OPEN

# Implicit Measures of Receptive Vocabulary Knowledge in Individuals With Level 3 Autism

*Emily L. Coderre, PhD,*† Mariya Chernenok, BA,†‡ Jessica O'Grady, MEd, BCBA,† Laura Bosley, MA,† Barry Gordon, MD, PhD,†§ and Kerry Ledoux, PhD†*

**Abstract:** Implicit measures of cognition are essential for assessing knowledge in people with Level 3 autism because such individuals are often unable to make reliable overt behavioral responses. In this study, we investigated whether three implicit measures—eye movement (EM) monitoring, pupillary dilation (PD), and event-related potentials (ERPs)—can be used to reliably estimate vocabulary knowledge in individuals with Level 3 autism. Five adults with Level 3 autism were tested in a repeated-measures design with two tasks. High-frequency 'known' words (eg, *bus*, *airplane*) and low-frequency 'unknown' words (eg, *ackee*, *cherimoya*) were presented in a visual world task (during which EM and PD data were collected) and a picture-word congruity task (during which ERP data were collected). Using a case-study approach with single-subject analyses, we found that these implicit measures have the potential to provide estimates of receptive vocabulary knowledge in individuals with Level 3 autism. Participants differed with respect to which measures were the most sensitive and which variables best predicted vocabulary knowledge. These implicit measures may be useful to assess language abilities in individuals with Level 3 autism, but their use should be tailored to each individual.

Received for publication January 4, 2018; accepted February 21, 2019.

From the *Department of Communication Sciences and Disorders, University of Vermont, Burlington, Vermont; †Department of Neurology, Division of Cognitive Neurology/Neuropsychology, The Johns Hopkins University School of Medicine, Baltimore, Maryland; ‡Center for Mind and Brain, University of California at Davis, Davis, California; and §Department of Cognitive Science, The Johns Hopkins University, Baltimore, Maryland.

Supported in part by a grant from the Nancy Lurie Marks Foundation; the Department of Defense Autism Research Program (AR093137); the Therapeutic Cognitive Neuroscience Fund; and the Benjamin and Adith Miller Family Endowment on Aging, Alzheimer's, and Autism Research.

Associate Editor Victor W. Henderson oversaw the review process for this article.

The authors declare no conflicts of interest.

Correspondence: Emily L. Coderre, PhD, Department of Communication Sciences and Disorders, University of Vermont, 489 Main Street, Burlington, Vermont 05405 (email: emily.coderre@med.uvm.edu).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.cogbehavneurol.com.



**Key Words:** level 3 autism, vocabulary, eye-tracking, pupillometry, ERP

(*Cogn Behav Neurol* 2019;32:95–119)

**ADI–R** = Autistic Diagnostic Interview—Revised. **ADOS** = Autism Diagnostic Observation Schedule. **ASD** = autism spectrum disorder. **EM** = eye movement. **ERP** = event-related potential. **ICA** = independent component analysis. **KBIT–2** = Kaufman Brief Intelligence Test, Second Edition. **PD** = pupillary dilation. **PPVT–4** = Peabody Picture Vocabulary Test, Fourth Edition. **ROI** = region of interest. **TD** = typically developing.

Autism spectrum disorder (ASD) is a neurodevelopmental disorder affecting one in 59 children (Centers for Disease Control and Prevention, 2018). Although ASD is defined by deficits in social communication and interaction, as well as restricted and repetitive behaviors or interests (American Psychiatric Association, 2013), the disorder is heterogeneous in its presentation and outcome. Individuals who are less severely affected by ASD may thrive in typical educational and vocational settings and live independently; those more severely affected by ASD may have social communication deficits and/or restrictive and repetitive behaviors that may require substantial environmental support. The individuals most severely affected by autism are discussed here as having "Level 3 autism," in accordance with the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, category Level 3, Severe Level of Autism (American Psychiatric Association, 2013). Among those individuals with Level 3 autism, there is incredible heterogeneity in features such as language (eg, some individuals develop fluent speech, whereas others may remain nonverbal throughout their lives) and other cognitive functions.

An understanding of cognitive functioning in autism is essential for developing therapeutic interventions and informing current models of the disorder at a basic level. Extensive research on cognition in autism has contributed to the direct development of interventional strategies to help remediate observed deficits (Dawson et al, 2010; Ozonoff and Miller, 1995). However, the inclusion of individuals who are most severely affected by ASD in research studies has remained extremely limited. Individuals with Level 3 autism (with or without accompanying speech and language deficits) pose significant challenges to cognitive testing for many reasons, including low reliability of behavioral responses,

inability or unwillingness to perform tasks, and/or sensitivity or aversion to testing equipment. Additionally, the very heterogeneity in function that is typical of individuals with Level 3 autism makes it difficult to create homogenous groups across which group averages would be statistically powerful or theoretically motivated. Because of such difficulties with testing and analyses, individuals with Level 3 autism are extremely underrepresented in studies of cognition, making our knowledge of autism consequently incomplete.

Because obtaining overt reports of these individuals' cognitive abilities may be difficult, implicit measures of these abilities, which can be collected and interpreted in the absence of behavioral responses, may provide important alternative assessments. The demonstration that implicit measures can be used to assess cognitive abilities and representations in individuals with Level 3 autism may allow this important segment of the autism spectrum to be included in a larger number of research studies, increase our understanding of the disorder, and enhance our ability to represent the wide heterogeneity in function that is at the very core of ASD.

The current work is an exploratory, proof-of-concept study to assess one aspect of cognitive function—receptive vocabulary knowledge—in those individuals who are most severely affected by ASD. We describe our method investigating the use of eye movement (EM) monitoring, pupillary dilation (PD), and event-related potentials (ERPs) to assess receptive vocabulary knowledge in five individuals with Level 3 autism. Given challenges to testing and the inevitable heterogeneity of this condition, we adopted a case-study methodology with single-subject analyses to demonstrate the utility of these implicit measures in individualized assessments and interventions. Importantly, we included in our study individuals with Level 3 autism with a range of verbal behaviors (from those who are minimally verbal to those who express more fluent speech) in an attempt to capture some of the heterogeneity of this population and to investigate the utility of these implicit measures for those individuals with and without functional speech.

## IMPLICIT MEASURES OF RECEPTIVE VOCABULARY

Our choice of implicit measures developed out of the widespread use of EMs, PD, and ERPs to assess cognitive processes in the absence of an overt behavioral response in typically developing (TD) adults and children as well as patient populations. EMs, PD, and ERPs have been established as valid implicit measures of receptive vocabulary in TD adults. The so-called "visual world paradigm," in which a visual display of pictures is followed by a spoken word or phrase, has become a canonical technique to assess online spoken language comprehension (Tanenhaus et al, 1995, 2000). Participants' eyes typically move toward a named picture as soon as it can be identified and disambiguated from other pictures. Similarly, when a written word is presented before the picture display (eg, *marriage*), EMs are faster to a semantically related picture (eg, ring) than to an unrelated picture (eg, pencil) (Odekar et al, 2009). Importantly, these EM patterns occur in the absence of a behavioral task (Odekar et al, 2009), indicating their utility as implicit measures of language comprehension.

PD (in keeping with the terminology in the pupillometry literature, dilation is referred to here as an increase in pupil diameter) in response to a stimulus typically increases with cognitive load (Beatty and Lucero-Wagoner, 2000; Granholm et al, 1996). PD is thus taken to reflect resource recruitment and has been used to assess processing demands in numerous cognitive tasks (Beatty and Lucero-Wagoner, 2000). In language comprehension studies, unrelated pairs of pictures and spoken words (eg, duck-"bed") have been shown to elicit greater PD than matched pairs (eg, duck-"duck"), indicating greater resource recruitment in unrelated conditions (Kuipers and Thierry, 2011, 2013). Such effects occur in the absence of a behavioral task (Kuipers and Thierry, 2011), demonstrating the utility of PD as an implicit measure of language comprehension.

ERPs are derived by time-locking changes in the EEG to a stimulus onset. Specific ERP components are associated with various aspects of language (Kutas et al, 2006; Sereno and Rayner, 2003). For current purposes, the N400 component is taken to reflect semantic processing and integration (Kutas and Hillyard, 1980; Lau et al, 2008) (see Kutas and Federmeier, 2011, for a broader discussion). N400 amplitude is reduced when a stimulus is easily integrated with its preceding context (eg, semantically related or congruent stimuli). This amplitude reduction, compared to conditions with more difficult semantic integration (eg, semantically unrelated or incongruent stimuli), is termed the *N400 effect* and is thought to index semantic integration.

The N400 effect occurs in the absence of behavioral responses (Kuipers and Thierry, 2011), demonstrating its utility as an implicit measure of language comprehension. Importantly, however, the N400 effect is only elicited when the target concept is within an individual's vocabulary range (Byrne et al, 1995; Connolly and D'Arcy, 2000; Connolly et al, 1995). No N400 effect is observed for words that are unknown to the participant because prior knowledge cannot ease integration in these cases.

We have previously demonstrated the complementary use of EMs, PD, and ERPs as implicit measures of receptive vocabulary knowledge in TD adults using high-frequency 'known' words (eg, *airplane*) and low-frequency 'unknown' words (eg, *cherimoya*) (Ledoux et al, 2016). In a visual world paradigm, during which EM and PD data were collected, four pictures were followed by a spoken word matching one of the pictures. EM data indicated that TD adults could more quickly identify the target picture for 'known' rather than 'unknown' words; that is, 'known' words had fewer fixations over the course of the trial; faster EMs to, and longer fixations on, the target; and more trials for which the target was the last picture to be fixated on. Pupillometry results showed greater PD for 'unknown' than 'known' words, suggesting greater resource recruitment. In a separate session, ERP data were collected during a picture-word congruency paradigm in which a picture was followed by a spoken word that matched (congruent) or did not match (incongruent) the picture. An N400 effect (reduced N400

amplitude for congruent vs incongruent pairs) occurred for 'known' but not 'unknown' words since the participants could not evaluate the congruency of 'unknown' concepts.

Overall, Ledoux et al (2016) demonstrated that all three measures showed effects that are consistent with prior work using EMs, PD, and ERPs as implicit measures of language, indicating that they can be used together to assess receptive vocabulary in TD adults. Critically, although the participants made behavioral responses throughout the tasks, all three implicit measures demonstrated the participants' ability to distinguish between 'known' and 'unknown' words without relying on behavioral responses. The implicit nature of these measures makes them potentially valuable for studying cognition in populations who are unable to provide overt responses.

Note that in both Ledoux et al (2016) and this report, we use 'known' and 'unknown' to refer to presumed participant knowledge of these sets of words. In the Methods section, we describe the assessments of true word status that were conducted with the caregivers of participants in this study to corroborate this presumption. However, we acknowledge that a true, objective assessment of participant knowledge of these words is an open question that is difficult to assess on the basis of behavioral response alone, which is the very motivation for the current study. Throughout this manuscript, we use single quotation marks to temper our categorization of the stimuli and to acknowledge that, although motivated by existing evidence, we cannot be completely certain about the word knowledge of our participants.

## IMPLICIT MEASURES OF RECEPTIVE VOCABULARY IN INDIVIDUALS WITH AUTISM

Implicit measures of receptive vocabulary have gained attention in recent years based on their potential to provide an assessment of language that behavioral responses or parental reports cannot capture (eg, Tager-Flusberg and Kasari, 2013). However, as we describe in the following paragraphs, assessment of the utility of implicit measures in individuals with autism has been tempered by findings of individual variability and typical use of only one assessment method at a time.

Several studies have attempted to use eye-tracking measures to assess receptive vocabulary in individuals with autism (Bavin et al, 2014; Brady et al, 2014; Plesa Skwerer et al, 2016; Venker et al, 2013). For instance, Brady et al (2014) found that children with ASD showed eye-gaze patterns that were similar to those of TD populations in a visual world paradigm: that is, longer looking times to target pictures for 'known' words, as determined by behavioral performance on the Peabody Picture Vocabulary Test, Fourth Edition (PPVT–4; Dunn and Dunn, 2007), than 'unknown' words.

Although these studies generally promote the utility of eye-tracking measures in this population, there is some indication that results are more variable, and may be less reliable, for individuals with Level 3 autism. For instance, Plesa Skwerer et al (2016) compared a variety of assessments of receptive vocabulary (the PPVT–4, caregiver report measures, an eye-tracking paradigm assessing word comprehension, and a computerized assessment of word comprehension using a touch screen) in minimally verbal children and adolescents with autism (ie, individuals with little functional speech). Using an individual approach, the authors found significant heterogeneity in receptive language abilities among participants and assessment methods, with no clear advantage of one method over another. This finding suggests that although eye-tracking may be a useful implicit measure for assessing receptive vocabulary in minimally verbal children with autism, it may not be a viable solution for all participants, and individual differences may still emerge.

Bavin et al (2014) used a visual world paradigm assessing single-word comprehension (eg, participants heard "Where's the x?" as their EMs were recorded) in children with a range of ASD symptom severity, including Level 3 autism. They found that greater symptom severity was associated with lower proportions of looking time at target pictures. This finding may suggest that eye-tracking paradigms are not as useful for assessing receptive vocabulary knowledge in individuals with Level 3 autism. On the other hand, Venker et al (2013) used a similar paradigm in young children with ASD with a range of symptom severities and found that, although there were notable individual differences in eye-gaze accuracy to the named picture, accuracy was not associated with symptom severity in this sample.

Although eye-tracking has been explored as an implicit measure of receptive vocabulary in individuals with Level 3 autism, to our knowledge, there are no studies investigating PD as an index of language comprehension in individuals with ASD. This is surprising because PD measures can easily be generated from eye-tracking software. The current study extends prior work to investigate whether PD can provide estimates of receptive vocabulary knowledge in individuals with Level 3 autism.

The use of ERPs to assess receptive vocabulary also has been proposed (Tager-Flusberg and Kasari, 2013), but to our knowledge, there is only one study that used this methodology in individuals with Level 3 autism. Cantiani et al (2016) tested nonverbal and minimally verbal children with ASD on an ERP picture-word matching paradigm using 'known' words. As a group, the children with ASD showed no significant N400 effect. However, when looking at individual patterns, half of the sample did show an N400 effect. This research finding suggests that although ERPs may be a useful implicit measure in assessing receptive vocabulary, they may not be a reliable indicator of knowledge for all participants. Furthermore, this work highlights the importance of considering individual data as well as group averages when testing individuals with autism.

Previous research has highlighted the benefit of using multiple measures when testing individuals with autism (Plesa Skwerer et al, 2016) because of testing difficulties and greater individual variability. Furthermore, by comparing these three different methods, we may identify one or more

 

that may be better suited for assessing receptive vocabulary in this population or among certain individuals. Ledoux et al (2016) demonstrated that EMs, PD, and ERPs all could estimate receptive vocabulary by discriminating between 'known' and 'unknown' words, meaning that if one methodology is unavailable (eg, a participant will not tolerate the EEG net, or the presence of glasses makes eye-tracking difficult), the other(s) may provide an alternative. Similarly, we used multiple EM and PD variables because some may be better indices of receptive vocabulary than others in certain individuals. The current study is the first to use these three implicit measures (EMs, PD, and ERPs) complementarily to assess receptive vocabulary knowledge in adolescents and adults with Level 3 autism.

