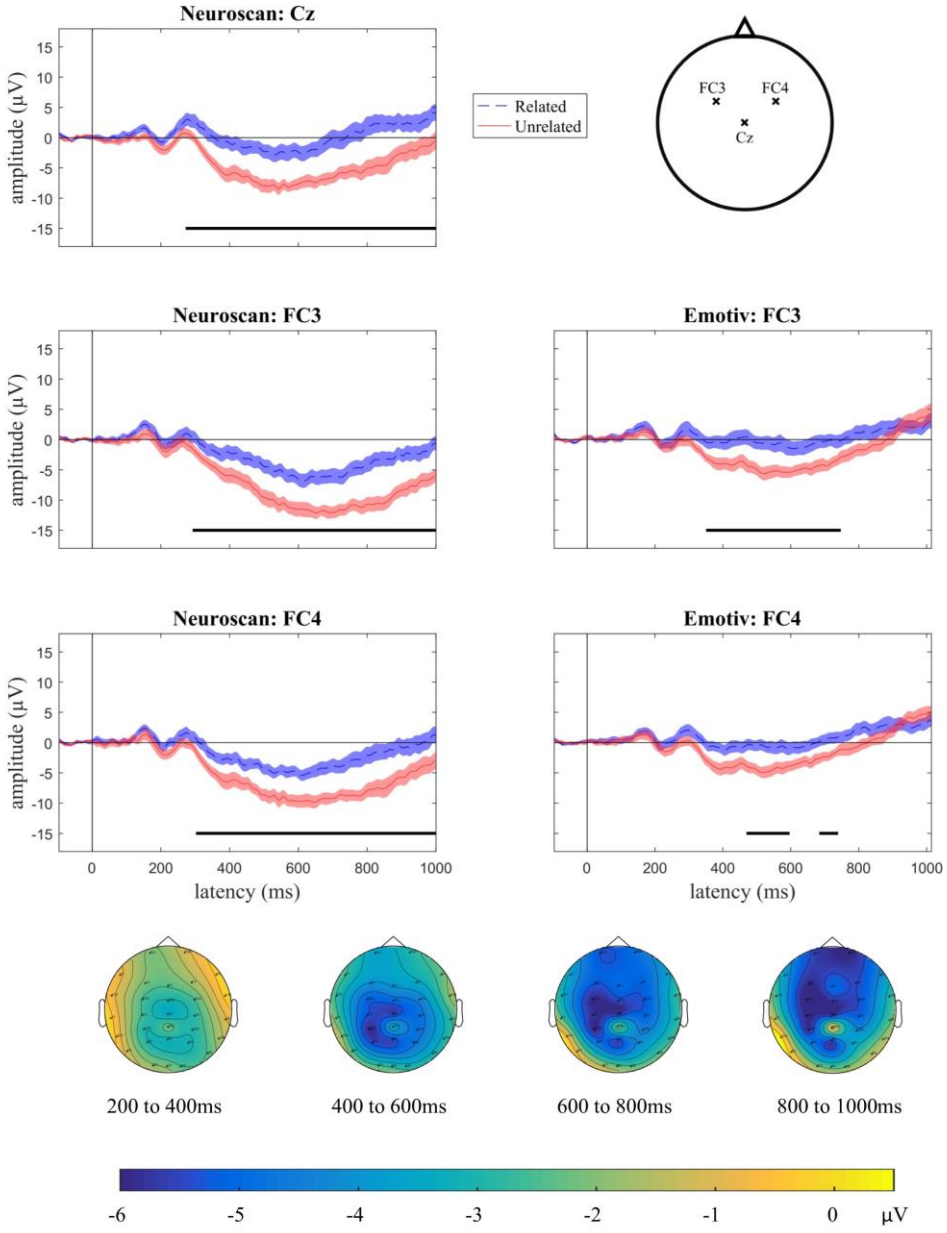

Fig. 2. Group N400 effects for Experiment 1 (word pairs). Plots display grand average ERPs (n=16), with related (dashed blue) and unrelated (solid red) conditions for Neuroscan electrodes Cz (top left panel), FC3 (middle left) and FC4 (bottom left), and EPOC+ electrodes F3 (middle right) and F4 (bottom right). Shading indicates standard error of the mean. Time points at which there was a statistical difference between the conditions are indicated with a solid black line under the plot ($p<0.05$, after cluster correction for multiple comparisons). Locations are shown on the top right panel. The bottom panel illustrates the topographic map of the N400 effect (unrelated minus related condition) from 200 to 1000 ms after target onset in the group for the Neuroscan system. Yellow colours indicate no

difference between the two conditions and blue colours indicate a more negative-going response for the unrelated condition. The N400 effect was distributed over central and centro-frontal regions.

Comparison of EEG systems

We compared the responses of the two systems directly using ICC and Spearman's rank correlation (see Table 1), and by comparing the areas under the difference curves. Waveforms across the two systems were qualitatively similar in shape, and positively correlated (all Spearman's rho $\geq$ 0.49, 95% CI not including zero for any of the comparisons). Mean ICC values ranged from 0.19 to 0.63 across the different conditions and sites. The ICC was greater than 0 (CIs did not include 0) for the related condition on both sides and for the unrelated condition on the left side, but not for the unrelated condition on the right side. We tested whether the amplitude of the effect was larger for Neuroscan compared to EPOC+. The area between the related and unrelated curves was numerically smaller for EPOC+ than for Neuroscan both in FC3 (140 for EPOC+ vs 294 for Neuroscan), and FC4 (82 for EPOC+ vs 251 for Neuroscan), but the difference was not significant (FC3: $t_{(14)} = 1.90$, $p = .0779$; FC4: $t_{(14)} = 1.74$, $p = .103$). Therefore, the ERPs recorded by the two systems were comparable in shape and not significantly different in amplitude.

Condition	Coeff.	Location	
		Left frontocentral (FC3)	Right frontocentral (FC4)
Related	ρ	0.63 [0.51, 0.75]	0.53 [0.37, 0.70]
	r	0.46 [0.31, 0.60]	0.25 [0.03, 0.48]
Unrelated	ρ	0.52 [0.33, 0.72]	0.49 [0.27, 0.71]
	r	0.27 [0.02, 0.52]	0.19 [-0.1, 0.47]
Area between curves (Arbitrary Units)		EPOC+ : 140 [-45, 324] Neuroscan : 294 [139, 448]	EPOC+ : 82 [-113, 276] Neuroscan : 251 [118, 384]

Table 1: Experiment 1 (word pairs paradigm, bottom) mean Spearman's rho (ρ) and ICC (r), and 95% confidence intervals, between waveforms simultaneously recorded with the research (Neuroscan) and gaming (EPOC+) EEG systems for the left (FC3) and right (FC4) frontocentral locations, in the semantically related and unrelated conditions. We also present the difference in area between the two condition curves between Neuroscan and EPOC+ averaged across subjects, with 95% confidence intervals.

Single subject ERP analyses

Our next goal was to assess the sensitivity to detect N400 effects in individual subjects. We defined a significant N400 effect as the presence of a statistically larger N400 in the unrelated compared to related condition (corrected for multiple comparisons across time points) in one or both of the two locations of interest that were present in both systems (FC3 or FC4). Table 2 shows the number of participants with significant N400 effects ("detection rate") at each electrode. A significant N400 effect was found at one or more electrodes of interest in 7 of the 15 (47%) participants' Neuroscan data, and in the same number of participants' EPOC+ data. Two participants showed an effect in the Neuroscan data but not the EPOC+ data, and vice versa. We show the individual waveforms recorded by the Neuroscan in Cz, which is where N400 effects are typically recorded (Fig. 3, first and fourth columns),

and in FC3, where we have both Neuroscan and EPOC+ data (Fig. 3, second, third, fifth and sixth columns); FC4 results were similar. In summary, the portable EPOC+ system had a comparable detection rate to the research grade EEG system, but the detection rate of an individual N400 effect was less than half with both systems, and a higher sensitivity would be needed for clinical applications.

Electrode	**FC3 / F3**	**FC4 / F4**	**Total (FC3 and/or FC4)**	**Cz**
Neuroscan	47%	33%	47%	47%
EPOC+	40%	40%	47%	N/A

Table 2: Experiment 1 (word pairs) detection rate (% of individuals) of statistically significant N400 effects in each of the three electrodes of interest, and the detection rate in a more lenient assessment where the effect was considered present if it occurred in either one or both of the two frontal electrodes.

