Vocabulary size and Native Speaker self-identification influence flexibility in linguistic prediction among adult bilinguals

Ryan Peters[a1], Theres Grüter[b2], Arielle Borovsky[a3]

[a] Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, West Lafayette, IN, USA

[b] Department of Second Language Studies, University of Hawai‘i at Mānoa, 1890 East-West Road, Honolulu, HI, USA

[1] peter495@purdue.edu, [2] theres@hawaii.edu, [3] aborovsky@purdue.edu

Corresponding Author:

Ryan Peters

Department of Speech, Language, and Hearing Sciences

715 Clinic Drive

West Lafayette, IN 47907

Email: peter495@purdue.edu

Abstract

When language users predict upcoming speech, they generate pluralistic expectations, weighted by likelihood (Kuperberg & Jaeger, 2016). Many variables influence the prediction of highly-likely sentential outcomes, but less is known regarding variables affecting the prediction of *less-likely* outcomes. Here we explore how English vocabulary size and self-identification as a Native Speaker (NS) of English modulate adult bi-/multilinguals' pre-activation of less-likely sentential outcomes in two visual-world experiments. Participants heard transitive sentences containing an agent, action and theme (*The pirate chases the ship*) while viewing four referents varying in expectancy by relation to the agent and action. In experiment 1 (N=70), spoken themes referred to highly-expected items (e.g., ship). Results indicate lower-skill (smaller vocabulary size) and less confident (not identifying as NS) bi-/multilinguals activate less-likely action-related referents more than their higher-skill/confidence peers. In experiment 2 (N=65), themes were one of two less-likely items (The pirate chases the *bone*/*cat*). Results approaching significance indicate an opposite but similar size effect: higher-skill/confidence listeners activate less-likely action-related (e.g., bone) referents slightly more than lower-skill/confidence listeners. Results across experiments suggest higher-skill/confidence participants more flexibly modulate their linguistic predictions per the demands of the task, with similar but not identical patterns emerging when bi-/multilinguals are grouped by self-ascribed NS-status versus vocabulary size.

Keywords: Sentence comprehension, Prediction, Bilingualism, Eye-tracking, Local coherence

Introduction

Fluent language comprehension occurs rapidly, requiring listeners to interpret language as it unfolds, in the moment. Ongoing incremental processes in language comprehension involve the probabilistic (pre)activation of an array of potential sentential outcomes that vary in likelihood. The lexical dynamics, or patterns of timing and degree of activation, of these various outcomes, however, can vary according to numerous contextual and individual differences that are not yet fully understood. Kuperberg and Jaeger (2016, p. 32) recently proposed that "the degree and level of predictive pre-activation might be a function of its expected utility" to a given processing goal. In other words, comprehenders are assumed to (unconsciously) weigh the costs and benefits of prediction in any given processing situation. This process involves considering "estimates of the relative reliability of their prior knowledge and the bottom-up input" (p. 32). Importantly, comprehenders' estimates of the reliability of their prior knowledge will be a function of the amount of knowledge they have about the domain over which predictions are made. In the case of sentence comprehension in English, the relevant knowledge is of English words and sentences, or more generally, English language skill and experience. Indeed, Borovsky and colleagues (2012) showed that within their respective age groups, (monolingual) English-speaking adults and children with larger vocabularies demonstrate faster anticipatory lexical activation than those with smaller vocabularies. This result aligns with the "pluralistic" view of prediction where multiple factors contribute simultaneously (Hintz, Meyer, & Huettig, 2017), and indicates that age *and* vocabulary skill may be responsible for some previously reported differences between children and adults in speed and efficiency in real-time sentence comprehension (e.g., Kail & Salthouse, 1994; Kidd & Bavin, 2007; Snedeker & Trueswell, 2004).

However, age and vocabulary size are critically interrelated among monolingual speakers, thereby obscuring which variable(s) support language processing skills. We attempt to address this issue here by looking at adults of relatively similar age and educational background, but who all speak one or more other language(s) in addition to English. Bi- and multilingual adults differ widely regarding their knowledge of and ability in a specific language (e.g., English), depending on the amount of experience with that language they have had across their lifespan. As a result, there will be substantial differences among such speakers in English vocabulary size, unrelated to their chronological age. These differences allow us to ask more directly how vocabulary size in a specific language contributes to predictive processing in that language.

In addition to further examining the role of vocabulary size, an objectively measurable aspect of language knowledge, in prediction during language comprehension, extending the investigation to adult bi-/multilinguals also offers a unique opportunity for exploring the role of comprehenders' *confidence* in their language knowledge. If, as proposed by Kuperberg and Jaeger (2016), engagement in prediction is determined at least in part by comprehenders' *estimates* of their own knowledge, we expect that their confidence in their own ability in a language may be just as important as objective measurements of that ability. The operationalization of 'confidence', however, is far from straightforward. For the exploratory purposes of this study, we attempt to capture one aspect of bilingual speakers' confidence in their knowledge of English by whether or not they consider themselves a *Native Speaker* of English, drawing on a rich literature in the fields of applied linguistics and language pedagogy that has discussed and debated the concept of the 'Native Speaker' (Cook, 1999; Davies, 1991, 2003, 2013; Rampton, 1990). In experimental psycholinguistics, the notion of the native (vs non-native) speaker is rarely put into question. Yet for many bi- and multilingual language users, determining their native language(s) is tied not only to objective proficiency in that language, but to several sociolinguistic variables such as self-confidence and cultural identity. Davies (2003) argues that "the distinction native speaker – non-native speaker, like all majority-minority power relations, is at bottom one of confidence and identity" (p. 213). Under this view, whether an individual self-identifies as a Native Speaker (NS) of a given language is inherently an issue of self-perception, which ties into numerous assumptions about the speaker's relationship with the language. For example, native speakers, in contrast to second language (L2) speakers, "assume that what is said to them... can be understood by them in principle" (p. 200). Native speakers also see themselves as repositories of knowledge regarding what ‘the language’ is, in essence laying claim to their native language, particularly in interactions with L2 learners. These assumptions that often accompany the claiming of native-speakerhood typically result from extensive language experience beginning in childhood, emerging into what may be one of the most long-lasting and deeply held self-perceptions regarding language skill. Here we assume that the claiming of native-speakerhood reflects at least one aspect of a speaker’s confidence in their ability to understand speech in that language effortlessly and comprehensively. We thus hypothesize that self-ascribed NS status will be an explanatory factor in the extent to which bi/multilingual adult speakers engage in prediction during language processing.