Prior research using implicit measures in individuals with ASD has documented notable differences between TD groups and groups with ASD. Participants with ASD, for example, showed abnormal EM patterns during visual tasks (Brenner et al, 2007; Goldberg et al, 2002; Mottron et al, 2007; Schmitt et al, 2014) and atypical viewing patterns in visual world paradigms, such as lower proportions of looking time at the target picture (Bavin et al, 2014; Brock et al, 2008). Pupillometry studies have documented abnormalities, such as larger baseline pupil size (Anderson and Colombo, 2009) and smaller change in pupil size, in response to social stimuli (Martineau et al, 2011) in individuals with ASD. In addition, ERP studies have reported reduced or absent N400 effects in response to linguistic stimuli in individuals with ASD compared with TD individuals (Dunn et al, 1999; McCleery et al, 2010; Pijnacker et al, 2010).

Given the documented atypical patterns in implicit measures in ASD, directly comparing individuals with Level 3 autism to TD groups or even to groups of individuals with mild to moderate autism could be problematic. In our study, we used assessment paradigms that have been widely validated with TD adults (including in Ledoux et al, 2016, using the same stimuli and methods). Critically, however, one advantage of the current work is that all of the measures are within-subject comparisons of 'known' and 'unknown' words, and each participant thus acts as his or her own "control." While atypical patterns of implicit measures may occur in populations with ASD, these measures might still distinguish between 'known' and 'unknown' vocabulary *within* each individual. For instance, even if individuals with ASD show reduced N400 effects compared with TD adults, an N400 effect may still be observed within *one* individual with ASD for 'known' but not 'unknown' words. The potential for these implicit measures to distinguish 'known' and 'unknown' vocabulary on a within-subject basis would inform their utility in assessing receptive vocabulary in individuals with Level 3 autism.

To summarize, although the use of implicit measures of receptive vocabulary with individuals with Level 3 autism has been promoted in the literature (Tager-Flusberg and Kasari, 2013), only a handful of studies have actually examined whether these measures can provide estimates of receptive vocabulary in this population. Furthermore, there are no studies employing multiple implicit measures across different domains (eg, from both EMs and ERPs), which might provide additional utility in individuals with Level 3 autism. Finally, many previous studies have tested only 'known' words and performed group comparisons between individuals with ASD and TD individuals. By including both 'known' and 'unknown' words, our study tested whether these implicit measures could provide estimates of receptive vocabulary knowledge for individual participants. In sum, based on previous demonstrations that EMs, PD, and ERPs can differentiate 'known' from 'unknown' words in TD adults (Ledoux et al, 2016), this exploratory, proof-of-concept study is the first to assess whether these measures can also serve as within-subject implicit indices of receptive vocabulary knowledge in individuals with Level 3 autism.

We intentionally included individuals with a range of verbal abilities in our sample to determine whether these measures can provide estimates of receptive vocabulary independent from verbal ability. For instance, given the documented atypical patterns resulting from these measures in individuals with ASD, it may be that they cannot discriminate between 'known' and 'unknown' vocabulary in individuals with Level 3 autism even in those individuals who have intact language abilities. On the other hand, given the fact that we took a within-subject approach to distinguishing between 'known' and 'unknown' vocabulary, these measures may be able to estimate vocabulary knowledge in individuals with or without functional speech.

To the extent that these implicit measures would show similar patterns in TD adults and individuals with Level 3 autism, we would predict similar results in the two groups: faster and more accurate EMs to 'known' words, greater PD to 'unknown' words, and an N400 effect for 'known' but not 'unknown' words. It is important to keep in mind that because this proof-of-concept study is meant to establish whether these measures can provide estimates of receptive vocabulary in these participants, any outcome will be beneficial and will further our understanding of cognition in this population. Even a finding that none of these measures show differences between 'known' and 'unknown' words would be informative, suggesting that we might be better served to seek alternative methods of assessment. Similarly, a finding of mixed results, such that some measures indicate significant results for some participants but not others, would suggest that the use of these implicit measures should be performed on an individualized basis. Given our limited understanding of cognition in individuals with Level 3 autism and the shortage of studies investigating implicit assessments of language in this population, the current study is a worthwhile addition to the literature regardless of its outcome.

## SPECIFIC CONSIDERATIONS FOR INDIVIDUALS WITH LEVEL 3 AUTISM

Cognitive testing in individuals with Level 3 autism can be challenging owing to idiosyncrasies of the autism disorder such as sensory abnormalities or difficulties understanding or following directions (Kasari et al, 2013; Kylliäinen et al, 2014; Tager-Flusberg and Kasari, 2013;

 

Tager-Flusberg et al, 2017). For example, participants may be unable to use a response box or mouse, may make responses haphazardly, or may display no motivation to complete the task (eg, Kylliäinen et al, 2014). Such difficulties can lead to high rates of data loss, participant attrition, and/or increased variability in the data. Furthermore, some research suggests that the EEG activity of ASD participants is inherently noisier than that of TD individuals (eg, Pérez Velázquez and Galán, 2013; although see Davis and Plaisted-Grant, 2015), which may require collecting more EEG data to improve signal-to-noise ratios or making modifications to data acquisition or cleaning. These challenges with data acquisition and quality have likely contributed to the shortage of research that has been performed with individuals with Level 3 autism. Autism is also an extremely heterogeneous disorder with significant variation among individuals in terms of cognitive abilities, expressive and receptive language, and symptom severity, making it difficult to categorize individuals. Although group analyses may be informative, single-subject examination is crucial, especially when testing more severely affected individuals.

Given these considerations for this population, and that the ultimate aim of this work was to determine vocabulary knowledge on an individual basis, we adopted a case-study approach with single-subject statistical methods to assess the utility of EMs, PD, and ERPs in distinguishing 'known' and 'unknown' words in five individuals with Level 3 autism. Single-subject analyses may elucidate which measures best predict vocabulary abilities in each participant, the strength of the effects, and the efficacy of each measure in estimating receptive vocabulary. Because implicit measures offer a promising method of accessing the latent constructs of language in certain clinical populations, this work is an important step in understanding cognition in individuals with Level 3 autism, whose behavioral responses are often unreliable or unattainable.

# METHODS

## Participants

Participants were five males with Level 3 autism (M age = 32 years, SD = 14.6, range = 15–48; four Caucasian, one Asian) who had been recruited from the Baltimore, Maryland, community. To protect patient privacy, the initials used hereafter to identify the participants are not their real initials. All of the participants had normal or corrected-to-normal vision and hearing as assessed by caregivers or self-report. Because none of the participants wore glasses, this was not a consideration for the eye-tracking procedure. Experimental procedures were approved by the Johns Hopkins School of Medicine Institutional Review Board. For those participants who were unable to provide their own informed consent (D.L. and H.D.), we followed the Maryland law applicable to surrogate decision-making for health care, stating that a legal guardian may provide consent on behalf of the participant. For those participants who were able to legally provide their own consent (W.F., S.E., and P.B.), we obtained written informed consent from the participants as well as from their group home benefits manager.

For our study, we defined "Level 3 autism" based on three considerations:

- The severity of core features of autism according to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, Level 3 (Severe Level of Autism; American Psychiatric Association, 2013) diagnostic criteria, which marks severe deficits in social communication and restricted and repetitive behaviors requiring substantial support throughout the individual's daily life
- The level of environmental support and supervision needed
- Scores on the Autism Diagnostic Observation Schedule (ADOS; Lord et al, 2000, 2012) and/or the Autistic Diagnostic Interview—Revised (ADI–R; Lord et al, 1994)

All of the participants exhibited restricted and repetitive behaviors and severe deficits in verbal and/or nonverbal social communication skills that significantly affected their level of daily functioning. Each participant required direct 24-hour support staff and/or parental supervision with a focus on activities of daily living and functional communication. All were enrolled in adult or educational programs that were targeted to individuals with autism.

D.L. and H.D. were functionally nonverbal (ie, nonspeaking and only exhibiting limited communication using topic boards or visual communication systems). W.F., S.E., and P.B. had functional speech, although it was marked by the presence of stereotyped/idiosyncratic words and phrases. D.L. and H.D. had received diagnoses of autism before the age of 4. Records on early language development and initial diagnosis for W.F., S.E., and P.B. were not available at their residential facilities, and additional attempts to track previous records were unsuccessful.

## Neuropsychological Assessments

Table 1 shows the test scores for each participant.

All of the participants had a current diagnosis of autism at the time of testing, which was verified using either the ADOS–1 (first edition) or ADOS–2 (second edition), depending on the version that was current at the time of testing, and/or the ADI–R. These assessments were administered by research team members who were clinically reliable (ie, had completed the official ADOS clinical training). No appropriate module of the ADOS was available for two participants (D.L. and W.F.) as the module that met the criteria for expressive language skills was developmentally inappropriate for the participants' chronological age. The researchers performed "adapted" modules by interacting with these two participants and identifying the specific behaviors measured by the ADOS. These adapted scores are noted in Table 1 but cannot be considered "official" ADOS scores. An ADOS could not be completed with H.D. at the time of initial assessment, and he was thereafter unavailable for additional testing. An ADI–R was completed for D.L. and H.D. The ADI–R was not performed for W.F., S.E., or P.B. because their legal guardians could not provide information about their infancy and early development (a major component of the ADI–R), and their parents were unavailable for contact.

**TABLE 1.** Participant Demographics, Autism Diagnostic Test Results, Intelligence Scores, and Vocabulary Scores

| Participant (Not Their Real Initials) | Age | Sex | ADOS | | | | ADI–R | | | KBIT–2* | | PPVT–4* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Version | Module | Social+ Communication Total | Classification | Social Interaction | Communication | Behaviors | Development | Verbal | Nonverbal | |
| D.L. | 18 | Male | 1 | 1 (adapted) | 20 | Autism | 22 | 14 | 2 | 5 | Not available | | Not available |
| H.D. | 15 | Male | | | Not available | | 22 | 20 | 6 | 5 | Not available | | 20 |
| W.F. | 39 | Male | 2 | 4 (adapted) | 22 | Autism | | Not available | | | 45 | 79 | 58 |
| S.E. | 40 | Male | 2 | 4 | 20 | Autism | | Not available | | | 40 | 60 | 43 |
| P.B. | 48 | Male | 2 | 4 | 19 | Autism | | Not available | | | 93 | 131 | 94 |

*Scores for the KBIT–2 and PPVT–4 are standard scores.
**ADOS** = Autism Diagnostic Observation Schedule. **ADI–R** = Autistic Diagnostic Interview—Revised. **KBIT–2** = Kaufman Brief Intelligence Test, Second Edition. **PPVT–4** = Peabody Picture Vocabulary Test, Fourth Edition.

Overall, these assessments confirmed the diagnosis of autism for all of the participants.

The Kaufman Brief Intelligence Test, Second Edition (KBIT–2; Kaufman and Kaufman, 2004) and the PPVT–4 were administered to assess the participants' intelligence and receptive vocabulary, respectively. All of the researchers had experience working with individuals with ASD in a research setting. D.L. and H.D. were unable to complete the KBIT–2 or PPVT–4 because of a lack of compliance and/or an inability to understand the test directions, although H.D. was able to complete some questions on the PPVT–4. The KBIT–2 and PPVT–4 standard scores for W.F. and S.E. indicated verbal and/or nonverbal abilities in the range of intellectual dysfunction (< 70). However, because of difficulties maintaining attention and understanding task instructions, these scores may not be accurate reflections of their true abilities. P.B.'s KBIT–2 and PPVT–4 scores did not indicate intellectual impairment.

Although intelligence and language ability were included in obtaining an overall picture of each participant, they were noted as possible associated features of autism and were not included in identifying these individuals as having Level 3 autism. While all participants were classified as having Level 3 autism for the purposes of this study, they varied in their intelligence and language abilities (Table 1); for example, D.L. and H.D. had little-to-no functional speech, whereas W.F., S.E., and P.B. expressed more fluent speech.

## Stimuli

As shown in Table 2, the stimuli for the 'known' and 'unknown' words used in the experiment consisted of 160 auditory words. As illustrated in Figure 1, each word had a matching picture. Half were high-frequency words (average frequency per million in the Corpus of Contemporary American English [Davies, 2008] = 56.5, SD = 84.1) such as *bus*. These words were classified as 'known,' as we expected most to be known by the participants. Half were extremely low-frequency words (average frequency per million = 0.4, SD = 0.7, although given their low frequency, many do not occur in language corpora), such as *avocet*. These words were classified as 'unknown' as we expected most to be unfamiliar to the participants. The 'unknown' words had slightly more letters (M = 6.8, SD = 1.6) than the 'known' words (M = 5.1, SD = 1.5). The influence of this difference, however, is likely to be minimal because the majority of our measures are proportional or do not otherwise depend on latency.

More information is available in Ledoux et al (2016) about the frequency norming supporting these 'known' and 'unknown' categories in TD adults. Such norming is virtually impossible with individuals with Level 3 autism for the very reason that we sought to use implicit measures of assessment: Their verbal and other behavioral responses are extremely variable and often unreliable. In addition to these objective classifications, each participant's parent or caregiver subjectively rated whether the individual knew

**TABLE 2.** Stimuli for Each of the 'Known' and 'Unknown' Words Used in the Visual World and Picture-Word Congruity Tasks

| 'Known' | | 'Unknown' | |
|---|---|---|---|
| airplane | flower | ablution | fossa |
| ant | fork | acerola | frieze |
| apple | frog | ackee | gelada |
| baby | girl | addax | gerenuk |
| ball | grapes | agouti | greengage |
| balloon | hammer | anemometer | harrow |
| banana | horse | angklung | homogenizer |
| bathtub | house | anole | jerboa |
| bed | key | argali | jicama |
| bicycle | kite | avocet | jujube |
| book | knife | babirusa | kinkajou |
| boots | ladder | balalaika | kohlrabi |
| bottle | leaf | banteng | kumquat |
| bowl | lion | barasingha | loquat |
| box | monkey | bilby | mead |
| boy | mouse | binturong | medlar |
| bread | orange | bolster | melee |
| brush | pencil | caiman | mendicant |
| bus | pig | cainito | millet |
| butterfly | pot | capybara | okapi |
| cake | pretzel | caracal | pangolin |
| camera | rabbit | carambola | panoply |
| candy | scissors | carboy | peccary |
| car | shoes | celeriac | persimmon |
| cat | slide | chayote | pillory |
| chair | snake | cherimoya | pinion |
| cheese | snowman | civet | quince |
| circle | sock | colugo | raiment |
| clock | spider | conflagration | ramekin |
| cloud | spoon | confluence | repast |
| coat | square | cudgel | rowan |
| cookie | star | douc | saguaro |
| cow | swing | drachma | specie |
| crayons | table | dugong | sylph |
| cup | telephone | durian | talisman |
| dinosaur | tiger | echidna | tamarillo |
| dog | train | effigy | tamarind |
| door | tree | epee | tarsier |
| drum | umbrella | feijoa | visage |
| elephant | watch | floe | yangmei |

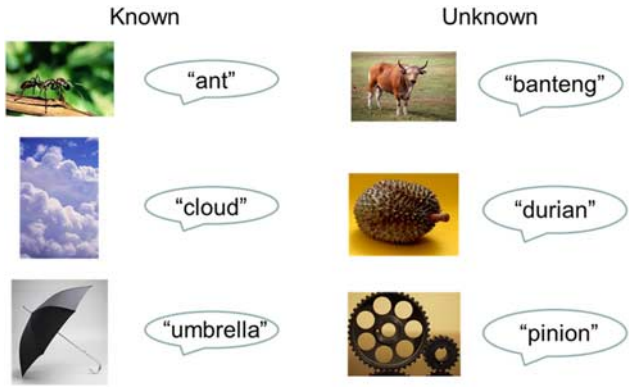


**FIGURE 1.** Examples of 'known' and 'unknown' stimuli.

each word receptively. These ratings estimated that all of the participants were familiar with most of the 'known' words and were unfamiliar with all of the 'unknown' words in the stimulus set.