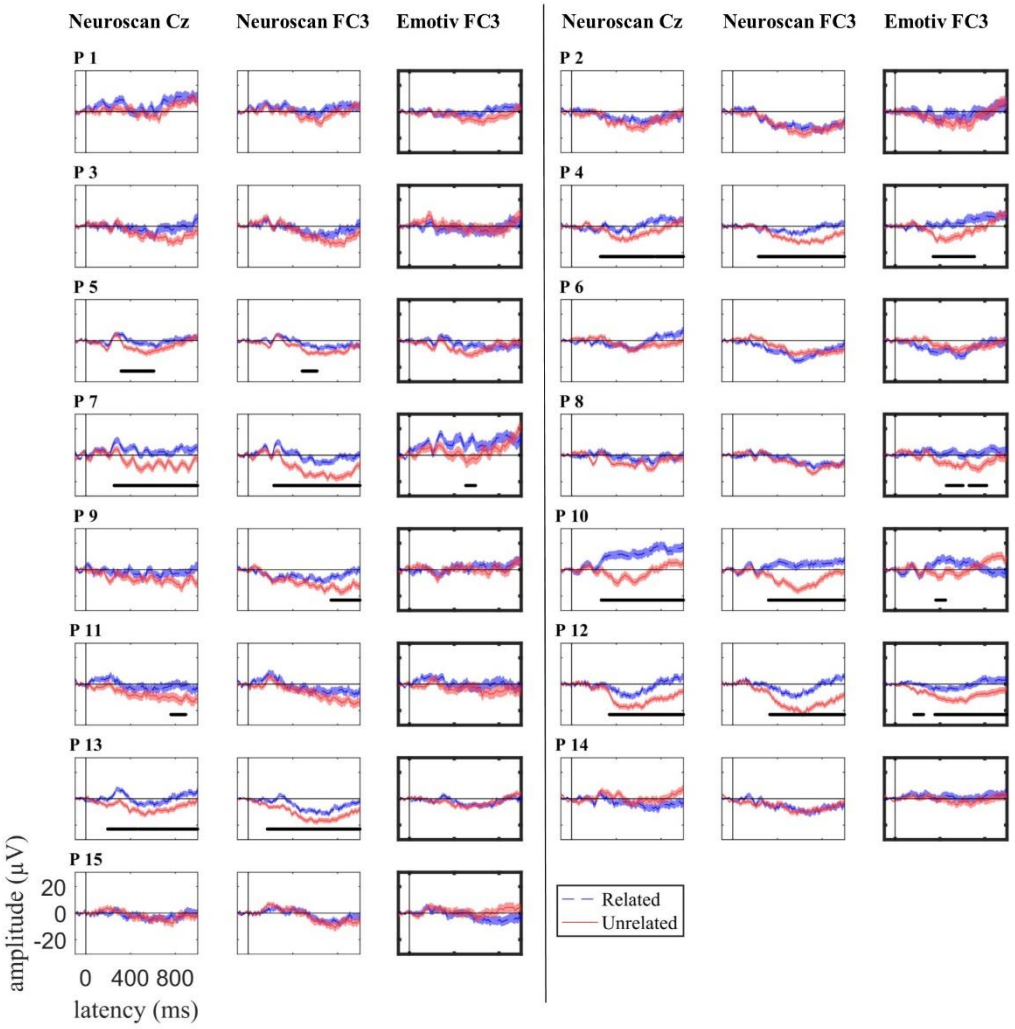

Fig. 3. Experiment 1 (word pairs paradigm) individual participant responses to target words following a related (dashed blue) and unrelated (solid red) word for Neuroscan electrode Cz (first and fourth column), and adjacent Neuroscan and EPOC+ electrode FC3 (second, third, fifth, and sixth column), plotted ± standard error (shaded area). Time points where there was a statistically significant N400 effect in each participant and sensor are indicated with a solid, horizontal, black line. EPOC+ results are outlined in bold. P indicates participant.

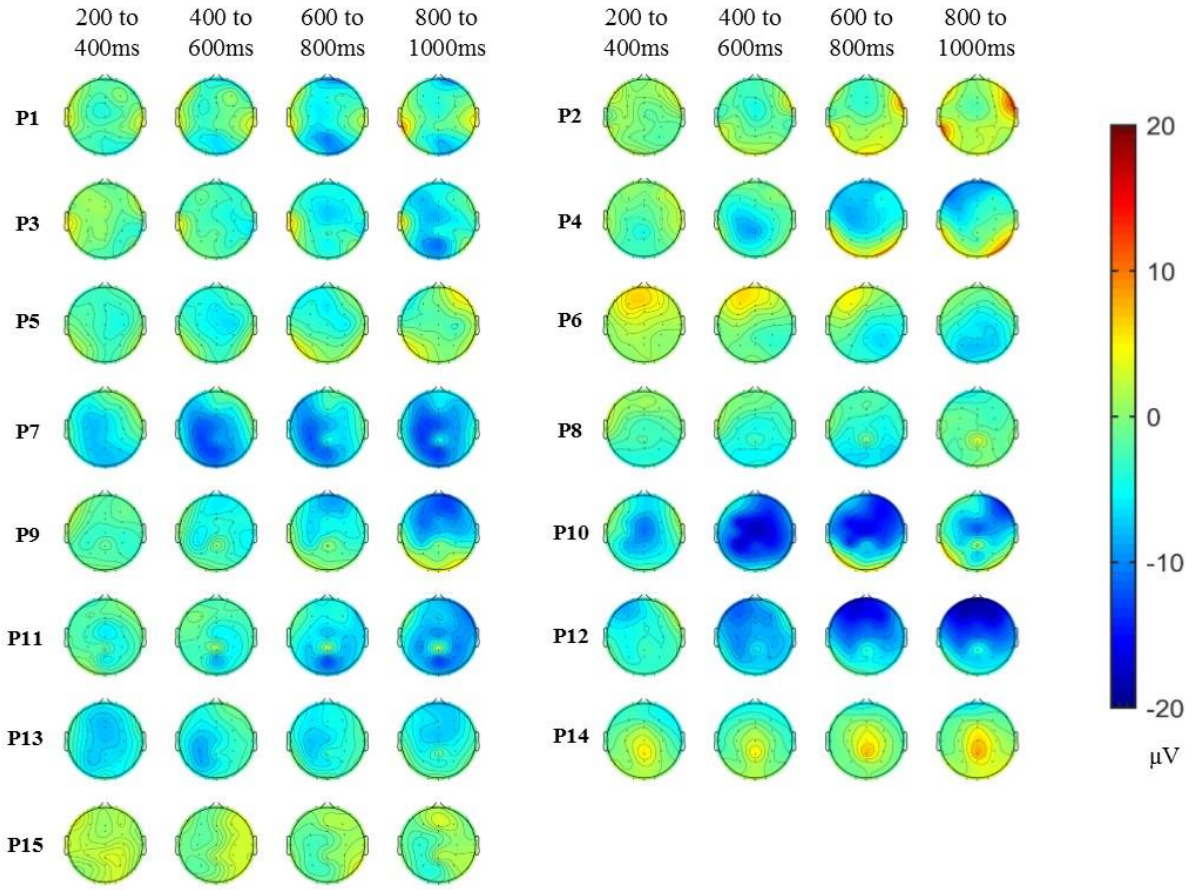

Fig 4. Experiment 1 (word pairs paradigm) individual topographic maps of the N400 effect (unrelated minus related condition) for 200 ms time windows from 200 to 1000 ms after target onset. Red areas indicate more negative-going response for the related condition and blue areas indicate more negative-going response for the unrelated condition. The topography of the N400 effect varied across individuals. P indicates participant.

Decoding analyses

We restricted our univariate analyses to three *a priori* sites of interest. However, individual topography plots (Fig. 4) suggested that the topography of the effect was highly variable between individuals, with effect location ranging from centroparietal to frontal sites. Therefore, we tested whether MVPA, which integrates information from across all sensors, would be a more sensitive method to detect differences in the EEG response to related and unrelated targets. Group level decoding performance (average over subjects) for the Neuroscan

and EPOC+ data is shown in Fig. 5. For Neuroscan (Fig. 5, purple), classifier accuracy was statistically above chance from 402 ms after target onset until the end of the epoch, indicating a reliable difference between the two conditions. There was no significant coding of associative context in the EPOC+ EEG data (Fig. 5, yellow).