Native speakers and adult sequential L2 learners differ in speed of information integration (e.g., Kilborn, 1992; cf. Kaan, Ballantyne, & Wijnen, 2015) and degree of anticipatory lexical activation (e.g., Grüter, Lew-Williams, & Fernald, 2012; Kaan, 2014; Lew-Williams & Fernald, 2010; Mitsugi & MacWhinney, 2016). However, the effect of proficiency on information processing speed is mixed, with some studies showing an effect of L2 proficiency (Chambers & Cooke, 2009; Dussias, Kroff, Tamargo, & Gerfen, 2013; Hopp, 2013; Leal, Slabakova, & Farmer, 2016) and others not (Dijkgraaf, Hartsuiker, & Duyck, 2016; Hopp, 2015). Less is known about speed and degree of anticipatory processing among adult bi- and multilinguals defined more broadly, namely speakers who have used one or more language(s) other than English for a substantial portion of their lives, either before exposure to English, or concurrently with English, starting in early childhood. Here we further investigate this issue by asking how vocabulary size, an objectively measured aspect of language skill, as well as self-ascribed NS status, which we use here as a measure of speakers’ confidence in their language skills, influence the timing and degree of lexical activation during spoken sentence comprehension in a more diverse population of speakers who do not fit squarely into traditional participant categories in language learning research, and are thus likely to have been underrepresented in such research to date.

We explore these questions using the Visual World Paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), in which gaze to a scene or set of referents, made while listening to spoken language, is taken as an index of online processing. Following Borovsky et al. (2012), we use an experimental design where participants listened to simple sentences of the form “The pirate chases the ship” while viewing a set of four images: the target object (SHIP), an object semantically related to the agent (TREASURE, agent-related distractor), a theme compatible with the verb (CAT, action-related distractor), and an unrelated distractor (BONE). The task is to select the image that “goes with the sentence.” On the assumption that visual attention reflects referential processing, we take proportion of looks to each image as an index of the amount of lexical activation of its referent. We take an increase in looks prior to the acoustic onset of the sentence-final referent as an indication of anticipatory lexical activation.

While most studies on predictive processing have focused on pre-activation of highly likely outcomes, the design of the study by Borovsky et al. (2012) and the extension in Troyer and Borovsky (2017), which we adopt here, was motivated by the question of whether language skill and experience would alter the dynamics of lexical activation for referents that are *less likely* given the cumulative evidence from the unfolding sentence at a given point, but which are “locally coherent” with the most recently encountered word (Kukona et al., 2011; Tabor, Galantucci, & Richardson, 2004). In this experiment, locally coherent lexical activation takes the form of increased looks to the action-related distractor after the onset of the verb (e.g., looks to CAT after hearing “The pirate *chases*”). Note that ‘cat’ is an unlikely continuation if the entire preceding sentence fragment (‘The pirate chases’) is considered, but if prediction is based only on the immediately preceding word (‘chases’), ‘cat’ becomes a highly likely continuation. Thus, while looks to the target object (SHIP) after the onset of the verb are considered cumulative or globally coherent anticipatory fixations, looks to the action-related distractor (CAT) are considered *locally* coherent anticipatory fixations. Looks to such locally coherent items are seen in native language processing regardless of the fact that they should already have been disqualified as likely targets by the agent (Borovsky et al., 2012; Kukona et al., 2011). While this pattern may seem less than optimal, locally coherent processing may play the important role of facilitating comprehension in the face of uncertainty and unexpected outcomes. For example, at the word level, the TRACE model accurately predicted/recognized words precisely because it was able to overcome initial predictions in the face of subsequent bottom-up stimuli (McClelland & Elman, 1986). Interestingly, this type of activation at a semantic/sentential level was found to be absent in children with specific language impairment (SLI; Borovsky, Burns, Elman, & Evans, 2013). Together, these findings suggest that the timing and degree of patterns of lexical activation during sentence processing may vary according to individual differences in language skill and experience. Here we explore whether the degree of activation of locally coherent referents is similarly related to differences in vocabulary size and self-ascribed NS status among adult speakers whose skill and experience with English varies substantially due to varying amounts of exposure to one or more other languages across their lifespan.

To this end, we conducted two visual-world experiments with adult bi- and multilingual speakers of English. In both experiments, we divided participants into higher and lower 'skill' subgroups by two different criteria: (1) according to a standardized test of vocabulary size, the Peabody Picture Vocabulary Test-Version 4 (PPVT; Dunn & Dunn, 2007), following the median-split procedure that Borovsky et al. (2012) used with monolingual children and adults, and (2) based on a simple dichotomous split according to participants' answer to the question "Do you consider yourself a native speaker of English?" (yes/no)

Experiment 1

Experiment 1 is designed to test three interrelated questions. First, do the higher-skill (PPVT-Higher, NS-Yes) groups show patterns of fixations indexing anticipatory lexical activation occurring sooner and to a greater degree than the lower-skill (PPVT-Lower, NS-No) groups? Such an outcome would be consistent with previous work on individual differences in lexical and sentential processing (e.g. Borovsky et al., 2012; Fernald, Perfors & Marchman, 2006; Hintz et al., 2017; Mani & Huettig, 2012). However, there have been inconsistent findings in the L2 processing literature as to whether native and non-native speakers differ in speed of language processing (e.g., Kilborn, 1992; Kaan et al., 2015) and whether L2 proficiency affects the degree of anticipatory processing in an L2 (e.g., Dijkgraaf, et al., 2016;

Dussias et al., 2013; Leal et al., 2016). Meanwhile, little is known about anticipatory processing in bilinguals that do not fit the traditional criteria of sequential L2 learners. Second, do higher- and lower-skill groups show different patterns of lexical activation for less-likely, locally coherent options across the sentence? This outcome would be consistent with studies that have reported individual differences in activation of locally-coherent (less-likely) outcomes that vary with language skill and other domain-general cognitive abilities, such as cognitive control (Nozari, Trueswell & Thompson-Schill, 2016; Woodard, Pozzan & Trueswell, 2016). Finally, we seek to explore to what extent an objective measure of vocabulary size on the one hand, and subjective self-categorization as a NS on the other result in groupings that show different patterns of results for the timing and degree of both anticipatory and locally coherent lexical activation. Moreover, we ask whether self-ascribed NS status as a proxy of speakers' *confidence* in their English language knowledge, can capture variance in predictive processing beyond what is accounted for by differences in objectively measurable language knowledge, operationalized here as vocabulary size assessed by the PPVT. Such an outcome would lend support to Kuperberg and Jaeger's (2016) proposal of prediction as a utility function, wherein speakers' estimates of their own knowledge plays a critical role in the degree to which they might engage in prediction.