For picture stimuli, high-resolution color photos were selected from online sources to represent each word (Figure 1). Pretesting with three TD adults confirmed that these images represented the corresponding concepts (dictionary definitions were provided for 'unknown' words). All of the words were highly imageable, as determined through pretesting. Picture luminance was matched across 'known' and 'unknown' words. For auditory stimuli, high-quality auditory recordings were made for each word using Audacity 1.3 and were edited using Computerized Speech Lab Model 4150 (KayPENTAX). The duration of the auditory stimuli ranged from 500 milliseconds to 1200 milliseconds.

## Task Procedure

The experiment consisted of a visual world task (EM and PD) and a picture-word congruity task (ERP), which were completed in separate sessions. Some participants underwent multiple sessions per task to ensure that there were adequate amounts of usable data.

### *Visual World Task*

The visual world paradigm was presented in E-Prime 2.0.8.74. In each trial, a central fixation cross was presented for 1000 milliseconds. Four pictures were then presented, one centered in each quadrant of the screen, followed 20 milliseconds later by an auditory word. 'Known' words were always presented with 'known' distractors, and 'unknown' words with 'unknown' distractors, to ensure that participants could not eliminate foils in the 'unknown' condition based on familiarity. All four pictures remained on the screen for a maximum of 5000 milliseconds after word presentation or until the participant selected a picture with a mouse click. These stimulus parameters are similar to those used in previous studies using the visual world paradigm or obtaining PD measures (Kuipers and Thierry, 2011; Odekar et al, 2009).

The experimental session consisted of 160 pseudo-randomized trials (one per item) in eight blocks of 20 trials each. The pictures were presented at 1.6 degrees to 9.5 degrees of visual angle on a MicroTouch 3M 15″ LCD monitor with $1024 \times 768$ resolution. EM and PD data were collected using an ASL Model 504 eye-tracking system. Pupil diameter was measured horizontally and was recorded every 17 milliseconds in pixels. The entire session lasted approximately 30 minutes, including approximately 15 minutes for equipment setup and calibration. To maintain attention, we asked the participants to use the computer mouse to indicate which picture matched the spoken word.

Note that although we were interested in the implicit measures and did not require behavioral responses, all of the participants (in keeping with their prior experience using computers) spontaneously sought a task or demonstrated a desire to have a task to complete during the visual world task. Therefore, the participants were allowed

to use the mouse as they wished, and presumably they attempted to click on the named picture. We are hesitant to interpret these behavioral responses because of nonstandardization in administration of the instructions, uncertainty in participants' interpretation of task demands, and so on. We do note, however, that for three participants (W.F., S.E., and P.B.), 'known' words showed significantly higher accuracy and faster reaction times on the visual world task than 'unknown' words. Behavioral results are provided in the Supplemental Digital Content 1, http://links.lww.com/CBN/A74.

### *Picture-Word Congruity Task*

The picture-word congruency paradigm was presented in E-Prime. A centrally presented picture was followed 700 milliseconds later by a spoken word. Each word was presented twice: once in an incongruent condition (word and picture did not match) and once in a congruent condition (word and picture matched), yielding a total of 320 trials. Incongruent picture-word pairs were drawn from the same knowledge condition ('known' or 'unknown') and did not share an initial phoneme. The picture was presented for 1000 milliseconds after the offset of the auditory stimulus. Pictures were presented at 2.4 degrees to 9.5 degrees of visual angle on a Dell 17″ LCD monitor with 1280 × 1024 resolution. ERPs were recorded at 250 Hz using a 256-channel Hydrocel Geodesic Sensor Net and NetStation 4.3. Impedances were kept under 50 kΩ. Videos were recorded from the front and back to code for any "bad" trials during data preprocessing (see the Data Preprocessing section). The entire session lasted approximately 35 minutes, including approximately 15 minutes for net application and setup. The (nonobligatory) behavioral task for this paradigm (as it had been developed for TD participants in Ledoux et al, 2016) required the participants to withhold their response until a delayed fixation cross appeared (to minimize movement artifacts) and then indicate whether the word and picture matched using a button press. D.L., H.D., W.F., and S.E. did not appear to understand these instructions and did not make behavioral responses. P.B. appeared to understand the instructions but was unable to reliably wait for the response fixation cross, and so the majority of his responses were not captured. Therefore, no interpretable behavioral data were collected from this task.

### *Number of Sessions*

Where necessary, the participants were desensitized to the equipment by using a Velcro strap and a contraption that resembled the eye-tracking equipment and/or a practice EEG net. Table 3 shows the number of trials that were collected and used in the final analyses. We required that approximately half of the total trials collected for each measure be usable. D.L. was unable to complete an entire eye-tracking session, with 120 trials collected (out of 160 total). Due to excessive movement in the first EEG session, D.L. performed two additional shorter sessions approximately 1 month later. H.D. performed one session each of the eye-tracking and EEG tasks. W.F. performed one eye-tracking session; due to movement and noise artifacts in the first EEG session, a second session was performed approximately 2 months later. S.E. required two eye-tracking and two EEG sessions due to excessive movement, talking, and difficulty with compliance in the first sessions; the second sessions were performed approximately 2 months later. P.B. performed one eye-tracking session; due to excessive movement during the first EEG session, a second session was run approximately 1 month later.

## Data Preprocessing

### *EM Data*

EM data from the visual world task were analyzed using ASL Results software (Applied Science Laboratories, 2009). Each visual display was divided into five regions of interest (ROIs), consisting of the four pictures and the central fixation. The "target" is referred to here as the named picture on each trial. A fixation was operationalized as a time period during which eye gaze remained at one location. A stable gaze duration for 100 milliseconds or more and a visual angle variation of less than or equal to 1 degree determined a *fixation onset*. Three or more sequential fixations deviating from the onset location by more than or equal to 1 degree of visual angle determined a *fixation offset*. *Dwell time* was operationalized as the time spent looking at the target with or without fixation. If less than half of the trial was detected by the eye tracker (ie, the sum of all fixation durations was <50% of the total trial length), that trial was removed.

In Ledoux et al (2016), all of the EM variables examined showed significant differences between the 'known'

**TABLE 3.** Total Number of Collected and Usable Trials for Each Participant

| | Eye Movement (EM) | | Pupillometry (PD) | | Event-related Potentials (ERPs) | |
|---|---|---|---|---|---|---|
| Participant | Total Number of Trials Recorded | Total Number of Usable Trials | Total Number of Trials Recorded | Total Number of Usable Trials | Total Number of Trials Recorded | Total Number of Usable Trials |
| D.L. | 120 | 62 | 120 | 56 | 780 | 190 |
| H.D. | 160 | 67 | 160 | 107 | 320 | 203 |
| W.F. | 160 | 134 | 160 | 155 | 640 | 462 |
| S.E. | 320 | 93 | 320 | 174 | 640 | 289 |
| P.B. | 160 | 114 | 160 | 122 | 640 | 375 |

For the EM and PD data, a full session was 160 trials; for the ERP data, a full session was 320 trials.

and 'unknown' words in the TD adults. In this study of individuals with Level 3 autism, all of the EM variables were included because some variables might be better indices of receptive knowledge than others. For each trial, the following variables were calculated. All duration and latency measures are in milliseconds.

- *Total number of fixations*: total number of fixations in the entire trial
- *Mean fixation duration*: average duration of fixations in the target ROI
- *First fixation duration*: duration of the first fixation in the target ROI
- *First dwell on stimulus*: total time spent in the target ROI, with or without fixation, during the first entry
- *Latency to first fixation*: time elapsed before the first fixation in the target ROI
- *Latency to first refixation*: time elapsed before the first refixation in the target ROI (ie, time to come back to the target ROI after leaving the target ROI)
- *Percentage of fixation duration on target*: total fixation duration on the target divided by total fixation duration for all of the pictures
- *Percentage of dwell time on target*: percentage of time spent in the target ROI with or without fixation (ie, total dwell time on the target/length of trial)
- *Percentage of trials first fixated*: percentage of trials on which the target was the first picture fixated
- *Percentage of trials last fixated*: percentage of trials on which the target was the last picture fixated

Because some participants had longer reaction times for 'unknown' than 'known' trials, and because trials ended upon response, 'known' trials were sometimes shorter than 'unknown' trials. Differences in trial length would likely not affect latency measures (eg, *latency to first fixation*); percentage measures, which divide by trial length, account for this difference automatically. *Number of fixations* is necessarily dependent on trial length and, as seen in the EM data, was often larger for 'unknown' than 'known' trials.

### *Pupillometry Data*

Pupillometry data from the visual world task were exported from ASL Results and were analyzed in R (R Core Team, 2013). Pupil diameter was converted to millimeters, and blinks were replaced by linear interpolation. For each trial, a "baseline" pupil diameter (obtained by averaging over the 200-millisecond prestimulus time window) was subtracted from each measurement following stimulus presentation. Based on the pupillometry variables used in Ledoux et al (2016), three measures were calculated: peak dilation, mean dilation, and maximum percent dilation. Trials in which 20 or more consecutive data points ($\geq$340 milliseconds) were missing due to lack of fixations were removed.

### *ERP Data*

ERP data were preprocessed using EEGlab 10.2.2 (Delorme and Makeig, 2004) and Matlab 8.1 (MathWorks Inc.). The data were bandpass filtered from 0.1 to 30 Hz and were transformed to the average reference. Continuous data were segmented from 800 milliseconds before the word to 1000 milliseconds after the word (with the picture presented at approximately 700 milliseconds). Videos recorded during the EEG session were reviewed to identify and remove any "bad" trials containing movement, speaking, or inattention to the stimulus (eg, not looking at the screen). Artifact correction was performed using independent component analysis (ICA) (Delorme et al, 2007; Jung et al, 2000). For participants with multiple sessions, the mean of each trial was removed before concatenating the sessions for ICA (Delorme and Makeig, 2004; Groppe et al, 2009). Before ICA decomposition, the data were reduced to 64 dimensions. ICA components were reviewed individually, and those contributing to sources of noise were removed from the data. Following ICA, a joint probability algorithm removed trials in which the amplitude at any channel or time point exceeded 3 SD above or below the average amplitude for that channel. Finally, the cleaned data were visually reviewed, and any further bad trials (eg, those containing artifacts not eliminated by the joint probability algorithm) were removed.

## Statistical Analyses

Single-subject statistical analyses were performed in R using permutation tests. In the behavioral, EM, and PD data, all of a participant's individual trials were permuted to create a distribution of simulated test statistics. For each variable, 5000 iterations were performed, which can estimate an alpha level of 0.01 to within 2% (Groppe et al, 2011; Manly, 1997). At each iteration, we permuted 'known' or 'unknown' labels between trials and ran a one-way (*trial type*: 'known'/'unknown') analysis of variance. This repeated-measures approach accounts for the intercorrelation of the data, which are not independent nor paired across trials. The $F$ statistics from each iteration were used to create a null distribution from which the critical $F$ value corresponding to an alpha of 0.05 was calculated. We compared the observed $F$ value to the critical $F$ value to determine statistical significance. For observed $F$ values exceeding the critical $F$ values, $P$ values were derived for the observed effect. Bonferroni corrections for multiple comparisons were performed for the number of variables in each measure (10 in EM, three in PD). All reported $P$ values are Bonferroni-corrected unless otherwise specified.

For the ERP data, nine topographic regions were defined—clustered around F3/Fz/F4, C3/Cz/C4, and P3/Pz/P4 (Figure 2). Data were collapsed over all electrodes within each cluster. Congruent versus incongruent comparisons were performed separately for 'known' and 'unknown' trials. Based on previous literature, we would expect an N400 effect from approximately 300 milliseconds to 500 milliseconds after word onset. However, because no previous studies have investigated N400 effects in individuals with Level 3 autism, it is unclear whether latency differences would occur in this population. Rather than restrict analyses to predefined time windows, permutation tests were performed at every time point. To reduce the number of comparisons (Groppe et al, 2011), the data were down-sampled to 125 Hz (one sample every 8 milliseconds), and analyses were restricted to a

 

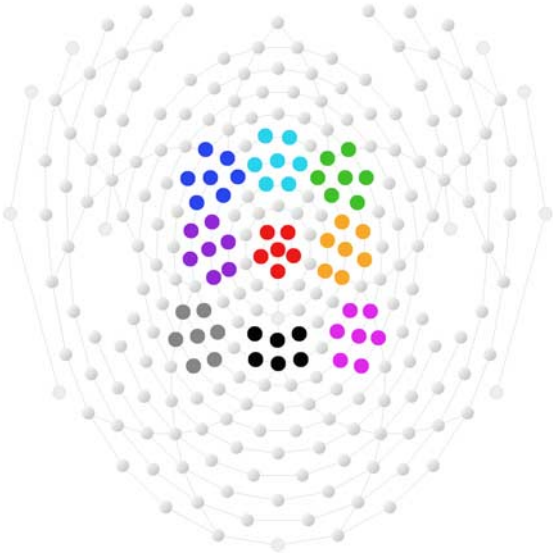

**FIGURE 2.** Illustration of the nine electrode clusters used for the EEG analysis.

time window from 200 milliseconds after word onset (as congruency differences should not occur earlier than this) until the trial's end. For each iteration, one-way (*congruency*: congruent/incongruent) analyses of variance were performed at each time point and electrode. Correction for multiple comparisons was performed using a cluster-based family-wise error correction at $P$ is less than 0.05 (full details in Groppe et al, 2011). Temporal clusters were defined as two or more consecutive time points showing effects at $P$ is less than 0.05. For each temporal cluster, $F$ values were summed to obtain the cluster "mass." The largest cluster-level $F$ mass from each iteration was used to create a null distribution from which we derived the critical cluster $F$ mass corresponding to an alpha of 0.05. We then compared each observed cluster-level $F$ mass to the critical cluster $F$ mass to determine statistical significance.

## RESULTS

### Individual Participant Analyses

#### *D.L.*

No significant differences between 'known' and 'unknown' words occurred in the EM variables (all $P$ values $>0.53$; Figure 3A) or PD variables (all $P$ values $>0.32$ uncorrected; Figure 3B).

In the ERP data (Figure 3C), no significant differences between congruent and incongruent conditions occurred for either word type.

#### *H.D.*

No significant differences between 'known' and 'unknown' words occurred in the EM variables (all $P$ values $>0.16$; Figure 4A) or PD variables (all $P$ values $>0.14$; Figure 4B).

In the ERP data (Figure 4C), no significant differences between congruent and incongruent conditions occurred for either word type.

#### *W.F.*

In the EM variables (Figure 5A), 'known' words showed larger *mean fixation duration* ($F_{1,132}=62.40$, $P<0.01$), *first fixation duration* ($F_{1,127}=8.52$, $P<0.05$), *first dwell* ($F_{1,127}=63.54$, $P<0.01$), *percent fixation duration* ($F_{1,132}=230.20$, $P<0.01$), and *percent last fixated* ($F_{1,132}=60.11$, $P<0.01$) than 'unknown' words. 'Unknown' words showed a larger *number of fixations* ($F_{1,132}=100.50$, $P<0.01$) than 'known' words. Note that some variables have different degrees of freedom because different numbers of trials were included in the analyses. On some trials, the participant did not look at the target picture at all. In such cases, the *mean fixation duration* would have a value of 0 and would be included in the analysis, but *first fixation duration* would be coded as "not applicable" (NA) and would not be included.