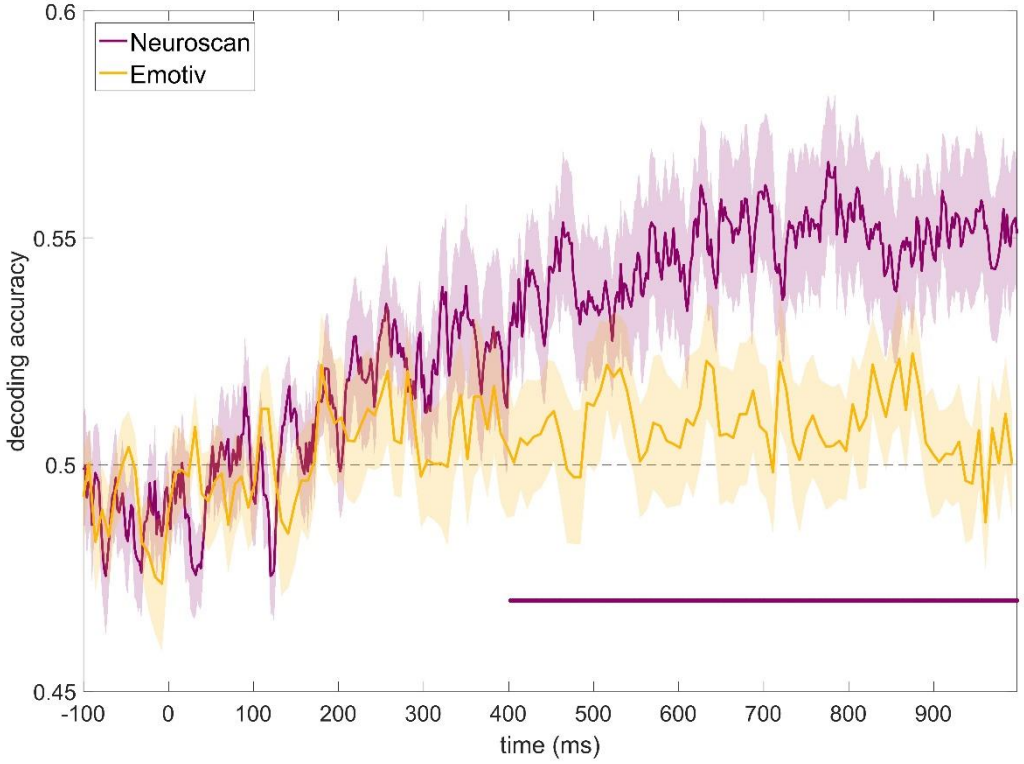

Fig. 5. Experiment 1 (word pairs) grand average decoding accuracy for discriminating between congruent and incongruent conditions over time for Neuroscan (purple) and EPOC+ (yellow) data, shown with standard error of the mean. Time points with significant decoding for Neuroscan ($p < 0.05$, assessed with TFCE permutation tests corrected for multiple comparisons, see Methods) are shown by a purple horizontal line. Decoding accuracy was significantly above chance for Neuroscan from 402 ms, but was not significant at any time point for EPOC+.

Individual decoding results are shown in Fig. 6. In the Neuroscan data, decoding was significant at some point in the epoch in 8/15 participants (53% detection rate). However, for

EPOC+, the classifier only detected a significant effect in 2 of the 15 participants (13% detection rate).

Fig. 6. Experiment 1 (word pairs) individual participant decoding accuracy for classification of congruent vs. incongruent conditions over time for Neuroscan (purple) and EPOC+ (yellow) data. Time points with accuracy significantly above chance ($p<0.05$ assessed with TFCE permutation tests corrected for multiple comparisons, see Methods) are shown as solid horizontal lines (in purple for Neuroscan, and yellow for EPOC+). Chance (50%) is indicated by the horizontal dashed line. Semantic condition could be decoded in 8/15 (53%) of individuals' Neuroscan data and in 2/15 (13%) of individual's EPOC+ data. P indicates participant.

Experiment 1 summary

We examined whether differential neural responses were elicited to identical spoken words presented in different normatively-associative contexts using two different EEG systems, Neuroscan and EPOC+. We were able to elicit N400 effects in a group of children, and to record them with a traditional EEG system, consistent with the previous literature (Berkum et al., 1999; Kutas and A. Hillyard, 1980; Friedrich and Friederici, 2004; Holcomb et al., 1992), as well as with a gaming EEG system, EPOC+. The waveforms recorded by the two EEG systems were similar.

We also sought to examine how robust N400 effects were at the individual level. Using traditional univariate approaches, we detected statistically significant individual N400 effects in some children with comparable detection rates in the two EEG systems. However, detection rates were low, relative to a desirable rate for clinical application. In two electrodes of interest, FC3 and FC4, we detected an N400 effect in 47% of our individual children, using either the Neuroscan or EPOC+ system. Multivariate pattern analyses returned quantitatively similar (53%) or weaker (13%) detection rates for Neuroscan and EPOC+, respectively.

Since our overarching aim was to derive a sensitive measure of semantic language processing for use in individual children, we next considered another avenue to increase our detection rate of N400 effects. A common way to elicit N400 effects is to present words in the context of semantically congruent or incongruent sentences. This may yield larger N400 effects than the word pairs task, as sentences provide a stronger semantic context compared to a single probe word (Kutas, 1993). Furthermore, in order to perfectly match the stimuli across conditions, we elected to repeat target words between the two conditions. However, Cruse et al. (2014) showed that this approach could lead to order effects in the case of normatively-related word pairs, which may reduce the N400 effect. Specifically, they argue that words

presented in an unrelated context first, then in a related context, produce a smaller N400 effect, impacting on the overall strength of the effect. Cruse et al. suggested that repeating target words presented in sentences is less likely to be a problem, as participants hear multiple words in each trial. For this reason, our second experiment used words presented in sentences. We also modified the task that participants performed to make it more demanding and encourage greater attention to the stimuli than for the word pairs task.

Experiment 2: congruent and incongruent sentences

Methods

Participants

Eighteen participants, aged 6 to 12 years, were recruited as described for Experiment 1. The data from two participants were excluded due to excessive artefacts in the EEG data. The final set of data thus came from 16 participants (mean age =10.3, SD=2.4, 8 male and 8 female), five of whom had also participated in Experiment 1.

Stimuli

We created two conditions: (1) *congruent sentences*, consisting of semantically correct sentences (e.g., "she wore a necklace around her *neck*"); and (2) *incongruent sentences*, consisting of sentences that ended with an anomalous word (e.g., "There were candles on the birthday *neck*"). The set of congruent sentences were based on 56 high-close probability sentences from the norms of Block and Baldwin (2010) and were chosen according to suitability for children and such that target words were high-frequency words acquired by age five (Kuperman et al., 2012). We recombined sentence stems and target words to form the set of incongruent sentences. We ensured that the incongruent target word was unexpected but grammatically correct. It also did not begin with the same phoneme or rhyme with the

corresponding congruent target word. Within one session, each sentence stem and target was used twice, once in the congruent condition and once in the incongruent condition. The final set of stimuli consisted of 56 sentences in each condition, 112 in total (see supplementary table S2 for the complete list).

Stimuli were digitally recorded by a female native British English speaker in a soundproof room and edited in Audacity®. To avoid co-articulation, the speaker recorded the sentence stems separately from the target words. This also introduced a lengthening in the final word of the sentence stem. Sentence stems and targets were combined online during stimulus presentation with a 100ms silence between the sentence frame and the target word.

EEG Equipment

The equipment and experimental setup were the same as in Experiment 1, including the completion of the matrices section of the Kaufman Brief Intelligence Test, Second Edition (K-BIT), and the Peabody Picture Vocabulary Test, Fourth Edition (PPVT). Participants who already took part in Experiment 1 did not complete these tests a second time.

Experimental procedure

Participants completed two EEG recording sessions of 25 minutes, separated by a 5-minute break. Each session included all 112 sentences (56 congruent, 56 incongruent). We presented the sentences in a pseudo-random order that was reversed in the second session. We optimised the order to avoid bias in the sequence of related and unrelated trials as described above, with all sentences presented once before being repeated in the alternate condition, and to maximise the distance between repeated presentations of the same target word. We allocated the sentences to this trial order pseudo-randomly with the additional constraint that there were at least two sentences between any repetitions of semantic content in the sentence frame or

target word. We presented an image of a satellite centrally on the screen to signal the onset of each trial, and kept this display on for the whole trial. This served as an alerting cue and encouraged children to fixate, reducing eye-movements. After 2 s, we presented the sentence through the speakers, and the satellite remained onscreen for a further 1.5 s after the presentation of the target word (Fig. 7). There was then a 2 s inter-trial-interval before the next trial. Each 20-minute session consisted of 16 blocks of 4 to 10 trials, after which children gave an answer to the experimenter (see below).

We designed a task that was strongly engaging for children while requiring minimal overt responses. It was embedded in the context of a story: an evil alien Lord had messed up some of the "messages" that we were trying to send to our extra-terrestrial friends. Participants were asked to pay attention to each sentence, and to count how many did not make sense. Accurate responses, given at the end of each block, would help "catch" an evil alien's henchman who appeared on the screen. This encouraged participants to pay attention and to make covert semantic judgments of sentences. Most of the children appeared to be highly engaged and motivated by the task, and reported that they found it entertaining. The whole experiment, including setup, took approximately 2 hours.