Method

Participants

Seventy college students (mean age: 21.6 years, 52 women) participated in this study in return for course credit. All participants indicated exposure to a language other than English either before exposure to English, or concurrently with English, in early childhood. In other words, these were all speakers who would NOT typically be included in studies on native-language processing, since they are or were bilingual at some stage in their lives. As such, they constitute a highly heterogeneous sample with various profiles of dominance in English versus their other language(s) and include both simultaneous and sequential L2 learners. This group heterogeneity is advantageous for addressing the main questions of this study as there is substantial variation among these participants regarding both English vocabulary size and self-identification as a native speaker of English. The reasons for this variability are manifold, including length of exposure, type of exposure (e.g., immersion vs. classroom), and age of onset. As the *source* of these speakers' variability in skill and experience with English is of less interest in this study than the consequences of speakers' current skills and confidence on real-time processing, we do not further differentiate the groups by these factors.

As participants in this group vary in whether English was strictly speaking the first or second language they were exposed to chronologically, we henceforth refer to the participants' other language(s) as "LX" for the purpose of this study. There were 20 different LXs reported.[1] The three most commonly reported languages were evenly distributed between those who answered "yes" and "no" to the question "Do you consider yourself a native speaker of English?" As with the specific profiles of dominance in English, the specific English-LX relations are of less interest in this study, and thus we do not differentiate the groups by these factors.

As a result of the variability in participants' language experience and use across their life spans, we expect considerable variability within this sample not only in terms of empirically quantifiable English language skills, such as vocabulary size, but also in terms of the confidence with which this knowledge is put to use in real-time language processing, which here we attempt to capture by self-identification as a NS. The central goal of this study is to explore the effects of this variability on these overall highly proficient speakers' processing of simple English sentences.

Participants reported normal hearing, normal or corrected-to-normal vision, and no history of diagnosis of mental illness or treatment for speech, language or cognitive issues. One participant was

[1] Spanish (n=17); Korean (n=14); Chinese (n=13); Vietnamese (n=4); Armenian, Farsi, Indonesian, Japanese (each n=2); Arabic, Filipino, French, Gujarathi, Khmer, Lithuanian, Polish, Q'anjob'al, Sinhi, Slovak, Telugu, Thai, Urdu (each n=1).

excluded for receiving prior speech therapy and one participant was removed for failing to complete both the language background questionnaire and the PPVT.

Stimuli

The stimuli used in this experiment were the same as in Borovsky et al. (2012), for which eight sentence quartets (32 total sentences) were developed by mixing two agents, two actions, and four themes appropriate for each agent-action combination. All sentences consisted of the standard structure: article, $\text{noun}_{(\text{agent})}$, $\text{verb}_{(\text{action})}$, article, $\text{noun}_{(\text{theme})}$. An example quartet is:

1. The pirate hides the treasure.

2. The pirate chases the ship.

3. The dog hides the bone.

4. The dog chases the cat.

Each quartet had an associated image that consisted of photo-realistic pictures of the four potential themes, each presented on a 400 × 400-pixel white square background in its own quadrant of a black screen (Figure 1). Across the quartet of sentences each of the potential theme images corresponded variously to each of four conditions: target, agent-related distractor, action-related distractor, and unrelated distractor. Thus, each word and image served as its own control across lists, balancing for differences in intrinsic saliency. Additionally, across lists each theme picture appeared with equal frequency in each quadrant, and in a given version the target image appeared with equal frequency in each quadrant.

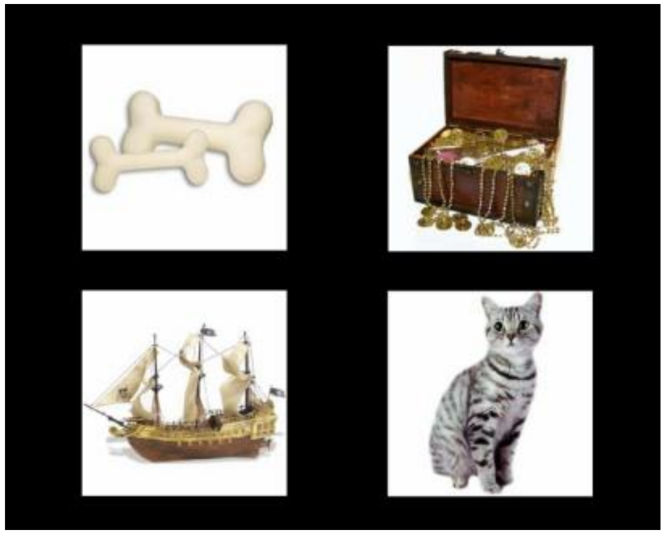

Figure 1. Screenshot of the visual stimulus associated with above example quartet.

The sentences were presented as auditory stimuli that were recorded by a female native English speaker ([redacted for review]) in a child-directed voice, sampled at 44,100 Hz on a single channel. Word durations were normalized to the following values: article-1, 134 ms; $\text{noun}_{(\text{agent})}$, 768 ms; $\text{verb}_{(\text{action})}$, 626 ms; article-2, 141 ms; $\text{noun}_{(\text{theme})}$/target, 630 ms. For a given list, each participant saw each of the eight images twice, each with a different associated sentence, so that any one participant heard 16 of 32 possible sentences.

Procedure

Experimental Task. The stimuli were presented on a 17-inch LCD display using a PC computer running EyeLink Experiment Builder software (SR Research, Mississauga, Ontario, Canada). Participants were told they would see sets of pictures while listening to sentences, and that they should click on the picture that "goes with the sentence." Before the experiment, the eye-tracker was focused and calibrated using a manual 5-point calibration and validation with a standard black-and-white 20-point bull's-eye image. Before each trial, participants were presented the same bull's eye in the center of the screen, with the trial starting once they had fixated on it. The images were presented for 2000 ms before sentence onset, and remained on the screen after sentence offset until participants clicked on an image with the mouse.

Eye Movement Recording. Eye movements were sampled at 500 Hz using an EyeLink 2000 remote eye-tracker attached directly below the LCD display. A remote arm configuration allowed for flexible adjustment of the camera and display to allow for reliable positioning within 580-620 mm from the participant's (typically right) eye. Head and eye-movement were automatically tracked by the system via a sticker affixed to each participant's forehead.