In the PD variables (Figure 5B), no significant differences between 'known' and 'unknown' words were observed (all $P$ values $>0.47$ uncorrected).

In the ERP data (Figure 5C), a significant N400 effect (incongruent more negative than congruent) occurred at the Pz cluster from approximately 200 milliseconds to 400 milliseconds. This effect occurred only for 'known' words; no congruency effects occurred for 'unknown' words.

#### *S.E.*

In the EM variables (Figure 6A), significant differences between 'known' and 'unknown' words occurred for *percent last fixated* ($F_{1,87}=10.62$, $P<0.05$), with 'known' words being the last picture fixated more often than 'unknown' words.

In the PD variables (Figure 6B), no significant differences between 'known' and 'unknown' words occurred (all $P$ values $>0.14$).

In the ERP data (Figure 6C), no significant differences between congruent and incongruent conditions occurred for either word type.

#### *P.B.*

In the EM variables (Figure 7A), 'known' words showed larger *percent fixation duration on stimulus* ($F_{1,112}=75.89$, $P<0.01$), *percent dwell* ($F_{1,112}=64.47$, $P<0.01$), and *percent last fixated* ($F_{1,112}=72.42$, $P<0.01$) than 'unknown' words. 'Unknown' words showed a larger *number of fixations* ($F_{1,112}=78.49$, $P<0.01$) than 'known' words.

In the PD variables (Figure 7B), 'unknown' words showed larger *peak dilation* ($F_{1,120}=4.50$, $P=0.10$) and significantly larger *max percent dilation* ($F_{1,120}=5.86$, $P<0.05$) than 'known' words.

In the ERP data (Figure 7C), a significant N400 effect (incongruent more negative than congruent) occurred at the C3 cluster from approximately 400 milliseconds to 550 milliseconds. This effect occurred only for 'known' words; no congruency effects occurred for 'unknown' words.

### Comparison of Individual Patterns

Table 4 summarizes each participant's results for each measure. To illustrate effect magnitudes for each variable and participant, within-subject 'unknown' – 'known' differences for each variable were scaled to normalized $z$ scores (Figure 8A). Normalization within subjects enables

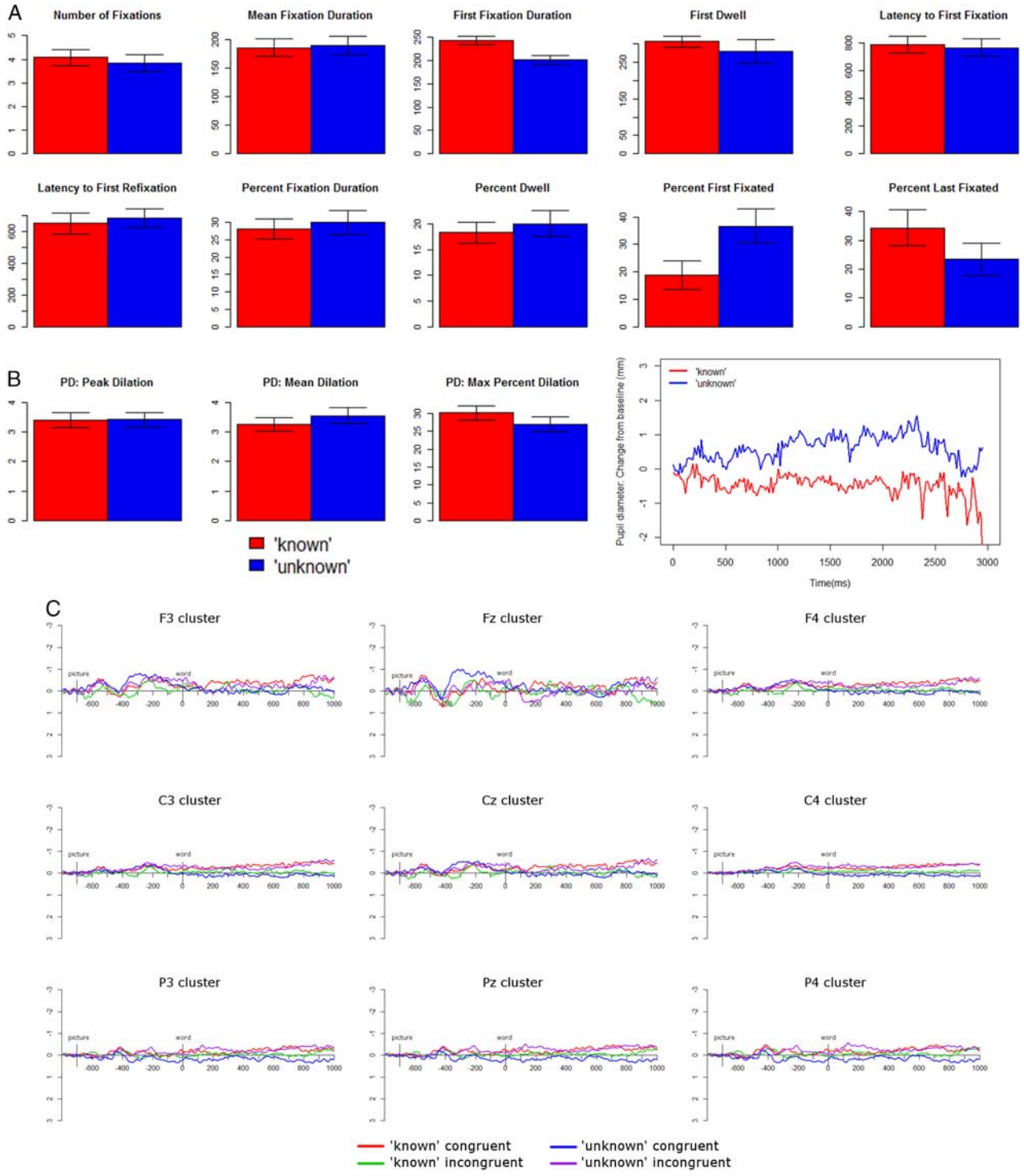


**FIGURE 3.** Results for D.L. **A:** Bar graphs comparing 'known' and 'unknown' word trials for each of the eye movement variables. **B:** Comparisons of 'known' and 'unknown' word trials for each of the pupillometry variables. Error bars indicate standard error of the mean. **C:** Event-related potential data for all conditions at the nine electrode cluster sites. Negative is plotted up.

comparison of effects on different scales and illustrates the strength of each effect in each participant. Topographic plots of incongruent – congruent differences illustrate ERP effects for 'known' and 'unknown' words (Figure 8B).

Figure 8 demonstrates the variability within and between participants with regard to which measure(s) best distinguished between 'known' and 'unknown' words. For example, for W.F., the EM measures showed much larger

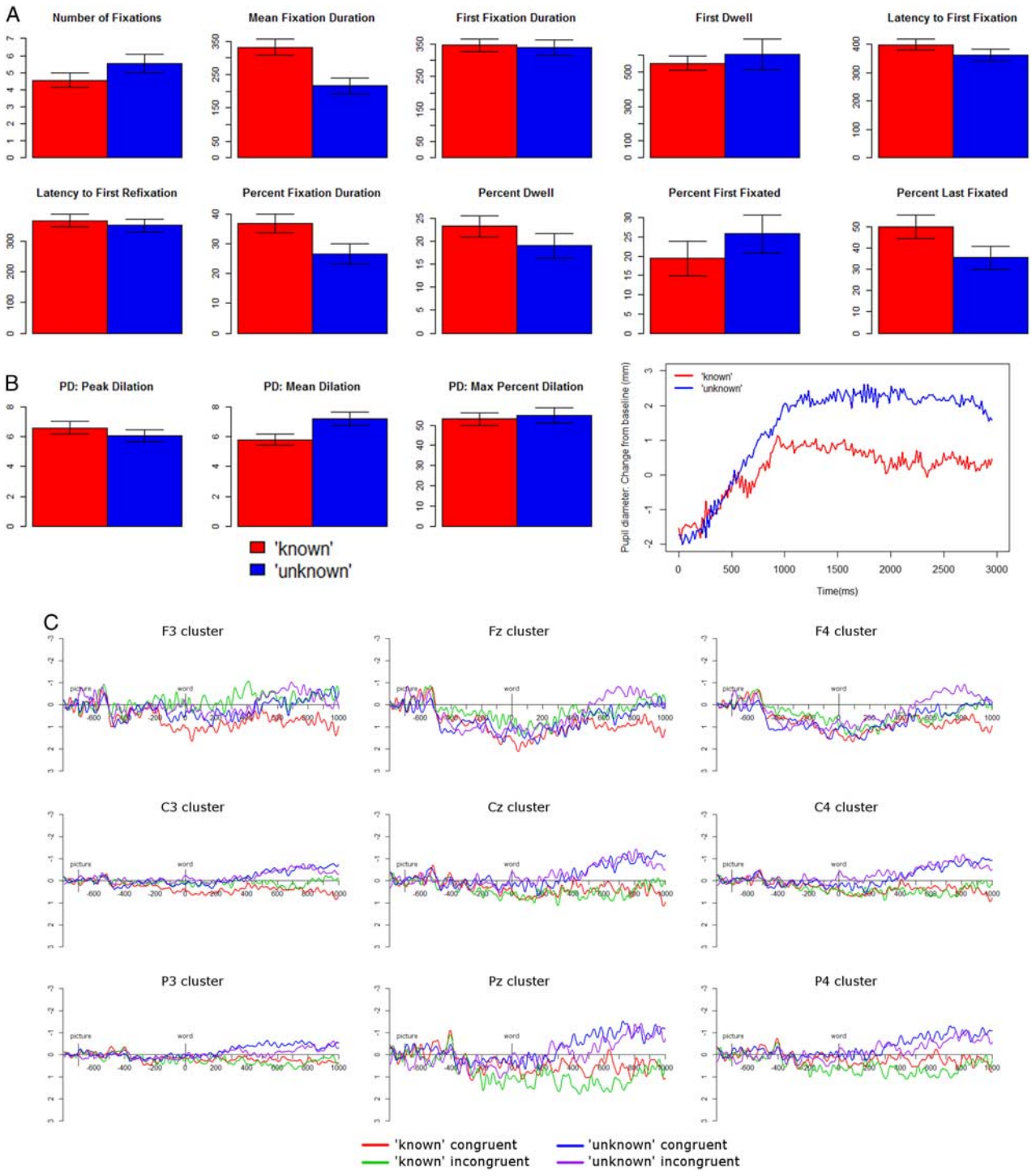


**FIGURE 4.** Results for H.D. **A:** Bar graphs comparing 'known' and 'unknown' word trials for each of the eye movement variables. **B:** Comparisons of 'known' and 'unknown' word trials for each of the pupillometry variables. Error bars indicate standard error of the mean. **C:** Event-related potential data for all conditions at the nine electrode cluster sites. Negative is plotted up.

effects than the PD measures. Likewise, *mean fixation duration* was the largest effect for H.D. but showed no effect for D.L. This variability also occurred in the EEG data: Although W.F. and P.B. showed large N400 effects, the other participants showed negligible effects. Overall, Figure 8 illustrates that the specific measures that best elicit differences between 'known' and 'unknown' vocabulary may differ among individuals.

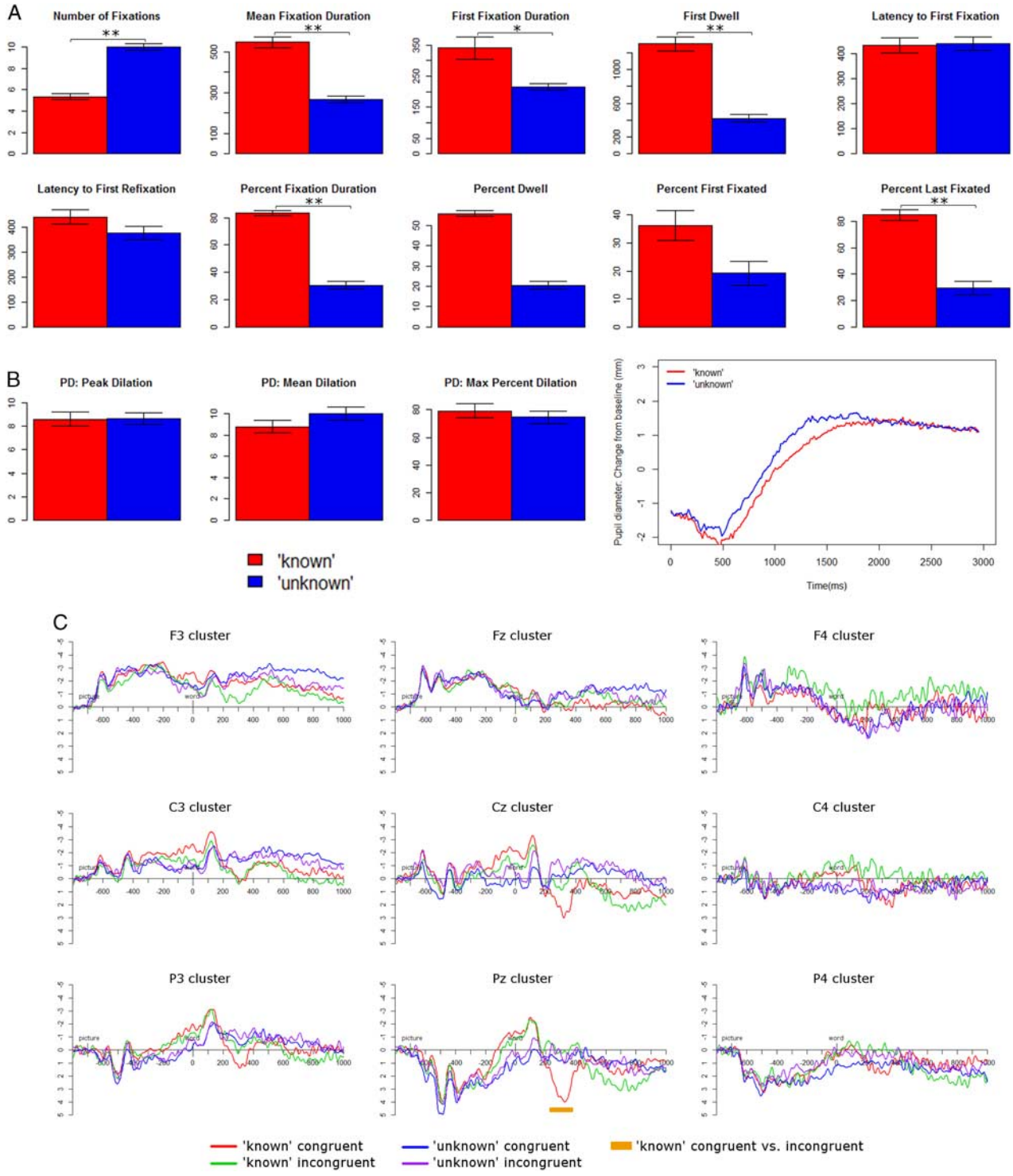


**FIGURE 5.** Results for W.F. **A:** Bar graphs comparing 'known' and 'unknown' word trials for each of the eye movement variables. **B:** Comparisons of 'known' and 'unknown' word trials for each of the pupillometry variables. Error bars indicate standard error of the mean. Significant differences between 'known' and 'unknown' words, based on permutation tests with Bonferroni corrections, are indicated by asterisks (*Significant at $P < 0.05$; **Significant at $P < 0.01$). **C:** Event-related potential data for all conditions at the nine electrode cluster sites. Negative is plotted up. The orange bar beneath the waveforms indicates significant differences between congruent and incongruent conditions for 'known' words, as determined by permutation tests with a cluster-based family-wise error correction at $P < 0.05$.