Offline EEG processing, ERP, and MVPA

The correlation scores, ERP analyses, and decoding analyses were performed as for Experiment 1.

Results

Behavioural results

All participants scored within the normal range for non-verbal reasoning (K-BIT score M= 111, 95% CI [101,120]) and receptive vocabulary (PPVT score M=117, 95% CI [111,122]).

Participants performed the behavioural task with a high degree of accuracy (mean percent correct: M=96.29%, SD=3.36%, range = [88.4%, 100%]), indicating that they understood the sentences, and were able to notice semantic anomalies.

Group ERP analyses

We recorded large N400 effects in the group using Neuroscan in all of the electrodes of interest (Fig. 8, left panels). We also recorded N400-like effects using the EPOC+ system at our two locations of interest, FC3 and FC4. However, these effects only reached significance at FC4 and not at FC3 possibly indicating a lesser sensitivity of the EPOC+ system (Fig. 7, right panels).

For the central location (Cz), the N400 effect started at 171 ms, and continued until 817 ms post-stimulus onset (Fig. 7, top panel). For the frontal sites, the effect started later, with a significant cluster from 409 – 697 ms for FC3 (Fig. 7, middle left panel), and a significant cluster from 387 – 875 ms for FC4 (Fig. 7, bottom left panel). For EPOC+, the N400 effect was significant in FC4 for a cluster from 418 – 752 ms. Potentials in both conditions and all three sensors also shifted in the positive direction over time, possibly corresponding to the Closure Positive Shift, an ERP component reflecting the processing that occurs at a prosodic boundary (Steinhauer and Friederici, 2001).

The topographic distribution of the N400 effect is shown in Fig. 7, bottom panel. The distribution was centro-frontal, with a slight right bias. This is in line with some previous reports that the N400 effect may be more frontal in children than in adults (Friedrich and Friederici, 2004; Henderson et al., 2011), although we did not observe this in Experiment 1.

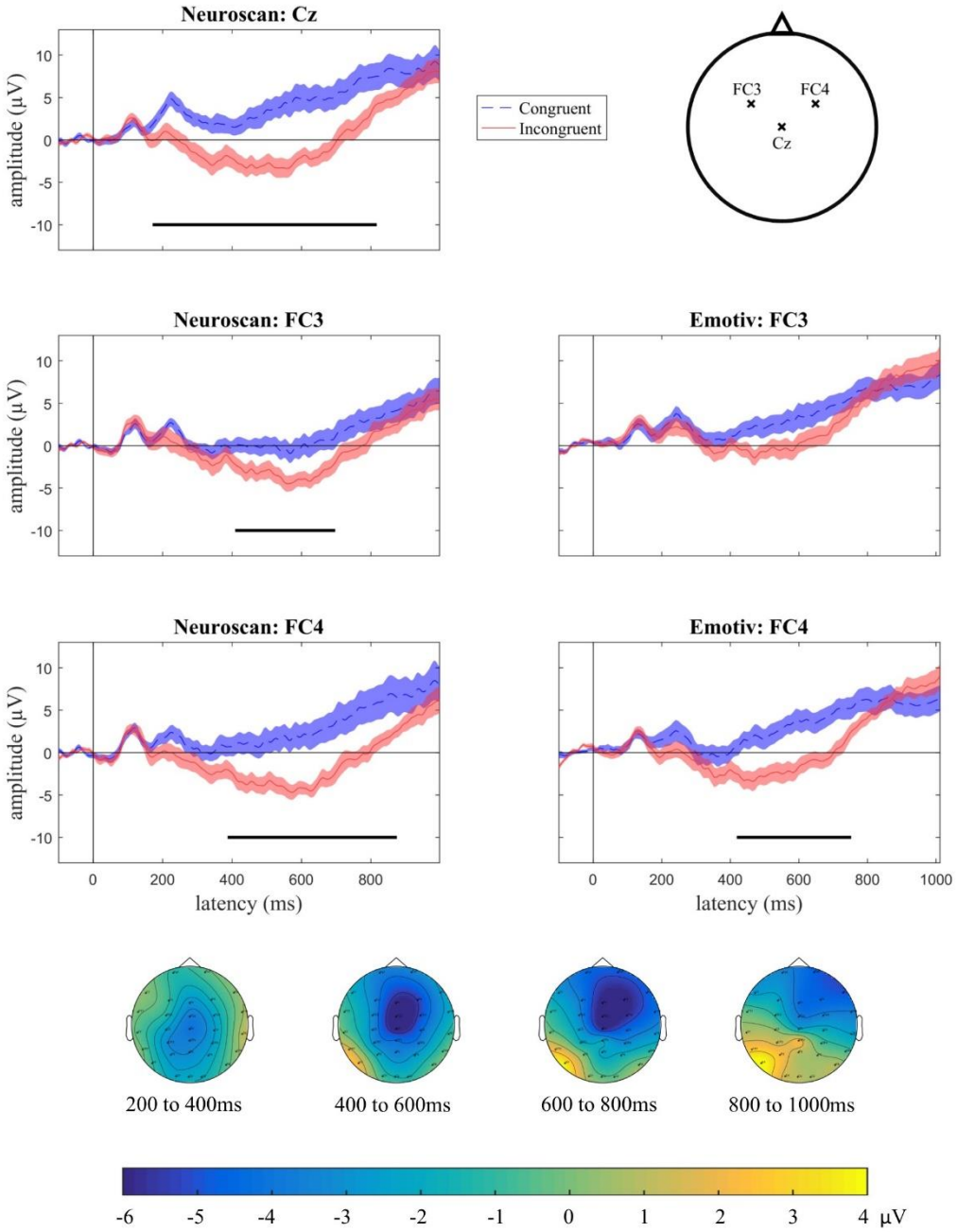

Fig. 7. Group N400 effects for Experiment 2 (sentence paradigm). Plots show grand average ERPs (n=16), with congruent (dashed blue) and incongruent (solid red) conditions for Neuroscan electrodes Cz (top left panel), FC3 (middle left) and FC4 (bottom left), and EPOC+ electrodes F3

(middle right) and F4 (bottom right). Shading indicates standard error of the mean. Temporal clusters at which there was a statistical difference between the conditions are indicated with a solid black line ($p<0.05$, after cluster correction for multiple comparisons). The bottom panel illustrates the topographic map of the N400 effect (incongruent minus congruent condition) from 200 to 1000 ms after target onset in the group (n=16). Yellow areas indicate a more negative-going response for the congruent condition and blue areas indicate a more negative-going response for the incongruent condition. The N400 effect was mainly distributed over the centro-frontal region.

Comparison of EEG systems

We again compared the responses of the two systems directly using ICC and Spearman's rank correlation and by comparing the area under the difference curve, at our two locations of interest (see Table 3). Waveforms for the sentences task across the two systems were qualitatively similar in shape and amplitude, as indicated by positive correlations (Spearman's rho ≥ 0.54, ICC ≥ 0.23 CI not including zero for any of the comparisons) for both sites and conditions. The area between the related and unrelated curves was again numerically smaller for EPOC+ than for Neuroscan both in FC3 (125 arbitrary units for EPOC+ vs 208 for Neuroscan), and FC4 (196 for EPOC+ vs 340 for Neuroscan), but the difference was not significant for either site (FC3: $t_{(15)} = 0.82$, p= .43, FC4: $t_{(15)}=1.44$, p= .17).

Condition	Coeff.	Location	
		Left frontocentral (FC3)	Right frontocentral (FC4)
Congruent	ρ	0.57 [0.38, 0.76]	0.54 [0.32, 0.77]
	r	0.23 [0.01, 0.46]	0.30 [0.02, 0.58]
Incongruent	ρ	0.71 [0.58, 0.85]	0.72 [0.58, 0.87]
	r	0.53 [0.34, 0.72]	0.62 [0.44, 0.80]
Area between curves (Arbitrary Units)		EPOC+ : 125 [-39, 290] Neuroscan : 208 [42, 373]	EPOC+ : 196 [52, 339] Neuroscan : 340 [138, 541]

Table 3: Experiment 2 (sentence paradigm) mean Spearman' (ρ) and ICC (r) and 95% confidence intervals between waveforms simultaneously recorded with the research (Neuroscan) and gaming (EPOC+) EEG systems for the left (FC3) and right (FC4) frontocentral locations, in the semantically congruent and incongruent conditions. We also present the difference in area between the two conditions between Neuroscan and EPOC+ averaged across subjects with 95% confidence intervals.

Single subject ERP analyses

We observed reliable N400 effects in one or both of FC3 and FC4 in 56% of the participants with the Neuroscan data, and 50% in the EPOC+ data (individual electrode rates shown in Table 4).