For each trial, eye-movements were recorded from image onset until participants clicked on a picture with the mouse. The eye-tracking system automatically classified recorded eye-movements into saccades, fixations and blinks using default settings. Fixations were then binned into 50 ms intervals for subsequent analyses.

Offline Measurements. Prior to the eye-tracking task participants completed a detailed language history questionnaire. After the eye-tracking task they were administered an offline measure of vocabulary skill, the Peabody Picture Vocabulary Test-Version 4 (PPVT; Dunn & Dunn, 2007), which has been normed for ages 2.5 to 90 years and used with adult bilingual populations in previous research (e.g., Bialystok & Luk, 2012).

Results

Assignment to Groups

We began with a correlational analysis of PPVT scores and five items from the language history questionnaire: (1) current age (Age), (2) age when first exposed to English (English Age of Acquisition, AoA), (3) length of time living in an English-speaking country (Length of Exposure to English; LoE), (4) self-rating of overall proficiency in English on a scale of 0 to 10 (following the procedure of the LEAP-Q; Marian et al., 2007; English skill), and (5) self-rating of overall proficiency in the participant's most proficient language other than English (LX skill). As shown in Table 1, objectively measured scores (PPVT) and self-report measures of English language skill were moderately but significantly correlated (*r*

= .47, *p* < .001). Both measures were correlated negatively with age of first exposure to English, and positively with length of exposure.

Table 1

Correlations Between Language History Questionnaire Items and PPVT scores for Experiment 1

Measure	1	2	3	4	5	6
1. Age (yrs)	—					
2. English AoA (yrs)	.35 **	—				
3. English LoE (yrs)	-.15	-.75 ***	—			
4. English skill *self-rating*	-.27 *	-.50 ***	.39 **	—		
5. LX skill *self-rating*	.11	.21	-.36 **	.07	—	
6. PPVT *age-normed*	-.30 *	-.47 ***	.26 *	.47 ***	-.04	—

Note. * *p* < .05. ** *p* < .01. *** *p* <.001

To explore our results using both objectively measured vocabulary size and self-identification as a NS as grouping criteria, participants were divided into subgroups according to (1) a median-split by PPVT score, and (2) by self-ascribed NS status. The make-up of the groups is displayed in Table 2. There was substantial but not complete overlap between the NS-No and PPVT-Lower groups, and between the NS-Yes and PPVT-Higher groups.

Table 2

NS Status and PPVT Group Make-up for Experiment 1

	NS-No	NS-Yes	Total	Comparison
PPVT-Lower	25	9	34	$\chi^2(1, 68) = 11.54$, *p* < **.001**
PPVT-Higher	10	24	34	φ = .44, odds ratio = .15
Total	35	33		

To determine the validity of using these two grouping criteria on the current sample we compared participants between groups for each grouping criterion on the same six items included in the correlational analysis (Table 3). For the NS-status groups, with the exception of age, there were significant group differences for all measures. However, for the PPVT groups, there were significant group differences only for PPVT, self-rated English skill and English AoA.

Table 3

Means of Questionnaire answers and PPVT scores by NS Status and PPVT Group for Experiment 1

Groupings	Native-Speaker-Status Groups			PPVT Median Split Groups		
	No n = 35	Yes n = 33	Comparison	Lower n = 34	Higher n = 34	Comparison
Age (yrs)	21.74 (2.48)	21.24 (1.52)	$t(56.95) = 1.01$ $p = .32$, $d = .27$	21.74 (2.4)	21.26 (1.68)	$t(58.94) = 0.94$ $p = .35$, $d = .24$
English AoA (yrs)	7.43 (3.75)	2.47 (2.70)	$t(61.84) = 6.28$ $p <$ **.001**, $d = 1.6$	6.59 (4.53)	3.46 (2.95)	$t(56.75) = 3.38$ $p =$ **.001**, $d = .90$
English exp. (yrs)	13.32 (6.25)	19.31 (2.99)	$t(49.76) = -5.08$ $p <$ **.001**, $d = -1.44$	15.23 (6.77)	17.10 (4.54)	$t(55.69) = -1.32$ $p = .19$, $d = -.35$
English skill *self-rating*	8.51 (1.15)	9.39 (0.75)	$t(58.84) = -3.77$ $p <$ **.001**, $d = -.98$	8.50 (1.11)	9.38 (0.82)	$t(60.70) = -3.65$ $p <$ **.001**, $d = -.96$
LX skill *self-rating*	7.86 (2.16)	6.58 (1.95)	$t(65.90) = 2.57$ $p =$ **.01**, $d = .63$	7.24 (2.55)	7.24 (1.69)	$t(57.31) = 0$ $p = 1$, $d = .0$
PPVT *age-normed*	90.17 (8.47)	101.16 (11.7)	$t(56.09) = -4.37$ $p <$ **.001**, $d = -1.17$	86.47 (5.96)	104.88 (7.92)	$t(61.29) = -10.83$ $p <$ **.001**, $d = -2.77$

Notes. Standard deviations are reported in parentheses. Comparisons assume unequal variances.

Overall, the results in Table 3 support the use of both self-ascribed NS status and PPVT scores as grouping criteria. Both grouping methods result in non-identical groups that differ in English skill as determined by both PPVT and self-rated English skill. However, the data also indicate that using NS status as a grouping criterion results in more cleanly differentiated groups, with the magnitudes of the differences between groups being larger for all but PPVT scores when using NS status as a grouping criterion. This is not entirely surprising given that NS status is a categorical variable, which allows for a well-motivated group split, in contrast to the split based on the continuous variable of PPVT, which is arbitrarily determined by scores of the sample. At the same time, this observation is noteworthy, as self-ascribed NS status is a very simple and subjective criterion, and as such one may question its validity. The fact that the NS-status subgroups here are well-differentiated in terms of more generally accepted individual-difference variables in bilingualism research, such as age of acquisition, length of exposure, and self-ratings of both English and LX skills (Marian et al., 2007), suggests that self-determined NS status may be a useful criterion to include in future research concerned with individual differences among bilingual speakers.

Behavioral Analysis

We verified that participants attended to and understood the sentences and task by calculating the accuracy with which participants selected the correct target picture in the experimental task. Accuracy was high, with only 6 incorrect responses out of 1088 trials (99.45% correct). The six incorrect responses were spread across 6 individuals, each with an accuracy of 93.75 %. Only accurate trials were included in all subsequent analyses.