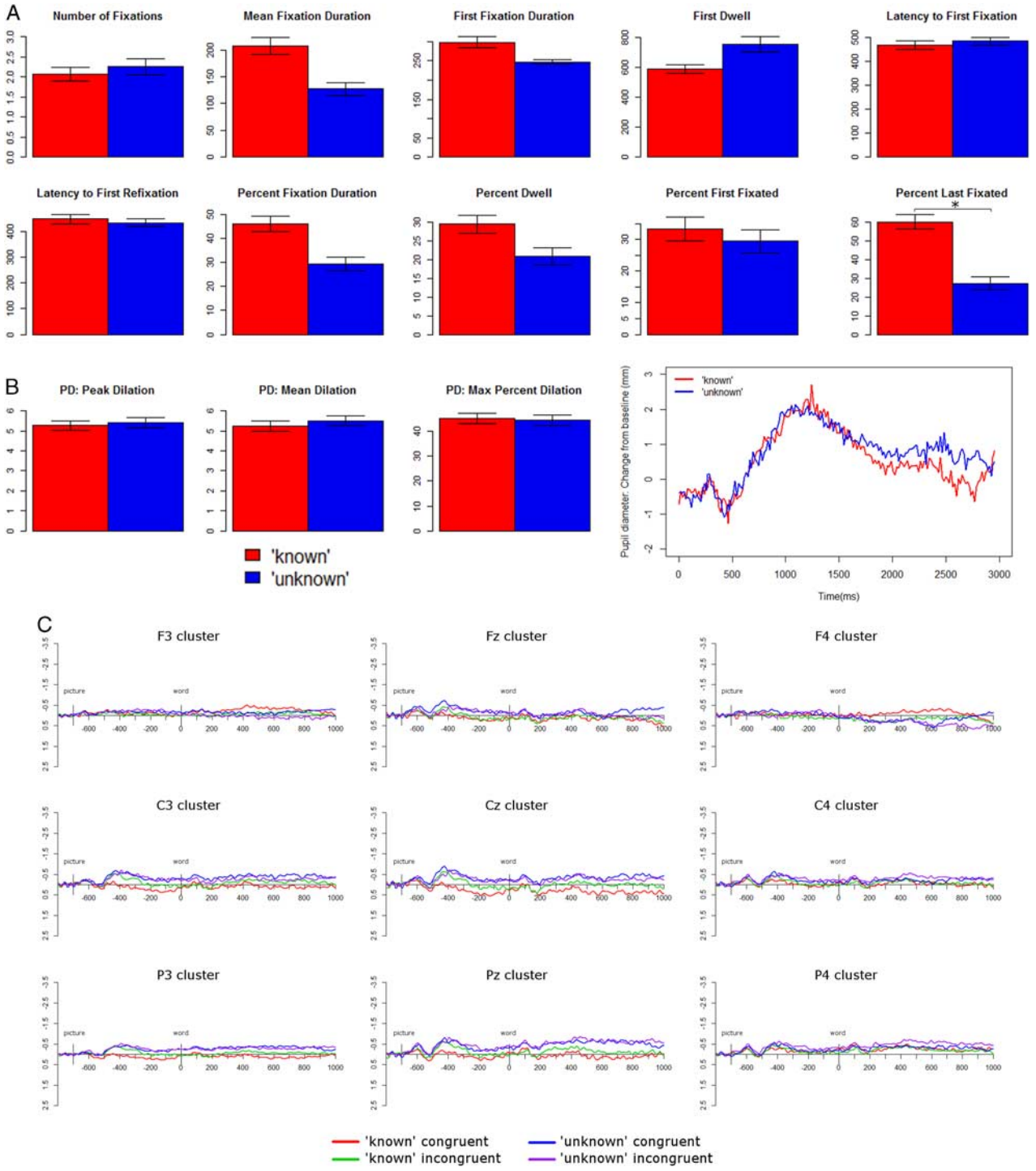


**FIGURE 6.** Results for S.E. **A:** Bar graphs comparing 'known' and 'unknown' word trials for each of the eye movement variables. **B:** Comparisons of 'known' and 'unknown' word trials for each of the pupillometry variables. Error bars indicate standard error of the mean. Significant differences between 'known' and 'unknown' words, based on permutation tests with Bonferroni corrections, are indicated by asterisks (*Significant at $P < 0.05$). **C:** Event-related potential data for all conditions at the nine electrode cluster sites. Negative is plotted up.

## Reliability Analyses

In an effort to evaluate the reliability of the measures, we calculated both within-subject reliability and test-retest reliability statistics for each measure for each participant. Correlations were run for the 'known' and 'unknown' trials separately and, for the ERP measures, for

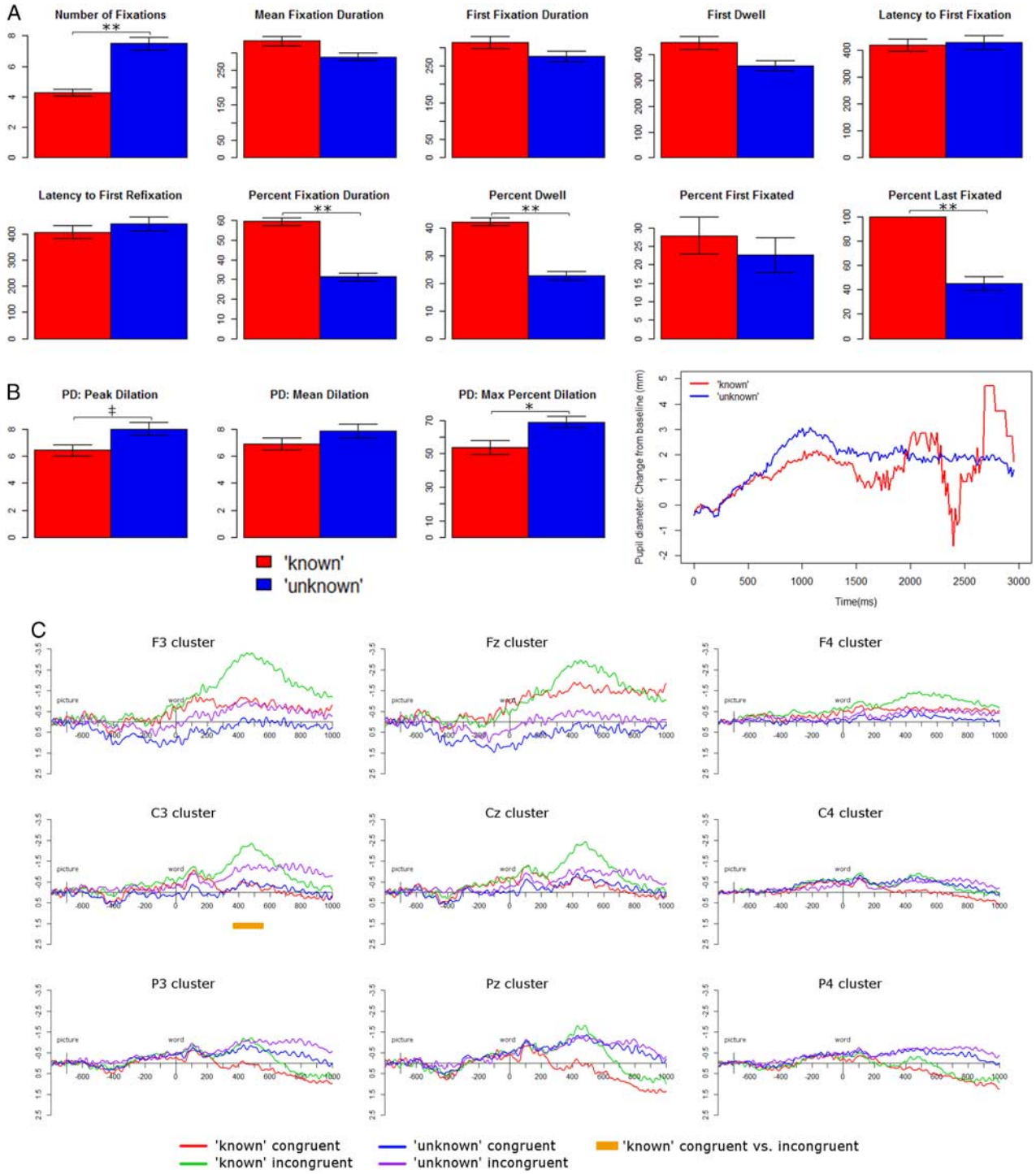


**FIGURE 7.** Results for P.B. **A:** Bar graphs comparing 'known' and 'unknown' trials for each of the EM variables. **B:** Comparisons of 'known' and 'unknown' trials for each of the pupillometry variables. Error bars indicate standard error of the mean. Significant differences or trends toward significance between 'known' and 'unknown' words, based on permutation tests with Bonferroni corrections, are indicated by asterisks (*Significant at $P < 0.05$; **Significant at $P < 0.01$; ‡Statistical trend at $P < 0.10$). **C:** Event-related potential data for all conditions at the nine electrode cluster sites. Negative is plotted up. The orange bar beneath the waveforms indicates significant differences between congruent and incongruent conditions for 'known' words, as determined by permutation tests with a cluster-based family-wise error correction at $P < 0.05$.

**TABLE 4.** Summary of Results for Each Measure for Each Participant

| | Participant | | | | |
|---|---|---|---|---|---|
| **Measure** | **D.L.** | **H.D.** | **W.F.** | **S.E.** | **P.B.** |
| Eye movement monitoring | No statistically significant effects | No statistically significant effects | Significant effects for *mean fixation duration, first fixation duration, first dwell, percent fixation duration, percent last fixated,* and *number of fixations* | Significant effect for *percent last fixated* | Significant effects for *percent fixation duration on stimulus, percent dwell, percent last fixated,* and *number of fixations* |
| Pupillary dilation | No statistically significant effects | No statistically significant effects | No statistically significant effects | No statistically significant effects | Significant effects for *peak dilation* and *max percent dilation* |
| Event-related potential | No statistically significant N400 effects | No statistically significant N400 effects | Significant N400 effect for 'known' words at Pz cluster from 200 to 400 ms | No statistically significant N400 effects | Significant N400 effect for 'known' words at C3 cluster from 400 to 550 ms |

congruent and incongruent separately. (An N400 effect cannot be estimated for individual trials because this effect is calculated by subtracting congruent from incongruent trials.)

For within-subject reliability, we ran split-half correlations between even and odd trials using Pearson correlations. A Spearman-Brown correction was applied to account for the reduction in power resulting from halving the data (Burnett, 1974). The *r* correlation coefficients for the split-halves reliability for the EM, PD, and ERP measures are shown in Tables 5 through 7. As can be seen in these tables, the correlation coefficients were overall quite low. The two exceptions were the latency measures in the EM data—*latency to first fixation* and *latency to first refixation*—which both showed very high correlations (all *r* values $>0.98$) for all of the participants and conditions. Aside from the EM latency measures, there were only a handful of correlations that were greater than 0.5 (indicating a moderate relationship), and there did not seem to be any systematic patterns between conditions, participants, or measures.

For test-retest reliability, we ran correlations between each session for each measure (again for 'known' and 'unknown' trials separately) using Pearson correlations, for those participants who completed more than one session. In the EM and PD data, only S.E. performed multiple sessions. In the ERP data, all of the participants except H.D. performed multiple sessions. The *r* coefficients for the EM, PD, and ERP measures are shown in Tables 8 through 10. Once again, the EM latency measures showed high test-retest reliability (all *r* values $>0.93$), but there were no other correlations greater than 0.5 (indicating a moderate relationship) in either the PD or ERP measures.

## DISCUSSION

Using a case-study approach, we investigated whether three implicit measures—EMs, PD, and ERPs—could provide within-subject assessments of receptive vocabulary knowledge in five individuals with Level 3 autism and varying levels of verbal ability. Based on previous results in TD adults (Ledoux et al, 2016), we predicted faster EMs and longer fixation durations, smaller PD, and larger N400 effects for 'known' than 'unknown' words. The results revealed notable differences among the participants in terms of which variables, if any, distinguished between 'known' and 'unknown' words.

### EM Monitoring

In Ledoux et al (2016), all EM measures showed differences between 'known' and 'unknown' words. Here, only three participants (W.F., S.E., and P.B.) showed significant effects on a subset of the EM variables. All three participants showed significant differences in *percent last fixated*. (Trends in the expected direction were also observed on this variable for D.L. and H.D.) Because this variable is a percentage measure, and therefore collapses across trials, it may be more robust to trial-by-trial variability, which may explain why it was better able to estimate vocabulary knowledge across participants in the current study. W.F. and P.B. also showed significant differences in *number of fixations* and *percent fixation duration*. Only W.F. showed significant differences in *average fixation duration*, *first fixation duration*, and *first dwell*. These effects all replicated those found in TD adults (Ledoux et al, 2016). Importantly, the fact that some convergence occurred across participants in the measures eliciting significant differences may indicate that certain EM variables (specifically *percent last fixated*, *number of fixations*, and *percent fixation duration*) may be more sensitive in distinguishing 'known' and 'unknown' words in individuals with Level 3 autism. These measures may be the most valuable for future studies using this paradigm to assess vocabulary knowledge in this group.