Electrode	**FC3**	**FC4**	**Total (FC3 and/or FC4)**	**Cz**
Neuroscan	38%	50%	56%	50%
EPOC+	38%	44%	50%	N/A

Table 4: Experiment 2 (sentence paradigm) detection rate (% of individuals) of statistically significant N400 effects in each of the three electrodes of interest, and the detection rate in a more lenient assessment where the effect was considered present if it occurred in either one or both of the two frontal electrodes.

Figure 8 shows individual participant waveforms for Neuroscan electrode Cz (Fig. 8, first and fourth column), and for Neuroscan (Fig. 8, second and fifth column) and EPOC+ at FC4 (Fig. 8, third and sixth column), which was the site with the highest detection rate. We again found that the topographical distribution of the effect was variable across individuals (Fig. 9).

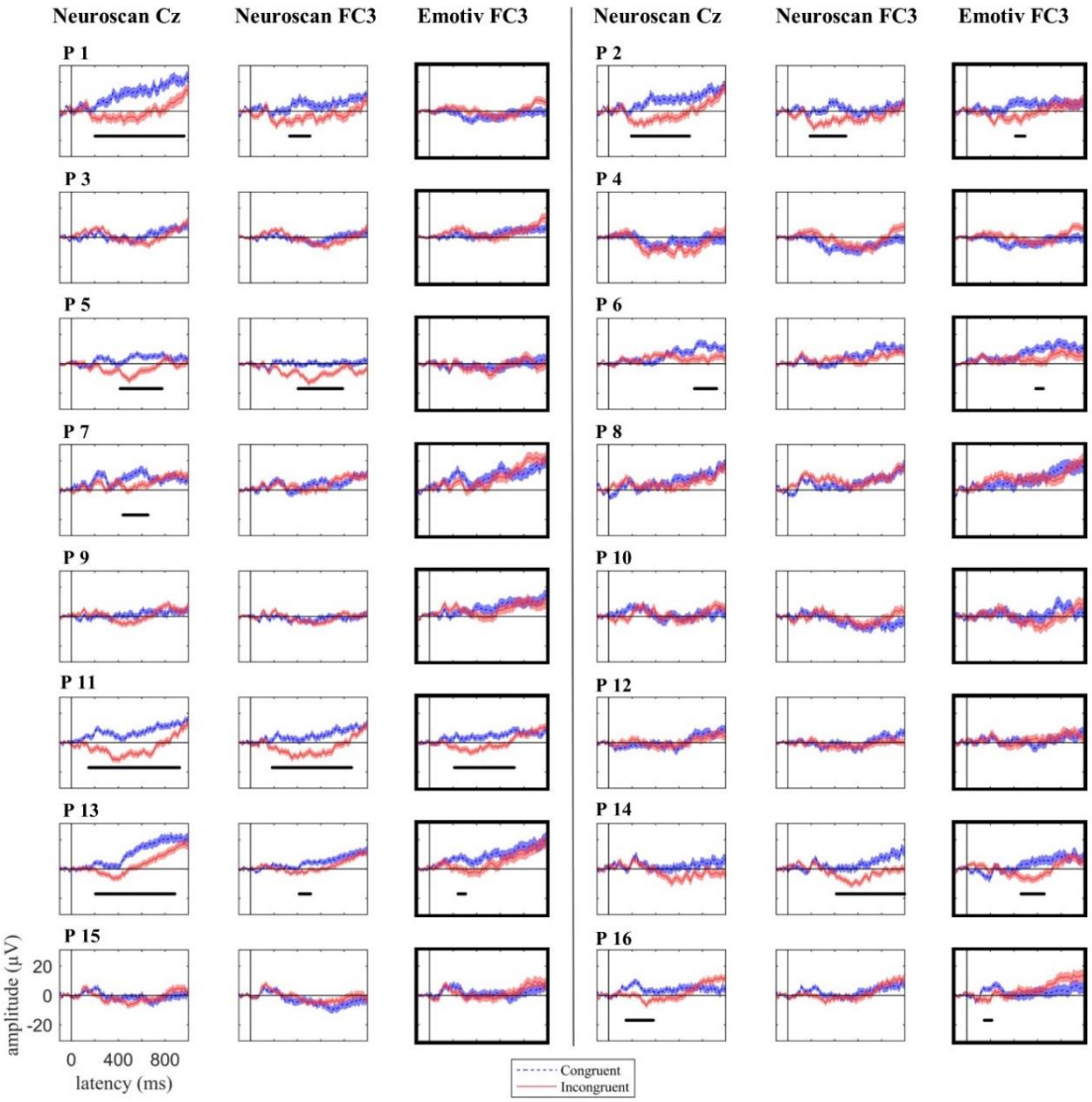

Fig. 8. Experiment 2 (sentence paradigm) individual responses to target words in congruent (dashed blue) and incongruent (solid red) sentences for Neuroscan electrode Cz (first and fourth column), and adjacent Neuroscan and EPOC+ electrode FC4, plotted ± standard error (second, third, fifth, and sixth column). Statistical N400 effect is shown as a solid black line. EPOC+ results are outlined in bold. P indicates participant.

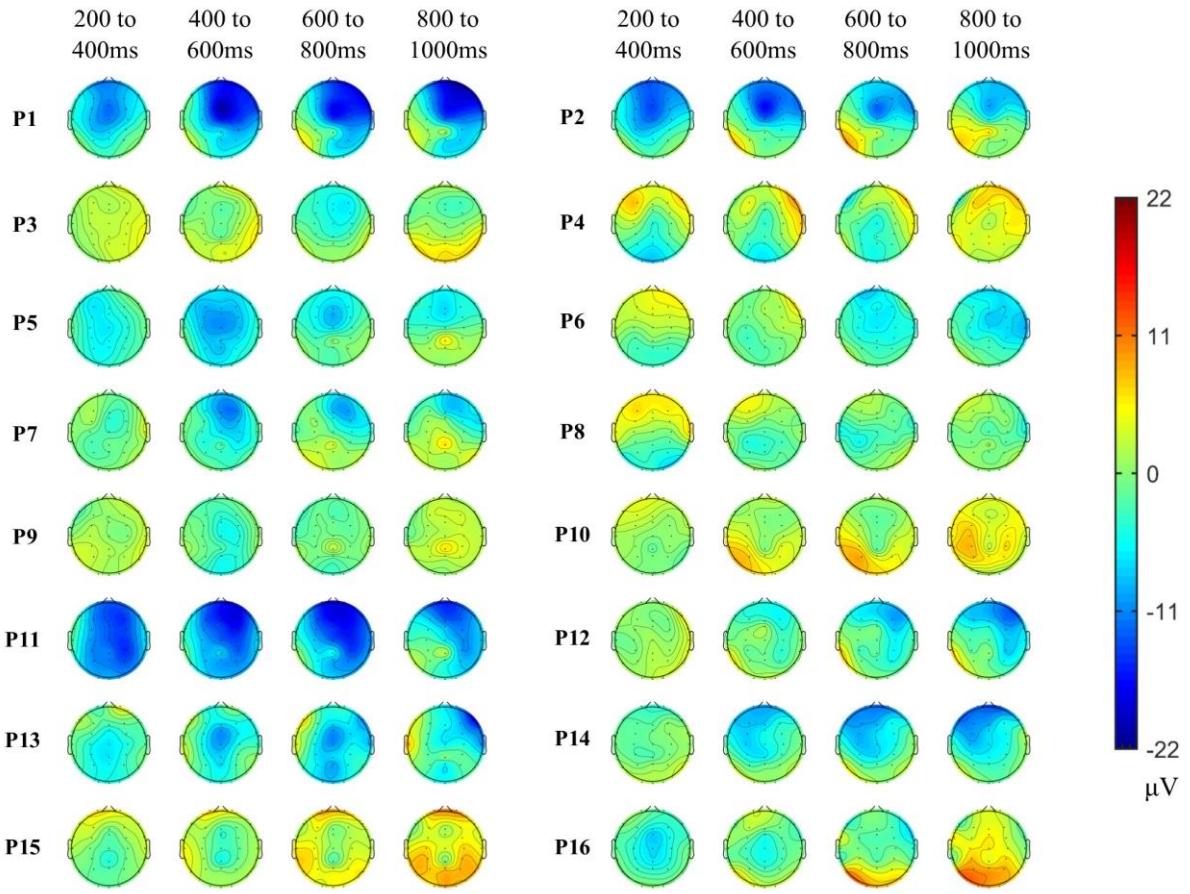

Fig 9. Experiment 2 (sentence paradigm) individual participant topographic maps of the N400 effect (incongruent minus congruent condition) for 200 ms time windows from 200 to 1000 ms after target onset. Red areas indicate more negative-going response for the congruent condition and blue areas indicate more negative-going response for the incongruent condition. The N400 effect location was variable across individuals. P indicates participant.