Eye-movement Analyses

Time Course by Group. To explore cumulative and locally coherent eye-movements during incremental sentence comprehension, we first visualized the time course of fixations by calculating the mean proportion of time spent fixating the four target areas in each image (the target, agent-related, action-related, and unrelated pictures). These means were then averaged across participants in each of the two NS-Status Groups and in each of the two PPVT Groups and plotted against time from sentence onset in Figure 2.

In these time course plots, there are two apparent visual patterns that have typically appeared in prior studies using similar sentential stimuli (Borovsky et al., 2012; Borovsky et al., 2013). First, there is a rise in fixations to the Target that begins as the agent is spoken and continues to the end of the trial. This rise is initially accompanied by an equal increase in fixations to the Agent-Related distractor. Second, there is a momentary increase in fixations to the Action-Related distractor. This increase begins near the end of the action time window and subsides in the theme time window, and appears more pronounced in both the NS-No and PPVT-Lower groups.

We carried out analyses that address two main questions, (1) does the timing of anticipatory fixations vary between lower skill (NS-No and PPVT-Lower) and higher skill (NS-Yes and PPVT-Higher) groups and (2) do patterns of fixations to the locally coherent referent vary between groups? For both sets of analyses, we compared a pair of relevant interest areas. For example, for (1) we compared the

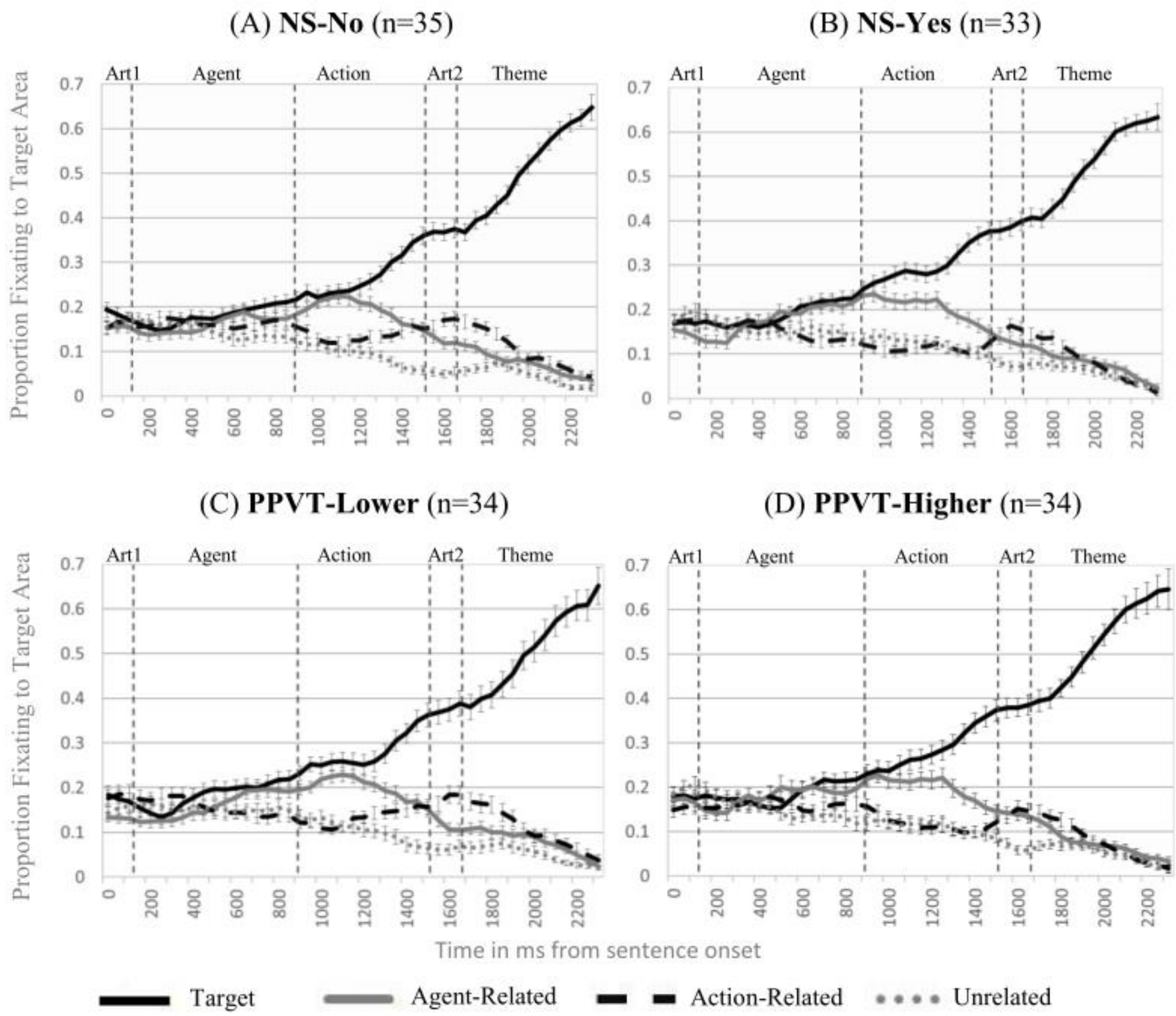

Figure 2. Time course of fixations to target and distractor interest areas for participants who did not claim native speaker status (A) and participants who did claim native speaker status (B), and for low (C) and high (D) PPVT median split groups, with mean fixation proportions calculated over 50-ms time bins (with SE bars).

Target to the Agent-Related distractor. However, one problem that arises with doing a direct comparison in this case is the violation of the assumption of linear dependence, given the fact that an increase in the proportion of looks to one interest area necessarily results in a decrease in the proportion of looks to another. To circumvent this problem, we used a dependent measure calculated by taking the log of the ratio of fixation proportions to relevant interest areas.[2] So, for (1) we took the log of the ratio of the

[2] Log ratios are undefined for 0, so every 0 in either the numerator or denominator was replaced with 0.01.

Target over the Agent-Related distractor. The resulting Log-Gaze ratio is a measure of relative bias that varies between positive and negative infinity. Using this ratio fixes the additional problem of the violation of the assumption of homogeneity of variance that results from the fact that simple proportion measures are bounded between 0 and 1. Positive and negative scores indicate a bias to look at the interest area in the numerator and denominator respectively, while a score of zero indicates equivalent looks between competing areas of interest. Accordingly, for (1), a positive score would indicate a bias to look at the Target, a negative score would indicate a bias towards the Agent-Related distractor, and a score of zero would indicate equivalent looks to both the Target and Agent-Related distractor. Thus, rather than comparing looks to two areas of interest directly, we can compare the resultant Log-Gaze ratio to 0 in order to determine periods of divergence (Arai, van Gompel & Scheepers, 2007; Borovsky et al., 2014; Knoeferle & Kreysa, 2012 for a similar approach).