In comparison, some variables (particularly latency measures) were less sensitive in distinguishing 'known' and 'unknown' words across participants. The nonsignificant effects in latency measures in our participants could reflect baseline abnormalities in EM patterns. For example, Schmitt et al (2014) observed slower, longer, and less accurate saccades in individuals with ASD compared with TD individuals. Such baseline differences could have

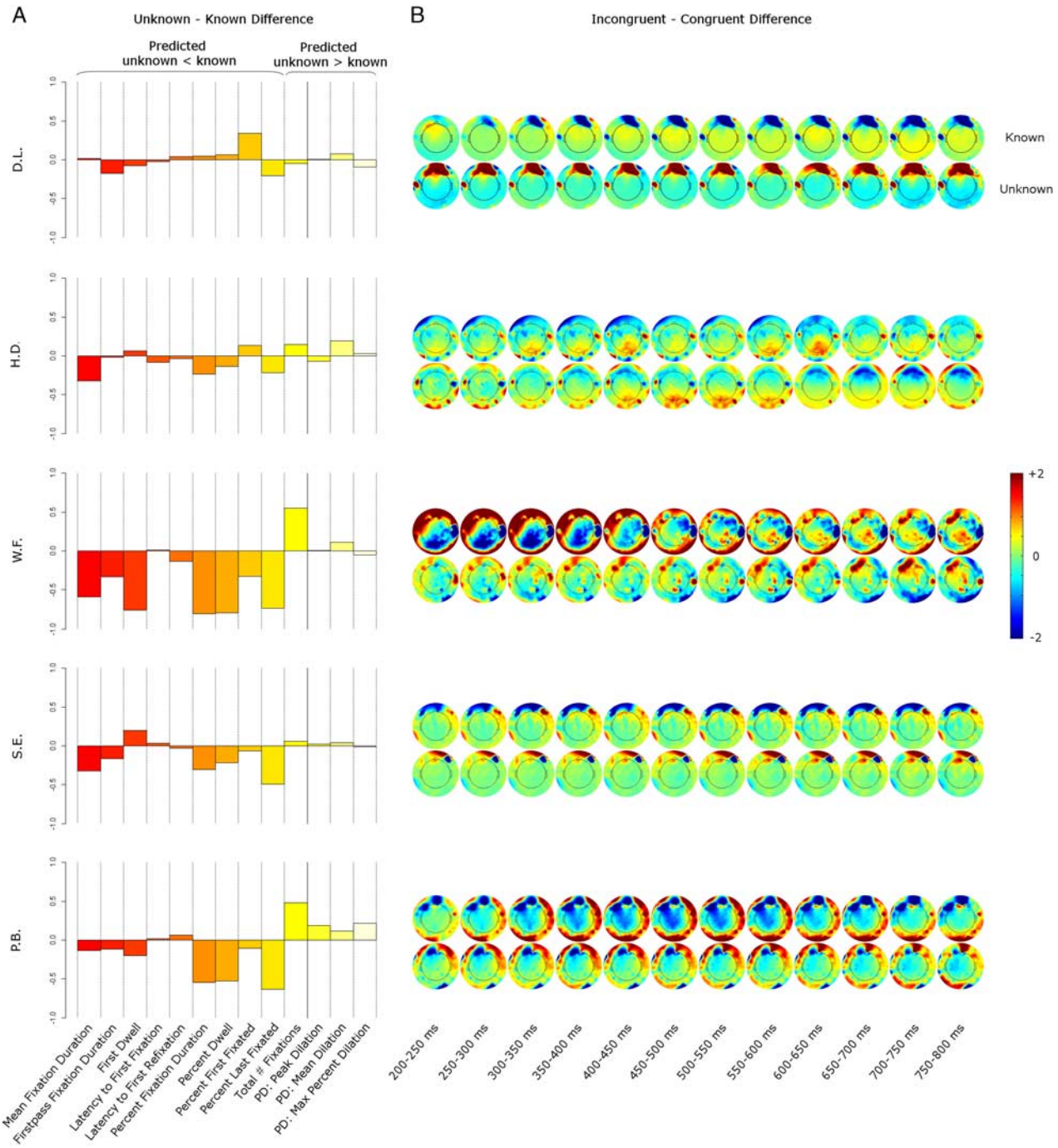


**FIGURE 8.** Summary descriptions of individual participant data. **A:** 'Unknown' – 'known' difference scores (scaled *z* scores) for each eye movement and pupil dilation variable. Variables on the left were predicted to be larger for 'unknown' than 'known' word trials, and so the 'unknown' – 'known' difference score should be negative. Variables on the right were predicted to be larger for 'known' than 'unknown' word trials, and so the 'unknown' – 'known' differences should be positive. **B:** Topographic plots of the event-related potential incongruent – congruent difference for 'known' and 'unknown' words in 50-millsecond windows from 200 to 800 milliseconds after sound presentation.

minimized 'known' and 'unknown' differences in the EM latency measures. We also observed no differences in *percentage of trials first fixated*, which could be explained by other idiosyncratic EM patterns in individuals with ASD such as strategic viewing patterns: These participants may be more likely to scan all pictures in the same order on every trial (eg, top-left, top-right, bottom-left, bottom-right) before returning to dwell on the target.

**TABLE 5.** Split-halves Reliability for the Eye Movement (EM) Measures

| | | EM Variables | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Participant | Condition | Total Number of Fixations | Mean Fixation Duration | First Fixation Duration | First Dwell on Stimulus | Latency to First Fixation | Latency to First Refixation | % Fixation Duration on Target | % Dwell on Target | % Trials First Fixated | % Trials Last Fixated |
| D.L. | Known | 0.70 | −0.22 | −0.45 | 0.48 | 0.99 | 0.99 | 0.10 | 0.38 | 1.00 | 0.07 |
| | Unknown | 0.13 | 0.13 | 0.29 | −0.41 | 0.99 | 0.97 | −0.51 | −0.57 | −0.72 | −0.46 |
| H.D. | Known | 0.47 | 0.31 | −0.28 | −0.14 | 1.00 | 0.99 | 0.33 | 0.35 | 0.51 | −0.46 |
| | Unknown | 0.73 | 0.66 | −0.22 | 0.05 | 0.99 | 1.00 | 0.47 | 0.33 | 0.24 | 0.45 |
| W.F. | Known | 0.26 | 0.00 | 0.07 | −0.27 | 1.00 | 1.00 | 0.38 | 0.26 | −0.26 | 0.11 |
| | Unknown | 0.11 | 0.02 | −0.25 | −0.30 | 1.00 | 1.00 | 0.19 | 0.26 | −0.38 | 0.02 |
| S.E. | Known | −0.02 | 0.37 | 0.04 | −0.57 | 0.99 | 0.99 | −0.05 | −0.09 | −0.40 | 0.25 |
| | Unknown | 0.17 | 0.33 | −0.48 | 0.39 | 0.99 | 0.99 | −0.15 | −0.13 | 0.08 | −0.20 |
| P.B. | Known | −0.01 | 0.26 | −0.06 | 0.13 | 1.00 | 1.00 | −0.10 | −0.16 | 0.18 | 1.00 |
| | Unknown | 0.40 | 0.14 | 0.36 | 0.15 | 1.00 | 1.00 | −0.04 | 0.06 | 0.24 | −0.50 |

Split-halves reliability was calculated for each condition by correlating odd trials with even trials using Pearson correlations. A Spearman-Brown correction was administered to account for the reduction in power. (For negative correlations, the Spearman-Brown correction was performed on the absolute value of the correlation coefficient and then negated.)

## Pupillary Dilation

Only P.B. showed differences between 'known' and 'unknown' words in the PD measures, specifically for *peak dilation* (although a statistical trend) and *max percent dilation.* These effects were larger for 'unknown' than 'known' words, replicating the pattern observed in TD adults (Ledoux et al, 2016). Although H.D. and W.F. showed nonsignificant trends in the expected direction for *mean dilation*, S.E. showed little difference between 'known' and 'unknown' words in any PD measures, and D.L. even showed a trend in the unexpected direction for *max percent dilation* (greater for 'known' than 'unknown' words).

**TABLE 6.** Split-halves Reliability for the Pupillary Dilation (PD) Measures

| | | PD Variable | | |
|---|---|---|---|---|
| Participant | Condition | Peak Dilation | Mean Dilation | Max % Dilation |
| D.L. | Known | −0.67 | −0.13 | −0.37 |
| | Unknown | −0.19 | −0.51 | 0.03 |
| H.D. | Known | 0.01 | 0.16 | 0.36 |
| | Unknown | 0.19 | −0.18 | −0.39 |
| W.F. | Known | −0.02 | −0.33 | −0.12 |
| | Unknown | −0.58 | −0.17 | −0.19 |
| S.E. | Known | 0.25 | −0.12 | 0.31 |
| | Unknown | 0.04 | −0.12 | 0.01 |
| P.B. | Known | −0.28 | 0.03 | −0.28 |
| | Unknown | 0.08 | −0.30 | 0.19 |

Split-halves reliability was calculated for each condition by correlating odd trials with even trials using Pearson correlations. A Spearman-Brown correction was administered to account for the reduction in power. (For negative correlations, the Spearman-Brown correction was performed on the absolute value of the correlation coefficient and then negated.)

Overall, these highly variable patterns suggest that the utility of PD measures for distinguishing 'known' and 'unknown' words differs between individuals, and that PD as an implicit measure of cognitive processing in individuals with Level 3 autism may be generally less informative than other measures.

## Event-related Potentials

W.F. and P.B. showed significant N400 effects only for 'known' words. This pattern replicates those observed in TD adults (Ledoux et al, 2016) and demonstrates that the N400 successfully distinguished between 'known' and 'unknown' vocabulary for these participants. The N400 effect occurred at the Pz cluster from approximately 200 milliseconds to 400 milliseconds for W.F. and at the C3 cluster from approximately 400 milliseconds to 550 milliseconds for P.B. The N400 typically occurs over the centroparietal scalp and anywhere from 200 milliseconds to 600 milliseconds in TD adults (Kutas and Federmeier, 2011). Thus, the N400 topographies and latencies for W.F. and P.B. are consistent with previous literature. D.L., H.D., and S.E. did not show N400 effects for either 'known' or 'unknown' words. These findings suggest that ERPs may be better suited as implicit measures of receptive vocabulary in some participants than others.

## Additional Considerations

Overall, these results suggest that EMs, PD, and ERPs can provide implicit assessments of receptive vocabulary in individuals with Level 3 autism, but that some measures are better suited for certain participants than others. Only P.B. showed significant effects in all three measures. W.F. showed effects only in the EM and ERP measures, and S.E. showed effects only in the EM

**TABLE 7.** Split-halves Reliability for the Event-related Potential Measures

| | | | Electrode Cluster | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Participant** | **Condition** | **Congruency** | **F3** | **Fz** | **F4** | **C3** | **Cz** | **C4** | **P3** | **Pz** | **P4** |
| D.L. | Known | Congruent | 0.27 | −0.12 | 0.17 | 0.29 | −0.16 | −0.04 | 0.25 | −0.08 | −0.27 |
| | | Incongruent | −0.14 | −0.06 | 0.01 | 0.10 | −0.40 | −0.10 | 0.41 | −0.47 | −0.36 |
| | Unknown | Congruent | 0.27 | −0.21 | −0.24 | −0.07 | −0.22 | 0.10 | −0.18 | 0.27 | 0.27 |
| | | Incongruent | 0.12 | 0.17 | 0.39 | −0.39 | 0.25 | 0.34 | −0.37 | −0.09 | 0.07 |
| H.D. | Known | Congruent | 0.12 | −0.25 | −0.35 | −0.03 | −0.39 | −0.30 | 0.01 | −0.12 | −0.23 |
| | | Incongruent | 0.07 | −0.39 | −0.56 | −0.32 | −0.23 | −0.35 | 0.15 | 0.22 | 0.14 |
| | Unknown | Congruent | −0.28 | −0.25 | −0.19 | 0.22 | 0.20 | 0.05 | 0.05 | 0.11 | −0.02 |
| | | Incongruent | 0.38 | 0.32 | 0.53 | 0.63 | 0.58 | 0.56 | 0.32 | 0.36 | 0.46 |
| W.F. | Known | Congruent | −0.03 | −0.24 | −0.24 | 0.06 | −0.13 | −0.21 | 0.08 | −0.18 | −0.31 |
| | | Incongruent | 0.11 | 0.45 | −0.18 | 0.08 | 0.25 | 0.16 | 0.17 | 0.11 | 0.41 |
| | Unknown | Congruent | 0.03 | 0.06 | 0.28 | −0.06 | −0.13 | 0.07 | 0.12 | 0.16 | 0.43 |
| | | Incongruent | −0.27 | −0.31 | −0.07 | −0.15 | 0.26 | −0.02 | −0.17 | −0.13 | 0.42 |
| S.E. | Known | Congruent | 0.13 | 0.30 | 0.18 | 0.15 | 0.35 | 0.48 | 0.25 | 0.15 | 0.42 |
| | | Incongruent | 0.15 | 0.23 | 0.55 | −0.46 | −0.35 | −0.15 | −0.61 | −0.11 | 0.07 |
| | Unknown | Congruent | 0.50 | 0.46 | 0.53 | 0.39 | 0.48 | 0.30 | 0.13 | 0.31 | −0.04 |
| | | Incongruent | −0.15 | −0.48 | −0.39 | 0.14 | −0.17 | −0.39 | 0.28 | 0.06 | −0.16 |
| P.B. | Known | Congruent | −0.23 | 0.14 | −0.05 | −0.02 | 0.20 | 0.42 | 0.37 | 0.34 | 0.25 |
| | | Incongruent | −0.18 | −0.51 | −0.04 | 0.18 | 0.24 | 0.12 | 0.22 | 0.29 | −0.12 |
| | Unknown | Congruent | −0.30 | 0.03 | −0.01 | −0.29 | −0.32 | 0.23 | 0.05 | −0.25 | 0.42 |
| | | Incongruent | −0.04 | 0.10 | 0.03 | 0.23 | 0.31 | 0.33 | 0.14 | 0.11 | 0.17 |

Split-halves reliability was calculated for each condition by correlating odd trials with even trials using Pearson correlations. A Spearman-Brown correction was administered to account for the reduction in power. (For negative correlations, the Spearman-Brown correction was performed on the absolute value of the correlation coefficient and then negated.) Note that an N400 effect cannot be estimated for individual trials because this effect is calculated by subtracting congruent from incongruent trials. For each trial, amplitude was averaged over a window from 350 to 450 milliseconds (which includes the windows of significant N400 effects observed in W.F. and P.B.).

measures. D.L. and H.D. showed no significant effects on any measure. Individual differences also occurred with regard to which variable(s) best distinguished 'known' and 'unknown' words. For instance, the PD data were highly variable between the participants, showing significant differences between 'known' and 'unknown' words only in one participant.

In the ERP data, variability occurred in the overall strength of the brain activity: Some participants had clear peaks in early perceptual components and/or robust N400 effects, whereas others showed generally reduced amplitudes. These differences could result from individual variability in the number of trials, the amount of endogenous neural noise, or atypical neural responses in general.

**TABLE 8.** Test-Retest Reliability for the Eye Movement (EM) Measures

| | | EM Variable | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Participant** | **Condition** | **Total Number of Fixations** | **Mean Fixation Duration** | **First Fixation Duration** | **First Dwell on Stimulus** | **Latency to First Fixation** | **Latency to First Refixation** | **% Fixation Duration on Target** | **% Dwell on Target** | **% Trials First Fixated** | **% Trials Last Fixated** |
| D.L. | Known | | | | | Not applicable | | | | | |
| | Unknown | | | | | | | | | | |
| H.D. | Known | | | | | Not applicable | | | | | |
| | Unknown | | | | | | | | | | |
| W.F. | Known | | | | | Not applicable | | | | | |
| | Unknown | | | | | | | | | | |
| S.E. | Known | −0.11 | −0.13 | −0.01 | 0.36 | 0.93 | 0.98 | −0.54 | −0.46 | −0.48 | −0.52 |
| | Unknown | −0.07 | 0.29 | 0.67 | 0.48 | 0.96 | 0.95 | 0.35 | 0.27 | 0.36 | 0.10 |
| P.B. | Known | | | | | Not applicable | | | | | |
| | Unknown | | | | | | | | | | |

Test-retest reliability was calculated for each condition by correlating trials on each session for participants who completed more than one session.

**TABLE 9.** Test-Retest Reliability for the Pupillary Dilation (PD) Measures

| | | PD Variable | | |
|---|---|---|---|---|
| **Participant** | **Condition** | **Peak Dilation** | **Mean Dilation** | **Max % Dilation** |
| D.L. | Known | −0.13 | −0.02 | −0.20 |
| | Unknown | −0.03 | 0.17 | 0.23 |
| H.D. | Known | −0.13 | −0.02 | −0.20 |
| | Unknown | −0.03 | 0.17 | 0.23 |
| W.F. | Known | −0.13 | −0.02 | −0.20 |
| | Unknown | −0.03 | 0.17 | 0.23 |
| S.E. | Known | −0.13 | −0.02 | −0.20 |
| | Unknown | −0.03 | 0.17 | 0.23 |
| P.B. | Known | −0.13 | −0.02 | −0.20 |
| | Unknown | −0.03 | 0.17 | 0.23 |

Test-retest reliability was calculated for each condition by correlating trials on each session for participants who completed more than one session.