Decoding analyses

We again tested whether our detection rate would improve by combining data across sensors using MVPA. At the group level, we saw significant decoding of semantic context (congruent or incongruent sentence frames) in the Neuroscan data (Fig 10, top panel) from 162 ms to 1000 ms. We also decoded the semantic category from the EPOC+ group level data (Fig. 10, bottom panel), at several clusters of time points between 414 ms and 860 ms. This matched

the timecourse of univariate decoding seen at the group level in Neuroscan and EPOC+ data (Fig 7).

At the individual level, we decoded semantic category from the Neuroscan data in all but two participants (88% detection, Fig 11, first and third columns). This was a marked improvement relative to the univariate detection rate at individual channels (38-50%, above). However, for EPOC+, the classifier only detected a significant effect in 4 of the 16 participants (25% detection rate, Fig 11, second and fourth columns). We examined whether the superior decoding of Neuroscan could be attributed to the larger number of electrodes (33 in Neuroscan versus 12 in EPOC+), by performing an additional MVPA analysis on the Neuroscan data using only the 12 electrodes closest to the EPOC+ ones. In these conditions, the MVPA again performed well for the Neuroscan data, identifying statistical differences in the signal for 88% of participants. Thus, the Neuroscan data appear to be more suitable for decoding than the EPOC+ data, and the difference cannot be attributed to the difference in the number or location of electrodes.

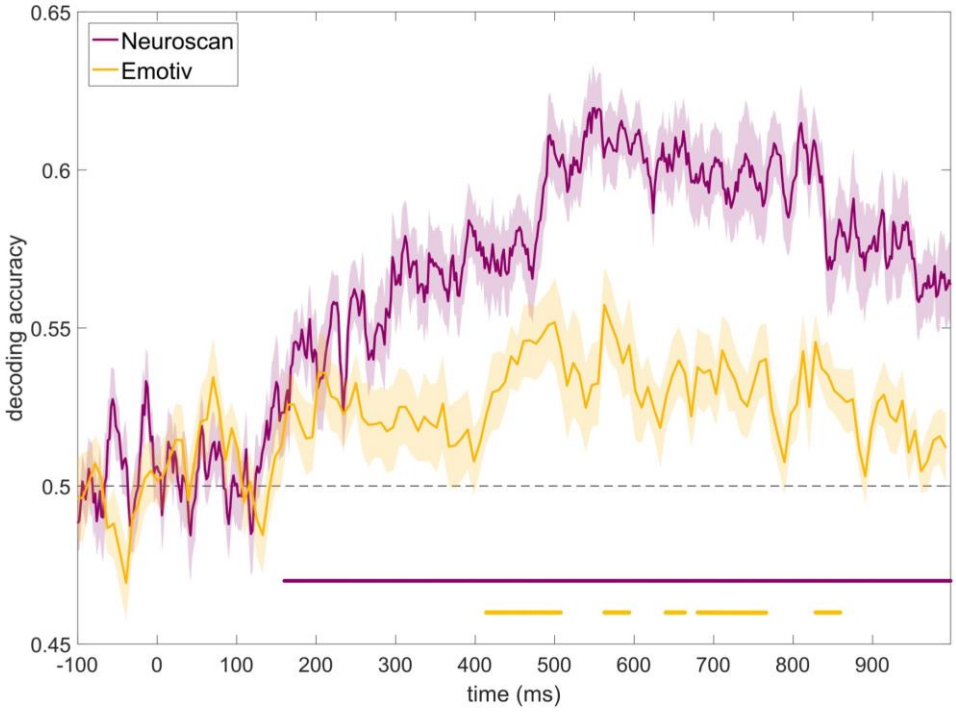

Fig. 10. Experiment 2 (sentence paradigm) grand average decoding accuracy for discriminating between identical target words presented in congruent and incongruent conditions at each time point for Neuroscan (purple) and EPOC+ (yellow) data. Shading indicates standard error of the mean. Clusters of significant decoding are shown by a purple (Neuroscan) and yellow (EPOC+) horizontal line. Decoding accuracy was significantly above chance for Neuroscan in a cluster from 162 ms to 1000ms, and for EPOC+ at several clusters of time points between 414 ms and 860 ms.

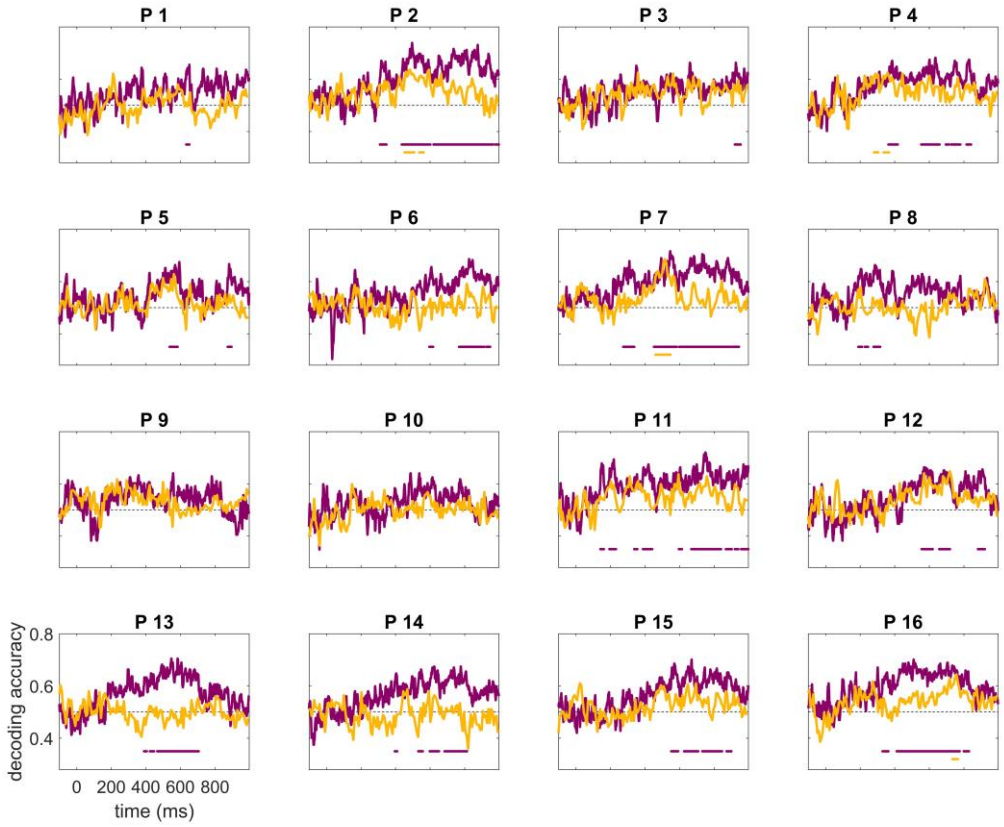

Fig. 11. Experiment 2 (sentence paradigm) individual decoding accuracy for discriminating between congruent and incongruent conditions over time for Neuroscan (purple) and EPOC+ (yellow) data. Time points with accuracy significantly above chance (temporal cluster correction, $p<0.05$) are shown as solid horizontal lines (in purple for Neuroscan, and yellow for EPOC+). Semantic condition could be decoded in 88% of the participants' Neuroscan data and in 25% of the participants' EPOC+ data. P indicates participant.

Experiment 2 summary

Using a paradigm that contrasted congruent and incongruent sentences, we elicited an N400 effect in a group of children, and recorded it using both the Neuroscan and the EPOC+ EEG systems. The shape and amplitude of the waveforms recorded by the two EEG systems were similar.

We also tested the sensitivity of our paradigm at the individual subject level. We elicited a statistically significant N400 effect in up to 50% of the participants using Neuroscan, and 44% of the participants using EPOC+ in the right frontocentral location.

Using multivariate analyses, we decoded the semantic condition in all but two individuals (88%) using the Neuroscan data. This suggests that multivariate analyses, which take the pattern across all electrodes into account, may be a sensitive way to detect effects in individuals, accounting for the topographic variability of effects across individuals. However, decoding was, if anything, slightly worse than univariate analyses for the EPOC+ system (25% with MVPA, as opposed to 50% with univariate analyses), suggesting that MVPA may be more sensitive to the differences in data quality from research-grade and gaming EEG systems than traditional analyses.