Analysis of Anticipatory Fixations toward the Target by Group. For this analysis, we calculated the mean log-gaze ratio of looks to the Target vs Agent-Related Distractor over the anticipatory time window, going from action onset to theme onset. A comparison assuming unequal variances found no significant difference between the NS-No group (M = .17, SD = .22) and NS-Yes group (M = .16, SD = .19), t (65.17) = .04, p = .97, d = 0.01. Likewise, there was no significant difference between the PPVT-Low group (M = .15, SD = .16) and PPVT-High group (M = .18, SD = .25), t (56.66) = -.62, p =.54, d = -0.16. These results indicate that higher-skill participants *do not* show patterns of fixations indexing anticipatory lexical activation occurring sooner and to a greater degree than lower-skill participants.

Analysis of Locally Coherent Fixations by Group. Consistent with prior findings from work with native English speakers, we see locally coherent lexical activation for lower and higher skill participants (for both grouping methods) following the onset of the Action. This pattern is characterized by increased looks to the Action-Related target area relative to the Unrelated target area in Figure 2. This pattern can also be considered anticipatory processing; however, it differs from the other instances of anticipatory processing in that it is not cumulative, ignoring what came earlier in the sentence, and resulting solely from information encoded in the verb.

The increase in looks to the action-related target area is clearly visible in the timecourse plots for both lower and higher skill groups, but is noticeably larger for the lower than for the higher skill groups. For this analysis, we calculated the mean log-gaze ratio of looks to the Action-Related vs Unrelated target areas over two time windows (Figure 3) at the trial level. While past work (e.g. Borovsky et al., 2013) indicates this effect takes place over the entire verb-phrase window (going from action onset to theme offset), we include exploratory analyses of a subset time window, the anticipatory time window (going from action onset to theme onset), to enable comparison with experiment 2. We entered group status as a categorical predictor of trial level log-gaze ratios in a linear mixed effects model using R (R Core Team, 2016) and lme4 (Bates, Maechler, Bolker, & Walker, 2015), with random intercepts for subjects and items. For all four analyses, the lower skill group (NS-No, PPVT-Lower) was set as the baseline, and thus the value presented as the mean for the lower skill group is the intercept of the model and the mean of the higher skill group is calculated by adding the coefficient for the fixed effect of group to the intercept. The presented t-tests, calculated using the Satterthwaite approximations to degrees of freedom using lmerTest (Kuznetsova, Brockhoff, & Christensen, 2016), measure whether the coefficient of the fixed effect of group is significantly different from zero, and thus whether the two groups are significantly different. For the anticipatory time window, going from action onset to theme onset, the analysis indicated a significant difference with a moderate effect size between the NS-No (M = .18, SD = .35) and the NS-Yes group (M = .01, SD = .40), t (65.57) = -2.33, p = .03, d = -0.45. Over the same time window, there was no significant difference with a small effect size, between the PPVT-Lower (M = .15, SD = .35) and the PPVT-Higher group (M = .05, SD = .41), t (65.85) = -1.38, p = .17, d = -0.26.[3] For the verb-phrase time

[3] A linear mixed effects analysis of the relationship between log-gaze ratio of proportion of looks to the Action-Related vs Unrelated in the anticipatory time window and PPVT (grand mean centered) entered as a continuous

window, going from action onset until theme offset, there was again a significant difference with a moderate effect size between the NS-No (*M* = .21, *SD* = .35) and the NS-Yes group (*M* = .06, *SD* = .34), *t* (65.26) = -2.35, *p* = .02, *d* = -0.43. Over the same time window, there was also a significant difference with similar moderate effect size between the PPVT-Lower (*M* = .21, *SD* = .35) and the PPVT-Higher group (*M* = .07, *SD* = .34), *t* (65.60) = -2.17, *p* = .03, *d* = -0.40.[4]

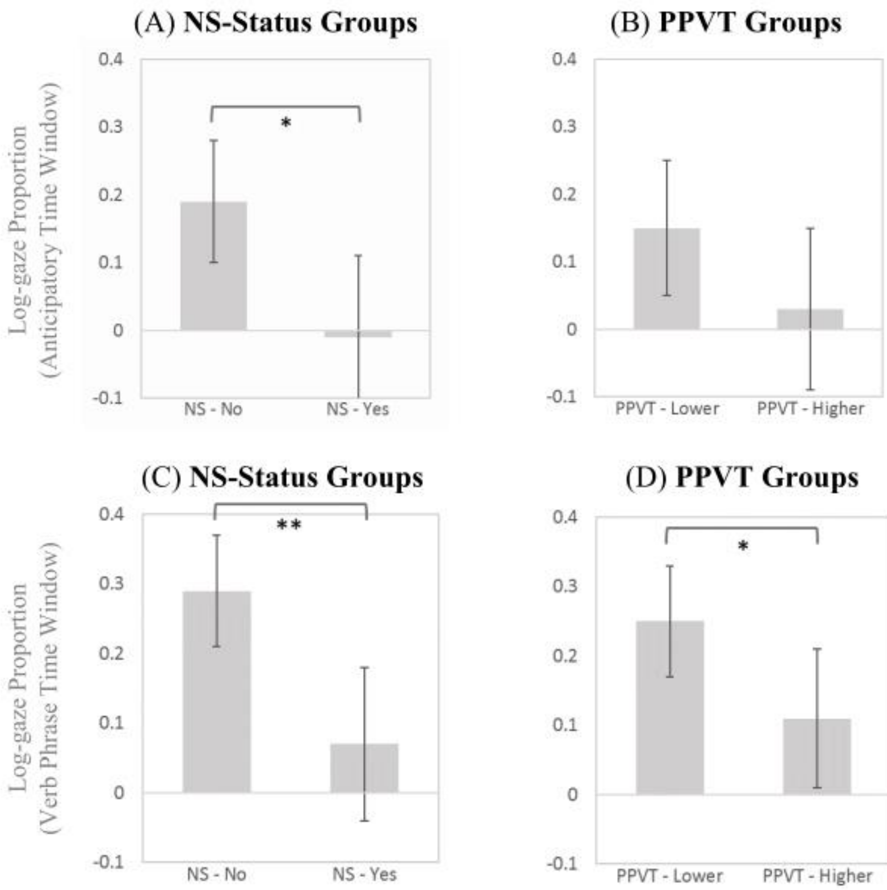

Figure 3. Between group comparisons of mean log-gaze in Anticipatory (A, B) and Verb Phrase (C, D) Time Windows for self-determined native speaker status groups (A, C) and PPVT median split groups (B, D). * *p* < .05. ** *p* < .01. Error bars represent 95% confidence intervals.