It did not escape our notice that the three participants who showed significant effects in some or all of the implicit measures (W.F., S.E., and P.B.) were also the three participants who were best able to complete the behavioral testing (ie, KBIT–2 and PPVT–4), although W.F and S.E. demonstrated intelligence and receptive language scores within the below-average range. As mentioned earlier, these participants may have had difficulties maintaining attention and/or understanding task instructions, and so these scores may not be accurate reflections of their true abilities. This potential lack of reliability in behavior on the standardized measures highlights the benefit of using implicit measures to assess receptive vocabulary even in individuals who have more functional expressive language.

P.B. was the only one with average to above-average scores on the standardized assessments; he also showed significant effects in all of the implicit measures. Although we did not perform an item-by-item analysis, nor directly compare the results of the behavioral standardized assessments and the ERP results, the relative alignment between the behavioral and implicit measures for P.B. may offer some confirmation that these implicit measures are reflecting receptive vocabulary knowledge range (Byrne et al, 1995; Connolly and D'Arcy, 2000; Connolly et al, 1995).

The two minimally verbal participants, D.L. and H.D., did not show any statistically significant effects on any measures (although several variables showed trends in the expected direction) and also were not able to complete the behavioral testing. This may suggest that these implicit measures are not as useful for individuals with Level 3 autism who have limited functional language.

With regard to the within-subject reliability and test-retest reliability analyses, the reliability estimates overall were quite low, with considerable negative values in the split-half reliability estimates (meaning that reliability was extremely low, as only positive correlations would be expected; see Burnett, 1974). This finding may suggest that these measures are not reliable at all, although it is also

**TABLE 10.** Test-Retest Reliability for the Event-related Potential Measures

| | | | Electrode Cluster | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Participant** | **Condition** | **Congruency** | **F3** | **Fz** | **F4** | **C3** | **Cz** | **C4** | **P3** | **Pz** | **P4** |
| D.L. | Known | Congruent | −0.10 | −0.21 | −0.29 | 0.10 | −0.34 | −0.21 | 0.20 | 0.24 | 0.07 |
| | | Incongruent | −0.11 | −0.13 | 0.29 | 0.02 | 0.10 | 0.23 | 0.19 | 0.21 | 0.02 |
| | Unknown | Congruent | 0.12 | −0.08 | −0.06 | −0.08 | −0.02 | −0.01 | −0.27 | 0.04 | −0.01 |
| | | Incongruent | −0.01 | 0.14 | 0.11 | 0.09 | 0.12 | 0.13 | −0.07 | −0.11 | 0.00 |
| H.D. | Known | Congruent | Not applicable | | | | | | | | |
| | | Incongruent | | | | | | | | | |
| | Unknown | Congruent | | | | | | | | | |
| | | Incongruent | | | | | | | | | |
| W.F. | Known | Congruent | 0.10 | −0.03 | −0.13 | 0.23 | 0.03 | −0.02 | 0.06 | 0.15 | −0.20 |
| | | Incongruent | −0.07 | −0.18 | 0.20 | −0.01 | 0.01 | 0.03 | 0.06 | −0.02 | 0.01 |
| | Unknown | Congruent | 0.08 | −0.17 | −0.18 | 0.07 | −0.16 | −0.36 | −0.03 | −0.36 | −0.08 |
| | | Incongruent | 0.07 | 0.08 | 0.23 | 0.11 | −0.04 | −0.04 | −0.19 | −0.26 | −0.22 |
| S.E. | Known | Congruent | 0.02 | −0.12 | −0.05 | 0.00 | −0.06 | −0.33 | 0.04 | −0.14 | −0.20 |
| | | Incongruent | 0.13 | −0.10 | 0.06 | 0.15 | 0.10 | −0.01 | 0.17 | −0.01 | −0.11 |
| | Unknown | Congruent | −0.07 | −0.03 | −0.08 | −0.04 | 0.06 | 0.13 | 0.14 | 0.42 | 0.25 |
| | | Incongruent | −0.12 | 0.00 | −0.17 | −0.14 | −0.17 | −0.09 | −0.05 | −0.20 | −0.11 |
| P.B. | Known | Congruent | 0.19 | 0.01 | −0.07 | 0.11 | 0.13 | −0.04 | −0.06 | −0.02 | −0.08 |
| | | Incongruent | 0.03 | 0.12 | 0.07 | −0.16 | −0.14 | −0.07 | −0.18 | −0.14 | −0.10 |
| | Unknown | Congruent | 0.21 | 0.30 | 0.21 | 0.11 | 0.00 | −0.09 | −0.02 | −0.02 | 0.15 |
| | | Incongruent | 0.17 | 0.18 | 0.02 | 0.02 | 0.14 | −0.08 | −0.06 | 0.02 | −0.16 |

Test-retest reliability was calculated for each condition by correlating trials on each session for participants who completed more than one session. (Note that an N400 effect cannot be estimated for individual trials since this effect is calculated by subtracting congruent from incongruent trials.) For each trial, amplitude was averaged over a window from 350 to 450 milliseconds (which includes the windows of significant N400 effects observed in W.F. and P.B.).

important to note that within-subject reliability measures for these types of paradigms are largely unavailable even for TD populations, making it difficult to determine how anomalous our results are for this highly heterogeneous population of individuals with Level 3 autism.

The EM latency measures were highly reliable within the study participants. This finding is interesting when contrasted with the findings in the main data, where the latency measures were the least able of all EM measures to distinguish between 'known' and 'unknown' words. This finding raises an important point: The lack of high reliability for these measures does not necessarily speak against their use as implicit measures of receptive vocabulary, that is, their ability to discriminate between 'known' and 'unknown' conditions more broadly. Rather, low reliability may be a result of intertrial variability, which is known to be high in individuals with Level 3 autism (Adamo et al, 2014; Pérez Velázquez and Galán, 2013). Because we were expecting differences between 'known' and 'unknown' conditions, we performed reliability estimates separately for each condition since combining them could have skewed the reliability ratings. Low reliability values thus indicate high variability between individual trials in the 'known' condition or between individual trials in the 'unknown' condition. However, even if there was significant intertrial variability *within* each condition, some participants and some measures still showed significant differences *between* 'known' and 'unknown' words. Therefore, the reliability estimates may speak more to the intertrial variability within either 'known' words or 'unknown' words, but not to the reliability of these implicit measures in distinguishing between 'known' and 'unknown' vocabulary. Ultimately, the results of these reliability analyses instead seem to support our main findings: There is significant individual variability both between participants and between variables within participants.

Our findings that some measures were better suited to some individuals than others echo the mixed findings of previous work using implicit measures in individuals with Level 3 autism. For instance, Plesa Skwerer et al (2016), using behavioral, caregiver report, and eye-tracking measures to assess receptive vocabulary in minimally verbal individuals, reported significant individual variability with no one method showing an across-the-board advantage for all of the participants. Venker et al (2013), using an eye-tracking paradigm in participants with a range of symptom severities, found notable individual differences in performance. When looking at individual data in an ERP paradigm, Cantiani et al (2016) also reported variability in the presence of an N400 effect in minimally verbal individuals.

Although the number of studies remains small, when combined with our results, this body of work quite consistently points to the notable individual variability in the utility of implicit measures of receptive vocabulary in individuals with Level 3 autism. This suggests that rather than continuing to test whether these implicit measures are useful for this population or show differences between individuals with ASD and TD individuals at the group level, research should instead begin to focus on how to optimize their utility on an individual basis in order to provide estimates of vocabulary knowledge on by-subject and by-item bases.

## Practical Recommendations for Testing

Implicit measures are attractive because of their potential to allow for by-subject and by-item assessments of words that an individual does or does not know, which may be difficult to assess using traditional behavioral testing. The current work suggests that, in some individuals, EMs, PD, and/or ERPs can provide estimates of receptive vocabulary knowledge. However, it also emphasizes the need to determine on an individual basis which measures are most informative for a given participant. Although we did not perform item-level determinations of vocabulary knowledge here, this technique offers that potential advantage, which we hope to explore in future research. This section provides some theoretical practical recommendations for pilot testing and intervention design for researchers who wish to use implicit measures to assess receptive vocabulary on single-subject and single-item bases in individuals with Level 3 autism.

First, researchers should perform pilot testing using a range of implicit measures to determine which measure (or measures) best predicts receptive vocabulary in a particular individual. Paradigms like the ones employed in the current study could be used, or others could be specially designed to test the construct(s) of interest. Stimuli should include words with an extremely high likelihood that the participant does or does not know (which can be approximated from parental report). Words selected to represent the extreme cases of knowledge will be more likely to allow clear differentiation of the implicit measures to the extent that such differentiation is possible. It is important at this stage of pilot testing to include a variety of implicit measures to allow for the possibility that some may be better indices of vocabulary knowledge than others. Based on the results from the current data, we suggest that *percentage of trials last fixated* be included in EM measures because this variable showed the most consistent results across all of our participants. At this stage of pilot testing, researchers can also determine any modifications needed to the testing procedures to facilitate data collection.

At this stage, researchers may also wish to perform power calculations to estimate the approximate number of trials needed to observe a given effect size. These calculations may help to determine the feasibility of the protocol for testing a particular individual. For example, if a measure would require hundreds of usable trials with a participant who has only a small vocabulary, a great deal of word repetition would be required, which should be taken into consideration before moving forward. Similarly, if a measure would require extensive amounts of data to yield reliable vocabulary estimates, and the participant would require extensive training or modifications to obtain clean data, researchers may need to consider the

 

time, effort, and costs associated with testing to determine feasibility.

The data from this pilot testing should be analyzed using single-subject statistical analyses (such as permutation tests as used here) to identify which implicit measures showed the largest differences between 'known' and 'unknown' words. At this stage, the researchers may wish to compute split-halves reliability to estimate the internal consistency of participant responses for each measure. From these initial reliability estimates, researchers can then determine the number of trials needed to attain a desired level of reliability. It is possible that the reliability estimates may vary between conditions and measures even for a single participant (Tables 5–10). In this case, more trials may be needed for certain measures than others. Nevertheless, such an approach would help guide researchers to approximate the number of trials needed to observe the desired effects and determine if more data should be collected in the pilot testing phase.

Researchers can also perform additional power analyses at this stage to estimate the number of trials needed to observe a desired effect size, given the magnitude of an individual's effect in the pilot data. As seen in the current data, effect sizes may differ between measures and participants. If no significant effects are observed after initial pilot testing, researchers may wish to revisit their power calculations and consider increasing the number of data points. It is possible that none of the chosen measures will show statistically significant differences, as was true for D.L. and H.D. in the current data, despite adequate amounts of data. In such a case, this might suggest that implicit measures may not be the best choice of assessment in that particular individual. It is left to the researchers' discretion whether overall patterns and statistical trends in the data are informative for selecting variables for future study in the absence of statistical significance.

Once the implicit measures of choice have been selected, the next step will be to select a pool of target words for which knowledge status is less clear (ie, words that the participant may or may not know), which will be the subject of study. In this case, these "uncertain" words can be included as stimuli in the same paradigms used in pilot testing, although potentially with a restricted focus on the variables of interest as determined from pilot testing. Ideally, analyzing data from the implicit measures of choice on these target words can inform researchers as to whether each target word is relatively familiar (eg, if it shows similar effect sizes as 'known' words in the pilot testing) or unfamiliar (eg, if it shows no effect, or smaller effects than 'known' words in the pilot testing). Results from the implicit measures can thus be used to separate the pool of target words into those that are more and less known to the participant, and can subsequently be used to inform further interventions (eg, provide more training on words that are less familiar).

Alternatively, target words could be the focus of intervention (eg, teaching new vocabulary words). In this case, these new 'unknown' words can be included in the paradigms in a similar manner as described above. Before training on these new vocabulary words, the words should show effect sizes in the implicit measures similar to the 'unknown' words in the pilot testing. Testing can be repeated at the end of training to determine whether the participant has successfully learned the meaning of the target words: after training, target words that show effect sizes similar to the 'known' words in the pilot testing can be said to have been learned by the participants, whereas those with smaller effect sizes or no differences would require further training. It would be of potential theoretical and practical importance to compare the participants' behavioral performance on these same words to changes noted on the implicit measures to determine, for example, whether behavioral gains are generally preceded by more implicit gains or whether implicit gains can act as predictors of behavioral change.

These practical recommendations are intended for researchers working with this population, rather than for incorporating these implicit measures into clinical practice. Although it is our hope that implicit measures may one day be used to quickly and objectively estimate an individual's receptive vocabulary abilities, at present, these techniques require further refinement and testing in clinical populations, particularly individuals with Level 3 autism, before being translated into clinical practice.

Based on our own experience testing individuals with Level 3 autism (including those who are functionally nonspeaking and/or nonverbal) with these implicit measures, we offer several suggestions in terms of possible modifications. (See also Kylliäinen et al, 2014, for a review of practical guidelines for testing individuals with ASD.) Because of the overall difficulty that many participants have with compliance to the testing protocols, it is useful to take steps to ensure that usable data are collected. First, it is helpful to conduct extensive procedures to orient and desensitize some participants to the equipment used to collect EM, PD, and ERP data. Our eye-tracking and ERP equipment required individuals to tolerate the application of some form of equipment to their head, something that is uncomfortable for many participants, especially for those with autism. We have designed or purchased variants of the head-mounted equipment to practice application with participants at the lab or in their homes before actual testing sessions. Depending on the severity of initial aversion to the head-mounted equipment, we have found that some participants might need extensive desensitization training over weeks or even months before they will tolerate the equipment. Researchers might even work with participants' family members or clinical team members to develop acclimation procedures and routines that align with other aspects of the participants' current behavioral interventions. Taking the time to work with individual participants to increase tolerance can dramatically improve data collection when actual testing begins.

Once participants are acclimated to the head-mounted equipment, we have found it beneficial to acclimate them to the laboratory testing space and research

 

personnel to the extent possible. Short visits to tour the lab space before testing with opportunities to sit in front of the eye tracker or the ERP testing computer could be helpful. Because continuity in staffing is beneficial to testing individuals with Level 3 autism, we have tried to ensure that during each visit, a participant is paired with the same research team member who is present at all visits and is a part of all research-related interaction. The development of rapport with the participant, the attendance to issues of comfort in the physical space, and the knowledge of participants' idiosyncratic needs are generally considered to be of benefit when testing all research participants, particularly when working with individuals with special needs such as those with Level 3 autism.

As part of our eye-tracking and ERP protocols with TD adults, we provided instructions beyond specific task demands that can improve data collection, such as minimizing movement and timing eye blinks to intervals between trials. Understanding such instructions may depend on the functional communication levels of the individual participants, and adherence to these instructions will vary even among those who might understand them. Most eye-tracking and ERP testing software allows online experimenter marking of trials with artifacts so trials may be analyzed separately or removed during analysis. Additionally, it has been beneficial to videotape participants during the experimental sessions in order to allow researchers to review the data offline to mark trials for rejection based on factors such as excessive movement or a clear lack of attention to the stimulus presentation computer (see the Methods section). Although the additional steps of video analyses and data coding can add time to data preprocessing, these steps have helped to maximize the signal-to-noise ratio on included trials.