General Discussion

We sought to examine the reliability and sensitivity of N400 paradigms as an avenue for inferring language processing from brain data in children. We developed two child-friendly N400 paradigms and compared the signal recorded by a low-cost gaming EEG system, Emotiv EPOC+, to that recorded by a traditional research EEG system, Neuroscan SynAmps2. We used two approaches to evaluate the electrophysiological response: univariate analysis of the N400 effect, and multivariate decoding. We found that the N400 effect could be observed at the group level, and detected at an individual level in about half of the children, using either paradigm (word pairs or sentences) and using either EEG system. The best individual participant detection rate, 88%, was given by multivariate analyses of the data recorded with Neuroscan in the sentences paradigm.

Despite an extensive body of literature on the N400 effect, only a few studies have carried out statistical testing at the individual subject level or report individual participants'

waveforms. In our data, statistically significant N400 effects were present in about half of our participants using either normatively associated word pairs or sentences to induce a violation of lexical-semantic predictions. This may seem low, given that the N400 is widely assumed to be a large and robust effect (e.g., Kutas and Federmeier, 2011), but is in fact similar to statistical detection rates in adults using similar tasks (Cruse et al 2014). This raises an interesting question about how robust and prevalent individual N400 effects actually are. Further work is needed to establish whether a failure to detect N400 effects in some individuals reflects true inter-subject variation in the neural response to lexical-semantic violations, or simply a lack of sensitivity to detect neural effects within testing constraints.

The details of the stimuli and the participants' task seem likely to affect the size of the N400 (e.g., Bentin et al., 1993; Chwilla et al., 1995; Ortega et al., 2008; Perrin and García-Larrea, 2003), and therefore, presumably, individual-subject detection rate. For example, Cruse et al (2014) reported superior individual-subject detection rates for active and overt tasks relative to passive ones. Cruse et al also reported superior rates for normatively associated word pairs (50%) than for sentences (17%) when heard passively. However, we found that the response to the two stimulus sets were comparable, when we considered an effect recorded from either of our sites of interest (47% in word pairs, and 56% in sentences). Our relatively high detection rate, particularly for sentences, is potentially attributable to our engaging covert tasks which required participants to attend to the sematic relatedness of the target word and encouraged participants to remain attentive throughout the experiment.

Our individual-participant data also emphasise the variability in topography and timecourse of individual N400 responses (previously suggested by Henderson et al. (2011)), which may be particularly important to consider when testing children and special populations, and might necessitate different analysis approaches. For example, in those participants not showing a reliable N400 effect, it is possible that univariate analyses failed to detect more

subtle changes in the neural responses, or that N400-like responses occurred in EEG channels that were not analysed. To overcome this, we employed MVPA to integrate information from across all the sensors. MVPA provides a sensitive method of assessing effects occurring at any location or combination of locations without increasing type I error (only one statistical test is performed on the pattern of response across all sensors) at the cost of specificity regarding where the effect arises from (e.g. Grootswagers et al., 2017). We ran our individual-subject multivariate analyses at every time point (with correction for multiple comparisons) allowing us to be sensitive to individual variability in both topology and timecourse.

With this approach, semantic condition could be decoded from the Neuroscan data in all but two individuals (88% detection rate, Experiment 2). In our pursuit of an individualised neural marker of language comprehension, this is encouraging. It also confirms the intuition that considering the signal recorded by all electrodes can be more powerful than restricting analysis to one or a few. Even though the N400 effect is well established as having a centro-parietal topography at the group level, our data suggest that there is variation in topology at the individual subject level, and we anticipate that the variation may be even greater amongst minimally-verbal children with autism. Moreover, we do not want to be too restrictive *a priori* because ultimately, any statistically reliable decodable difference between the brain responses to identical auditory tokens presented in different semantic contexts will be meaningful for our purpose, independently of *where* (and to a certain extent *when*) this effect occurs.

Finally, we assessed the sensitivity of low-cost portable EEG technology, Emotiv EPOC+ EEG for detecting N400 effects. EPOC+ was able to record N400 effects at the group level and showed a similar detection rate to the research system at the individual subject level. Moreover, when we formally compared the data between systems, we found that the waveforms recorded by EPOC+ were similar in shape and amplitude to those recorded by the research system Neuroscan. This makes it a promising avenue for research, particularly where

full EEG setups are unfeasible. Our result adds to the previous literature demonstrating the usability of the EPOC+ to record early ERPs (Badcock et al., 2015, 2013; de Lissa et al., 2015; Duvinage et al., 2013), and shows that late ERPs such as the N400 can also be recorded with a low-cost, portable system. Nonetheless the effects recorded with EPOC+ tended to be numerically lower amplitude than for Neuroscan, and detection rate did not improve with MVPA with the best detection rate for Neuroscan (88%, MVPA, Exp 2) far above the best with EPOC+ (50%, univariate, Exp 2).

A few limitations of the EPOC+ system could be mitigated in future research. First, due to a limitation of the software (EPOC+ Testbench software), the impedance of the EPOC+ electrodes was not assessed precisely, and we were only able to ensure that it remained lower than 220kOhm. It is thus likely that impedances were much higher for EPOC+ than for Neuroscan (for which impedances were adjusted to < 5 kOhms), although the differences between the systems (e.g., active vs. passive electrodes respectively) makes differences in impedance difficult to interpret. Use of a more precise measure of impedance to ensure the best possible connection in every participant would be beneficial. Second, the N400 effect was centrally distributed on the scalp, at least at the group level, in accordance with previous literature (Friedrich and Friederici, 2004; Kutas and Federmeier, 2011), and the Emotiv EPOC+ headset does not have any central sensors. In future research, researchers should consider wiring an additional electrode in a centro-parietal location where the effects were the largest. It is also important to consider the trade-off between the sensitivity and the reliability we want to achieve (in which case Neuroscan may be more suitable), and the level of portability and accessibility (in which case EPOC+ is more suitable). In particular, when testing children or special populations, the possibility of recording EEG outside of the lab, with an easy and fast setup procedure, may outweigh EPOC+'s apparently lower sensitivity in the context of multivariate analyses.

Taken together, our results indicate that it may be possible to index lexical-semantic processing in individual children using EEG. However, contrasting the results we obtained from different paradigms, EEG systems, and analysis methods, several trade-offs have to be considered. We found that using MVPA with data recorded by a research-grade EEG system in response to identical spoken words presented in the context of congruent and incongruent sentences (Experiment 2) yielded the best individual-subject detection rate of differential brain responses driven by lexical-semantic congruency (significant in 88% of individuals). Our variable individual subject data emphasise the importance of analysing and reporting individual ERP results, in addition to the grand average data, to illustrate the variability in the presence, location, and timing of ERPs.

Conclusion

In this study, we set out to devise a paradigm and analysis approach that elicited a reliably different neural signal to identical spoken language tokens presented in different lexical-semantic contexts, and to test whether we could use an inexpensive portable device, EPOC+, to measure it. Future research can build on this work to develop a method for reliably detecting language comprehension in people who struggle to overtly indicate their understanding.

Our results suggest that the sentence paradigm may be most suitable to evoke differential responses based on semantic congruency. We were able to replicate group-level N400 effects in typically-developing children using both the EPOC+ and the research grade Neuroscan system. At an individual level, there was considerable variability but nonetheless we have promising indicators that an individual-level N400 can be evoked in about half of typically developing children using either Neuroscan or EPOC+ systems. MVPA analyses further allowed us to reach near-perfect detection rate with Neuroscan EEG, with only two

participants not showing a reliable electrophysiological signature to semantic anomalies in sentences. This gives us a firm basis for future research in which we can test for receptive language processing in people who are unable to communicate.

Author contributions

Conceptualisation: A.W., A.N.R, N.A.B, J.B., L.N.; Stimulus and task development: A.W., N.A.B., A.R, D.M., N.D., S.Y., E.S., L.N., S.P.; Data acquisition: S.P., A.W., N.A.B.; Formal analysis: S.P., A.W., T.G.; Writing (original draft): S.P; Writing (reviewing and editing): all authors; Supervision: A.W.

Acknowledgments

This work was funded by an Australian Research Council (ARC) Centre of Excellence Neural Markers Training Scheme grant to A.W., N.A.B., A.N.R., J.B and L.N. A.W. was supported by an ARC Future Fellowship (FT170100105) and MRC intramural funding SUAG/035/RG91365. L.N. was supported by an ARC Future Fellowship (FT120100102).

We thank Polly Barr and Nickolas Williams for help with data acquisition.

References

Audacity(R) software is copyright (c) 1999-2014 Audacity Team. The name Audacity(R) is a registered trademark of Dominic Mazzoni.