(rather than categorical) predictor, with random intercepts for subjects and items, found PPVT significantly affected mean log-gaze ($\chi^2(1)=9.08$, p=.003): $LogGaze_{trial} = 0.1 - 0.01 * PPVT_i + r_{subject} + r_{item} + \varepsilon_{trial}$

[4] A linear mixed effects analysis of the relationship between log-gaze ratio of proportion to looks to the Action-Related vs Unrelated in the verb phrase time window and PPVT (grand mean centered) entered as a continuous predictor, with random intercepts for subjects and items, found PPVT significantly affected mean log-gaze ($\chi^2(1)=11.71$, p<.001): $LogGaze_{trial} = 0.13 - .01 * PPVT_i + r_{subject} + r_{item} + \varepsilon_{trial}$

Analysis of interaction between PPVT and NS Group status. To further delve into the issue of objectively measured English vocabulary size versus self-identified English Native Speaker status we carried out one final set of analyses exploring the interaction between PPVT and NS group status in predicting the mean log-gaze ratio of proportion of looks to the Action-Related vs Unrelated items. We performed a linear mixed effects analysis of the relationship between trial level mean log-gaze ratio of proportion of looks to the Action-Related vs Unrelated items and the fixed effects of PPVT (grand mean centered), NS group status, and the interaction between PPVT and NS group status. For random effects, we included intercepts for subjects and items. We calculated $R^2_{\beta*}$, an estimate of the variance explained by fixed effects in the context of random effects, for each model via penalized quasi-likelihood estimation (Jaeger, Edwards, Das, & Sen, 2017) using r2glmm (Jaeger, B., 2017). For the anticipatory time window, a model including only PPVT as a fixed effect had $R^2_{\beta*}$= .13, CI = [.02, .29], while the maximal model including PPVT, NS status, and the interaction term, had $R^2_{\beta*}$= .20, CI = [.08, .39]. A likelihood ratio test comparing the models was marginally significant ($\chi^2(2) = 5.69$, $p = .06$), tentatively supporting the inference that the models are different. Likewise, for the verb phrase time window, the simple model with only PPVT had $R^2_{\beta*}$= .16, CI = [.04, .33], while the maximal model had $R^2_{\beta*}$= .23, CI = [.10, .42]. A likelihood ratio test comparing the models was also marginally significant ($\chi^2(2) = 5.97$, $p = .05$). These results tentatively suggest that Native Speaker status may account for unique variance, unaccounted for by PPVT score, in determining the likelihood of looking at the locally coherent item.

Experiment 1 Discussion

In experiment 1 we set out to test three interrelated questions about how vocabulary size and self-ascribed native speaker status influence the timing and degree of lexical activation during spoken sentence comprehension.

The first question asked: Do higher skill (i.e., PPVT-Higher and NS-Yes, respectively) participants show patterns of fixations indexing anticipatory lexical activation occurring sooner and to a greater degree than lower skill (i.e., PPVT-Lower and NS-No, respectively) participants? The results of the analysis of anticipatory fixations toward the target by group clearly demonstrate that the answer is: no, participants who identified as NS and those with larger vocabularies were indistinguishable from participants who did not identify as NS and those with smaller vocabularies.

The second question asked: Do higher- and lower-skill groups show different patterns of lexical activation for less-likely, locally coherent options across the sentence? The results of the analysis of locally coherent fixations offer evidence that the answer is: yes. There were consistent small to moderate effects across all statistical analyses, including group level analyses based on both NS status and PPVT and participant level analyses with PPVT treated as a continuous variable, demonstrating that lower skill participants show a greater bias to look at Action-related, locally coherent items relative to higher skill participants. In other words, participants with smaller vocabularies and those who do not consider themselves to be native speakers, who may experience more uncertainty in everyday language interpretation than more highly skilled participants, appear to adaptively activate less-likely locally coherent referents.

The third question asked: Do objective measures of vocabulary knowledge and self-identification as NS result in groupings that show slightly different patterns of results for the timing and degree of both anticipatory and locally coherent lexical activation? Two pieces of evidence provide tentative evidence that the answer is yes. First, the effect size in the analysis of locally coherent fixations was greater when using native speaker status as a grouping criterion in comparison to using PPVT. Second, the analysis of the interaction between PPVT and native speaker status indicates that a model predicting locally coherent fixations including such an interaction term is marginally better than a model without one. Taken together, there is tentative evidence that the effect is not solely determined by individual differences measurable by PPVT.

Given the relationship between language skill and locally coherent lexical activation seen in this experiment, one might ask if this pattern reflects different strategies in language comprehension. One possibility suggested by research and modeling at the word level is that locally coherent processing may facilitate recovery in the face of uncertain language input and unexpected linguistic outcomes (McClelland & Elman, 1986). In this case, it suggests that lower skill participants, who may have greater uncertainty in their everyday experience of spoken language comprehension, may therefore activate a broader range of linguistic continuations during sentence processing. (Pre)-activating a broader range of continuations would, in turn, lead lower-skill comprehenders to show less difficulty than higher skill participants in interpreting sentences that contain plausible, but unexpected outcomes. In this case, one might expect reduced recovery costs for lower skill participants than higher skill participants for the processing of sentences ending with an action-related item compared to those ending with an unrelated item. We explore this hypothesis in experiment 2.

Experiment 2

In experiment 1, participants with smaller vocabularies and those who did not self-identify as NS showed greater locally coherent lexical activation than higher skill (PPVT-Higher, NS-Yes) participants. This pattern of results led us to the somewhat counter-intuitive prediction that lower skill (PPVT-Lower, NS-No) participants will show reduced recovery costs compared to higher skill participants, demonstrated by faster and/or more robust fixations to the theme when sentences end with a less-expected, but locally-coherent action-related item. However, another possibility is that higher skill participants are more efficient than lower skill participants at flexibly responding to changes in the likelihood statistics of the language they encounter and modifying their patterns of predictive activation on the fly. In this case, higher skill participants may have exhibited less locally coherent activation to non-target items in experiment 1 precisely because the sentential outcomes always adhered to highly expected outcomes, and therefore, it was not beneficial to the task at hand to consider other less-plausible endings. Thus, in a situation where activation of alternative outcomes facilitates linguistic processing, we may see that higher skill participants are faster than lower skill participants to recognize unexpected, but locally-coherent outcomes. Experiment 2 is designed to test these opposing predictions.