During data collection, consideration should also be given to maintaining the motivation of the participant. There is a temptation to collect as much data as possible in a single session, especially as the placement of the equipment for measuring EMs and ERPs can be time consuming and there is a benefit to ensuring that the placement is consistent across trials (ie, removing the equipment for breaks can lead to differences in calibration). However, these issues must be balanced with issues of participant comfort and motivation. We have achieved some success with participants with Level 3 autism by presenting stimuli in short blocks, separated by short intervals during which they have access to reinforcing stimuli (pictures or short video clips). Attempts possibly could be made to include motivating items among the stimuli either as target items or among filler items. In the current experiment, we included a number of foods and animals in our 'known' and 'unknown' word categories that might be more inherently interesting to participants while still meeting the desired stimulus characteristics. For all participants, including Level 3 participants, session length issues must be addressed with care because multiple testing sessions have disadvantages yet may be preferred in order to ensure participant motivation and comfort. Limiting testing sessions to approximately 1 hour has worked well for Level 3 participants, even if testing must occur over multiple sessions.

The costs and effort associated with data acquisition is an important consideration when determining whether to pursue implicit measures as a means of assessment. Use of these implicit measures requires access to sophisticated equipment, which may come with high price tags. Eye-tracking systems typically cost several thousands of dollars, and high-end EEG systems can run upwards of several hundreds of thousands. Partnering with research labs or hospitals may alleviate some of these costs. Data preprocessing and analyses may require specific expertise. There is also likely to be increased effort in testing participants, depending on the protocol modifications that are required. For instance, some participants may require several hours of training just to be able to tolerate an EEG net and may then require several repeat sessions to collect enough usable data. Although the increased costs and effort involved in using implicit measures are certainly not trivial, they may be well worth it if such measures offer the potential for accessing otherwise unavailable insight into an individual's cognitive functioning and language abilities.

Overall, the use of implicit measures for item-by-item identification of 'known' and 'unknown' words could have important clinical implications such as facilitating targeted therapeutic approaches. Knowing which words an individual comprehends could allow a language therapist to focus instruction on less understood words, thereby maximizing the use of clinical time and minimizing patient boredom and disengagement. Similarly, the more parents and caregivers know about an individual's comprehension abilities, the more successful their daily interactions and communication will be. Ultimately, on a case-by-case basis, researchers and clinicians should weigh the increased costs and effort associated with the collection of interpretable implicit data against the value of insight into an individual's language abilities that implicit measures may provide.

## Limitations

As noted in the Methods section, our stimuli differed in terms of length; that is, 'unknown' words were slightly longer (had more letters) than 'known' words. Because most of our comparisons were proportional or did not otherwise depend on latency, we do not believe that word length likely contributed to the differences (or lack thereof) observed in our implicit measures across word knowledge conditions. Although we cannot say for certain that this was the case, we acknowledge this difference as a potential limitation.

Similarly, our word knowledge categories are based on presumptions of a word being 'known' or 'unknown' to participants, which may be a limitation. Our presumptions were based initially on word frequency (with 'known' words used very frequently and 'unknown' words used very infrequently) and were corroborated by ratings of likelihood of knowledge by the individual participants and/or parents or caregivers who were familiar with the

 

participants' vocabulary. The true status of individual stimulus items known to each participant is something that we cannot determine. The extent to which 'known' items are actually unknown to individual participants, or vice versa, could work against confirmation of our hypothesis that implicit measures can detect word knowledge. In other words, some of our observations that some implicit measures did not distinguish between 'known' and 'unknown' conditions could be due to improper stimulus categorization rather than to the ineffectiveness of these measures in assessing word knowledge. A more individualized determination of words that are expected to be 'known' and 'unknown' on a participant-by-participant basis might better assess the suitability of these methods, and we suggest this direction for future research (eg, Coderre et al, unpublished data).

As noted in the literature (Kasari et al, 2013; Kylliäinen et al, 2014; Tager-Flusberg and Kasari, 2013; Tager-Flusberg et al, 2017), there are many difficulties with cognitive testing in individuals with Level 3 autism that have contributed to some limitations in the current study. For example, challenges to eye-tracking and EEG data collection, such as movement artifacts, are heightened. These challenges may require modifications to testing protocols to ensure participant comfort and engagement, to the number of testing sessions, and/or to data cleaning procedures to ensure maximum data retention. Our EM measures proved particularly sensitive to motion, as reflected in the low data retention rate. For EEG data, some individuals lost a large percentage of trials despite our extensive cleaning and preprocessing procedure. Four of our five participants performed two or more sessions. This repetition may have influenced the data; for example, the N400 effect is sensitive to repetition (Kutas and Federmeier, 2011). However, no participant saw the same stimulus more than three times, and multiple sessions were performed at least 1 month apart. Given the importance of collecting enough usable trials, the need for multiple sessions outweighed the potential repetition effects. Nevertheless, this factor should be considered in future studies using similar paradigms.

Despite these challenges, it is our hope that the contributions of this study to the relatively limited understanding of language comprehension in individuals with Level 3 autism outweighs the limitations that arise from issues of repeated testing. Our focus on individuals at the more severe end of the spectrum provides important information about a population that is woefully underrepresented in the autism literature. This work also demonstrates the importance of using case-study approaches with clinical populations and the utility of single-subject analyses for establishing implicit assessments in individual participants.

## CONCLUSIONS

We demonstrated that EMs, PD, and ERPs can provide implicit estimates of receptive vocabulary knowledge in individuals with Level 3 autism, although the participants differed in their individual sensitivity to specific measures, and some measures proved more able than others in discriminating 'known' and 'unknown' vocabulary between participants. This variability highlights the importance of tailoring these assessments to each individual. Despite the inevitable heterogeneity of our limited number of participants, this work is one of the few studies to use sophisticated neuropsychological methodologies, such as EEG and eye tracking, to examine language processing in individuals with Level 3 autism, thereby offering a rare insight into this population.

## ACKNOWLEDGMENTS

*The authors thank Ishanti Gangopadhyay, PhD, for her help with data collection and Nancy Grund, MBA, for her help with editing. The authors especially thank all of the participants and their families and/or caregivers and the Linwood Center of Ellicott City, Maryland, an approved institutional review board research site that was instrumental in research involving individuals on the autism spectrum.*

## REFERENCES

Adamo N, Huo L, Adelsberg S, et al. 2014. Response time intra-subject variability: commonalities between children with autism spectrum disorders and children with ADHD. *Eur Child Adolesc Psychiatry*. 23:69–79.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition: DSM–5. Washington, DC: American Psychiatric Association.

Anderson CJ, Colombo J. 2009. Larger tonic pupil size in young children with autism spectrum disorder. *Dev Psychobiol*. 51:207–211.

Applied Science Laboratories. 2009. ASL Results [Computer Software]. Bedford, Massachusetts: Applied Science Laboratories.

Bavin EL, Kidd E, Prendergast L, et al. 2014. Severity of autism is related to children's language processing. *Autism Res*. 7:687–694.

Beatty J, Lucero-Wagoner B. 2000. The pupillary system. In: Cacioppo JT, Tassinary LG, Berntson GG, eds. *Handbook of Psychophysiology*, 2nd ed. New York, New York: Cambridge University Press; 142–162.

Brady NC, Anderson CJ, Hahn LJ, et al. 2014. Eye tracking as a measure of receptive vocabulary in children with autism spectrum disorders. *Augment Altern Commun*. 30:147–159.

Brenner LA, Turner KC, Müller R-A. 2007. Eye movement and visual search: are there elementary abnormalities in autism? *J Autism Dev Disord*. 37:1289–1309.

Brock J, Norbury C, Einav S, et al. 2008. Do individuals with autism process words in context? Evidence from language-mediated eye-movements. *Cognition*. 108:896–904.

Burnett. 1974. Parallel measurements and the Spearman-Brown formula. *Educ Psychol Meas*. 34:785–788.

Byrne JM, Dywan CA, Connolly JF. 1995. Assessment of children's receptive vocabulary using event-related brain potentials: development of a clinically valid test. *Child Neuropsychol*. 1:221–223.

Cantiani C, Choudhury NA, Yu YH, et al. 2016. From sensory perception to lexical-semantic processing: an ERP study in non-verbal children with autism. *PLoS One*. 11:e0161637. doi.org/10.1371/journal.pone.0161637

Centers for Disease Control and Prevention. 2018. Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2014. *MMWR Surveill Summ*. 67:1–28.

Connolly JF, Byrne JM, Dywan CA. 1995. Assessing adult receptive vocabulary with event-related potentials: an investigation of cross-modal and cross-form priming. *J Clin Exp Neuropsychol*. 17:548–565.

Connolly JF, D'Arcy RC. 2000. Innovations in neuropsychological assessment using event-related brain potentials. *Int J Psychophysiol*. 37:31–47.

Davies M. 2008. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English: 450 Million Words,

 

1990–Present. Available at: https://www.english-corpora.org/coca/. Accessed January 11, 2015.
Davis G, Plaisted-Grant K. 2015. Low endogenous neural noise in autism. *Autism*. 19:351–362.
Dawson G, Rogers S, Munson J, et al. 2010. Randomized, controlled trial of an intervention for toddlers with autism: the early start Denver model. *Pediatrics*. 125:e17–e23.
Delorme A, Makeig S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods*. 134:9–21.
Delorme A, Sejnowski TJ, Makeig S. 2007. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage*. 34:1443–1449.
Dunn LM, Dunn DM. 2007. *Peabody Picture Vocabulary Test*, Fourth Edition. Bloomington, Minnesota: NCS Pearson Inc.
Dunn MA, Gaughan H Jr, Kreuzer J, et al. 1999. Electrophysiologic correlates of semantic classification in autistic and normal children. *Dev Neuropsychol*. 16:79–99.
Goldberg MC, Lasker AG, Zee DS, et al. 2002. Deficits in the initiation of eye movements in the absence of a visual target in adolescents with high functioning autism. *Neuropsychologia*. 40:2039–2049.
Granholm E, Asarnow RF, Sarkin AJ, et al. 1996. Pupillary responses index cognitive resource limitations. *Psychophysiology*. 33:457–461.
Groppe DM, Makeig S, Kutas M. 2009. Identifying reliable independent components via split-half comparisons. *Neuroimage*. 45:1199–1211.
Groppe DM, Urbach TP, Kutas M. 2011. Mass univariate analysis of event-related brain potentials/fields I: a critical tutorial review. *Psychophysiology*. 48:1711–1725.
Jung TP, Makeig S, Humphries C, et al. 2000. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*. 37:163–178.
Kasari C, Brady N, Lord C, et al. 2013. Assessing the minimally verbal school-aged child with autism spectrum disorder. *Autism Res*. 6:479–493.
Kaufman A, Kaufman N. 2004. *Kaufman Brief Intelligence Test*, Second Edition (KBIT–2). Circle Pines, Minnesota: American Guidance Service.
Kuipers J-R, Thierry G. 2011. N400 amplitude reduction correlates with an increase in pupil size. *Front Hum Neurosci*. 5:1–5.
Kuipers J-R, Thierry G. 2013. ERP-pupil size correlations reveal how bilingualism enhances cognitive flexibility. *Cortex*. 49:2853–2860.
Kutas M, Federmeier KD. 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu Rev Psychol*. 62:621–647.
Kutas M, Hillyard S. 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*. 207:203–205.
Kutas M, van Petten CK, Kluender R. 2006. Psycholinguistics Electrified II (1994–2005). In: Traxler M, Gernsbacher MA, eds. *Handbook of Psycholinguistics*, 2nd ed. Cambridge, Massachusetts: Academic Press; 659–724.
Kylliäinen A, Jones EJH, Gomot M, et al. 2014. Practical guidelines for studying young children with autism spectrum disorder in psychophysiological experiments. *Rev J Autism Dev Disord*. 1:373–386.
Lau EF, Phillips C, Poeppel D. 2008. A cortical network for semantics: (de)constructing the N400. *Nat Rev Neurosci*. 9:920–933.
Ledoux K, Coderre EL, Bosley L, et al. 2016. The concurrent use of three implicit measures (eye movements, pupillometry, and event-related potentials) to assess receptive vocabulary knowledge in normal adults. *Behav Res Methods*. 48:285–305.
Lord C, Risi S, Lambrecht L, et al. 2000. The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord*. 30:205–223.
Lord C, Rutter M, DiLavore PC, et al. 2012. *Autism Diagnostic Observation Schedule, Second Edition (ADOS–2) Manual (Part 1): Modules 1–4*. Torrance, California: Western Psychological Services.
Lord C, Rutter M, Le Couteur A. 1994. Autism Diagnostic Interview–Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord*. 24:659–685.
Manly BFJ. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*, 2nd ed. London, United Kingdom: Chapman & Hall.
Martineau J, Hernandez N, Hiebel L, et al. 2011. Can pupil size and pupil responses during visual scanning contribute to the diagnosis of autism spectrum disorder in children? *J Psychiatr Res*. 45:1077–1082.
McCleery JP, Ceponiene R, Burner KM, et al. 2010. Neural correlates of verbal and nonverbal semantic integration in children with autism spectrum disorders. *J Child Psychol Psychiatry*. 51:277–286.
Mottron L, Mineau S, Martel G, et al. 2007. Lateral glances toward moving stimuli among young children with autism: early regulation of locally oriented perception? *Dev Psychopathol*. 19:23–36.
Odekar A, Hallowell B, Kruse H, et al. 2009. Validity of eye movement methods and indices for capturing semantic (associative) priming effects. *J Speech Lang Hear Res*. 52:31–48.
Ozonoff S, Miller JN. 1995. Teaching theory of mind: a new approach to social skills training for individuals with autism. *J Autism Dev Disord*. 25:415–433.
Pérez Velázquez JL, Galán RF. 2013. Information gain in the brain's resting state: a new perspective on autism. *Front Neuroinform*. 7:1–10.
Pijnacker J, Geurts B, van Lambalgen M, et al. 2010. Exceptions and anomalies: an ERP study on context sensitivity in autism. *Neuropsychologia*. 48:2940–2951.
Plesa Skwerer D, Jordan SE, Brukilacchio BH, et al. 2016. Comparing methods for assessing receptive language skills in minimally verbal children and adolescents with autism spectrum disorders. *Autism*. 20:591–604.
R Core Team. 2013. *R: A Language and Environment for Statistical Computing [Computer Software]*. Vienna, Autria: R Foundation for Statistical Computing.
Schmitt LM, Cook EH, Sweeney JA, et al. 2014. Saccadic eye movement abnormalities in autism spectrum disorder indicate dysfunctions in cerebellum and brainstem. *Mol Autism*. 5:1–13.
Sereno SC, Rayner K. 2003. Measuring word recognition in reading: eye movements and event-related potentials. *Trends Cogn Sci*. 7:489–493.
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.
Tager-Flusberg H, Skwerer DP, Joseph RM, et al. 2017. Conducting research with minimally verbal participants with autism spectrum disorder. *Autism*. 21:852–861.
Tanenhaus M, Magnuson JS, Dahan D, et al. 2000. Eye movements and lexical access in spoken-language comprehension: evaluating a linking hypothesis between fixations and linguistic processing. *J Psycholinguist Res*. 29:557–580.
Tanenhaus M, Spivey-Knowlton M, Eberhard K, et al. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*. 268:1632–1634.
Venker CE, Eernisse ER, Saffran JR, et al. 2013. Individual differences in the real-time comprehension of children with ASD. *Autism Res*. 6:417–432.

 