Badcock, N.A., Mousikou, P., Mahajan, Y., de Lissa, P., Thie, J., McArthur, G., 2013. Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. PeerJ 1, e38. https://doi.org/10.7717/peerj.38

Badcock, N.A., Preece, K.A., Wit, B. de, Glenn, K., Fieder, N., Thie, J., McArthur, G., 2015. Validation of the Emotiv EPOC EEG system for research quality auditory event-related potentials in children. PeerJ 3, e907. https://doi.org/10.7717/peerj.907

Barham, M.P., Clark, G.M., Hayden, M.J., Enticott, P.G., Conduit, R., Lum, J.A.G., 2017. Acquiring research-grade ERPs on a shoestring budget: A comparison of a modified Emotiv and commercial SynAmps EEG system. Psychophysiology 54, 1393–1404. https://doi.org/10.1111/psyp.12888

Bentin, S., Kutas, M., Hillyard, S.A., 1993. Electrophysiological evidence for task effects on semantic priming in auditory word processing. Psychophysiology 30, 161–169. https://doi.org/10.1111/j.1469-8986.1993.tb01729.x

Berkum, J.J.A. v, Hagoort, P., Brown, C.M., 1999. Semantic Integration in Sentences and Discourse: Evidence from the N400. J. Cogn. Neurosci. 11, 657–671. https://doi.org/10.1162/089892999563724

Block, C.K., Baldwin, C.L., 2010. Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. Behav. Res. Methods 42, 665–670. https://doi.org/10.3758/BRM.42.3.665

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Int.

Borovsky, A., Elman, J.L., Kutas, M., 2012. Once is Enough: N400 Indexes Semantic Integration of Novel Word Meanings from a Single Exposure in Context. Lang. Learn. Dev. Off. J. Soc. Lang. Dev. 8, 278–302. https://doi.org/10.1080/15475441.2011.614893

Brainard, D.H., 1997. The Psychophysics Toolbox. Spat. Vis. 10, 433–436.

Cantiani, C., Choudhury, N.A., Yu, Y.H., Shafer, V.L., Schwartz, R.G., Benasich, A.A., 2016. From Sensory Perception to Lexical-Semantic Processing: An ERP Study in Non-Verbal Children with Autism. PLOS ONE 11, e0161637. https://doi.org/10.1371/journal.pone.0161637

Chwilla, D.J., Brown, C.M., Hagoort, P., 1995. The N400 as a function of the level of processing. Psychophysiology 32, 274–285. https://doi.org/10.1111/j.1469-8986.1995.tb02956.x

Cruse, D., Beukema, S., Chennu, S., Malins, J.G., Owen, A.M., McRae, K., 2014. The reliability of the N400 in single subjects: Implications for patients with disorders of consciousness. NeuroImage Clin. 4, 788–799. https://doi.org/10.1016/j.nicl.2014.05.001

Davis, C.J., 2005. N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. Behav. Res. Methods 37, 65–70. https://doi.org/10.3758/BF03206399

de Lissa, P., Sörensen, S., Badcock, N., Thie, J., McArthur, G., 2015. Measuring the face-sensitive N170 with a gaming EEG system: A validation study. J. Neurosci. Methods 253, 47–54. https://doi.org/10.1016/j.jneumeth.2015.05.025

Dunn, L.M., Dunn, D.M., 2007. PPVT-4: Peabody picture vocabulary test. Pearson Assess.

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., Dutoit, T., 2013. Performance of the Emotiv Epoc headset for P300-based applications. Biomed. Eng. OnLine 12, 56. https://doi.org/10.1186/1475-925X-12-56

Elsawy, A.S., Eldawlatly, S., Taher, M., Aly, G.M., 2014. Performance analysis of a Principal Component Analysis ensemble classifier for Emotiv headset P300 spellers, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Presented at the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5032–5035. https://doi.org/10.1109/EMBC.2014.6944755

Friedrich, M., Friederici, A.D., 2005. Semantic sentence processing reflected in the event-related potentials of one- and two-year-old children. Neuroreport 16, 1801–1804.

Friedrich, M., Friederici, A.D., 2004. N400-like Semantic Incongruity Effect in 19-Month-Olds: Processing Known Words in Picture Contexts. J. Cogn. Neurosci. 16, 1465–1477. https://doi.org/10.1162/0898929042304705

Giacino, J.T., Smart, C.M., 2007. Recent advances in behavioral assessment of individuals with disorders of consciousness: Curr. Opin. Neurol. 20, 614–619. https://doi.org/10.1097/WCO.0b013e3282f189ef

Grootswagers, T., Wardle, S.G., Carlson, T.A., 2017. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time-series neuroimaging data. J. Cogn. Neurosci. 29, 677–697. https://doi.org/10.1162/jocn_a_01068

Guthrie, D., Buchwald, J.S., 1991. Significance testing of difference potentials. Psychophysiology 28, 240–244.

Haynes, J.-D., 2015. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron 87, 257–270. https://doi.org/10.1016/j.neuron.2015.05.025

Hebart, M.N., Baker, C.I., 2018. Deconstructing multivariate decoding for the study of brain function. NeuroImage 180, 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005

Henderson, L.M., Baseler, H.A., Clarke, P.J., Watson, S., Snowling, M.J., 2011. The N400 effect in children: Relationships with comprehension, vocabulary and decoding. Brain Lang. 117, 88–99. https://doi.org/10.1016/j.bandl.2010.12.003

Kaufman, A.S., Kaufman, N.L., 2004. Kaufman brief intelligence test KBIT 2 ; manual.

Kiang, M., Patriciu, I., Roy, C., Christensen, B.K., Zipursky, R.B., 2013. Test-retest reliability and stability of N400 effects in a word-pair semantic priming paradigm. Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol. 124, 667–674. https://doi.org/10.1016/j.clinph.2012.09.029

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., 2007. What's new in psychtoolbox-3. Perception 36, 1–16.

Kuperman, V., Stadthagen-Gonzalez, H., Brysbaert, M., 2012. Age-of-acquisition ratings for 30,000 English words. Behav. Res. Methods 44, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kutas, M., 1993. In the company of other words: Electrophysiological evidence for single-word and sentence context effects. Lang. Cogn. Process. 8, 533–572. https://doi.org/10.1080/01690969308407587

Kutas, M., A. Hillyard, S., 1980. Kutas, M. & Hillyard, S. A. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207, 203-205. Science 207, 203–5. https://doi.org/10.1126/science.7350657

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Oosterhof, N.N., Connolly, A.C., Haxby, J.V., 2016. CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front. Neuroinformatics 10, 27. https://doi.org/10.3389/fninf.2016.00027

Ortega, R., López, V., Aboitiz, F., 2008. Voluntary modulations of attention in a semantic auditory-visual matching Task: an ERP study. Biol. Res. 41, 453–460. https://doi.org/10.4067/S0716-97602008000400010

Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10, 437–442.

Perrin, F., García-Larrea, L., 2003. Modulation of the N400 potential during auditory phonological/semantic interaction. Cogn. Brain Res. 17, 36–47. https://doi.org/10.1016/S0926-6410(03)00078-8

Rämä, P., Sirri, L., Serres, J., 2013. Development of lexical–semantic language system: N400 priming effect for spoken words in 18- and 24-month old children. Brain Lang. 125, 1–10. https://doi.org/10.1016/j.bandl.2013.01.009

Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44, 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Steinhauer, K., Alter, K., Friederici, A.D., 1999. Brain potentials indicate immediate use of prosodic cues in natural speech processing. Nat. Neurosci. 2, 191–196. https://doi.org/10.1038/5757

Steinhauer, K., Friederici, A.D., 2001. Prosodic boundaries, comma rules, and brain responses: the closure positive shift in ERPs as a universal marker for prosodic phrasing in listeners and readers. J. Psycholinguist. Res. 30, 267–295.

Tager-Flusberg, H., Kasari, C., 2013. Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum. Autism Res. 6, 468–478. https://doi.org/10.1002/aur.1329

Thie, J., 2013. A wireless marker system to enable evoked potential recordings using a wireless EEG system (EPOC) and a portable computer (No. e32v1). PeerJ PrePrints.

Torkildsen, J. von K., Syversen, G., Simonsen, H.G., Moen, I., Lindgren, M., 2007. Electrophysiological correlates of auditory semantic priming in 24-month-olds. J. Neurolinguistics 20, 332–351. https://doi.org/10.1016/j.jneuroling.2007.02.003

Vos, M.D., Kroesen, M., Emkes, R., Debener, S., 2014. P300 speller BCI with a mobile EEG system: comparison to a traditional amplifier. J. Neural Eng. 11, 36008. https://doi.org/10.1088/1741-2560/11/3/036008

Yau, S.H., McArthur, G., Badcock, N.A., Brock, J., 2015. Case study: auditory brain responses in a minimally verbal child with autism and cerebral palsy. Front. Neurosci. 9. https://doi.org/10.3389/fnins.2015.00208