Method

Participants

Sixty-five college students (mean age: 21 years, 49 women) participated in this study in return for course credit. Participants were drawn from the same population as those in experiment 1, namely all participants indicated exposure to a language other than English either before exposure to English, or concurrently with English, in early childhood. Thus, once again, participants constitute a heterogeneous sample, which includes both simultaneous and sequential bilinguals, and speakers with various profiles of dominance in English versus their other language(s) (LX). There were 13 different LXs reported.[5] The three most commonly reported languages were distributed in similar proportions between those who answered "yes" and "no" to the question "Do you consider yourself a native speaker of English?"

Participants reported normal hearing, normal or corrected-to-normal vision, and no history of diagnosis of mental illness or treatment for speech, language or cognitive issues. Five participants were removed due to issues with the eye-tracking task: failure to complete the task (n=2), eye-tracking error (n=1), and below chance levels of accuracy (n=2).

Stimuli

The stimuli used in the critical trials for experiment 2 consisted of a similar set of sentence quartets and visual scenes to those used in experiment 1, except that rather than ending in what would be the typical target, the sentences ended in either the action-related or unrelated targets. Examples of quartets for the Action-Related as target and Unrelated as target conditions are:

Action-Related Ending	Unrelated Ending
1. The pirate hides the *bone*.	1. The pirate hides the *cat*.

[5] Spanish (n=17); Chinese (n=15); Korean (n=15); Vietnamese (n=6); Farsi (n=3); Arabic, Hebrew, Kannada, Portuguese, Punjabi, Somali, Tagalog, Thai (each n=1).

2. The pirate chases the *cat*.	2. The pirate chases the *bone*.
3. The dog hides the *treasure*.	3. The dog hides the *ship*.
4. The dog chases the *ship*.	4. The dog chases the *treasure*.

As in experiment 1, each word and image served as its own control across lists, balancing for differences in intrinsic saliency, with each target picture appearing with equal frequency in each quadrant. Also, in a given version, the target image appeared with equal frequency in each quadrant. Word durations were normalized to the following values: article-1, 98, ms; $\text{noun}_{\text{(agent)}}$, 940 ms; $\text{verb}_{\text{(action)}}$, 967 ms; article-2, 165 ms; $\text{noun}_{\text{(theme)}}$/target, 884 ms. For a given version of the study, participants saw 16 critical trials, 8 ending with the action-related target and 8 ending with the unrelated target, as well as 32 filler sentences that ended with the typical target and were unrelated to the current study. The filler items were included to counteract the effect of including sentences with anomalous endings.

Procedure

The procedure for the experimental task and eye movement recording were identical to experiment 1.

Offline Measurements. As with experiment 1, prior to the eye-tracking task participants completed a language history questionnaire and afterwards they were administered the PPVT.

Results

Assignment to Groups

As in experiment 1, we chose to use both self-determined NS status and PPVT as grouping factors, running parallel group level analyses. One participant is missing from the NS groups due to not completing the language history questionnaire. Five participants are missing from the PPVT groups due to not completing the PPVT (n=3) or being one of the participants with a median score on the PPVT (n=2). We once again began with an exploratory correlational analysis of the same six items as in experiment 1 to verify whether or not there is a similar pattern of relationships. As shown in Table 4, objectively measured scores (PPVT) and self-report measures of English language skill once again correlated moderately but significantly ($r = .52$, $p < .001$). Both measures are correlated negatively with age of first exposure to English, and positively with length of exposure.

Table 4

Correlations Between Language History Questionnaire Items and PPVT scores for Experiment 2

Measure	1	2	3	4	5	6
1. Age (yrs)	—					
2. English AoA (yrs)	.19 *	—				
3. English exp. (yrs)	.40 ***	-.52 ***	—			
4. English skill *self-rating*	.17 *	-.37 ***	.31 ***	—		
5. LX skill *self-rating*	.21 *	.36 ***	-.16 *	.00	—	
6. PPVT *age-normed*	-.08	-.39 ***	.32 ***	.52 ***	-.20*	—

Note. * $p < .05$. ** $p < .01$. *** $p <.001$

Given the similarity between the structure of relationships in Tables 1 and 4, we moved forward with the usage of self-determined NS status and PPVT as grouping variables. The make-up of the groups is displayed in Table 5. As in experiment 1, there was substantial but not complete overlap between the NS-No and PPVT-Lower groups. Unlike experiment 1, the PPVT-Higher group was made up by nearly equal numbers of participants from the NS-Yes and NS-No groups.

Table 5

NS Status and PPVT Group Make-up for Experiment 2

	NS-No	NS-Yes	Total	Comparison
PPVT-Lower	25	2	*27*	$\chi^2(1, 54) = 10.75$, $p =$ **.001**
PPVT-Higher	13	14	27	φ = .49, odds ratio = .07
Total	38	16		

Notes. Only participants who completed both grouping measures are included in this table.

We next compared lower skill and higher skill (within grouping criterion) on the same six items included in the correlational analysis (Table 6). With the exceptions of age for the NS groups, and age and self-rated LX skill for the PPVT groups, there were significant group differences for all other measures.

Table 6

Means of Questionnaire answers and PPVT scores by NS Status and PPVT Group for Experiment 2

Groupings	Native-Speaker-Status Groups			PPVT Median Split Groups		
	No n = 42	Yes n = 17	Comparison	Lower n = 28	Higher n = 27	Comparison
Age (yrs)	20.86 (1.59)	21.29 (3.67)	$t(18.47) = -.47$ $p = .64$, $d = -.22$	21.21 (2.04)	20.81 (2.91)	$t(46.49) = .59$ $p = .56$, $d = .17$
English AoA (yrs)	6.83 (4.21)	1.35 (2.29)	$t(51.88) = 6.41$ $p <$ **.001**, $d = 1.78$	7.63 (4.81)	3.19 (3.23)	$t(45.51) = 3.98$ $p <$ **.001**, $d = 1.18$
English exp. (yrs)	12.63 (6.43)	20.11 (5.33)	$t(35.91) = -4.57$ $p <$ **.001**, $d = -1.53$	12.08 (6.26)	17.28 (6.61)	$t(50.99) = -2.94$ $p =$ **.005**, $d = -.82$
English skill *self-rating*	8.38 (1.21)	9.18 (0.88)	$t(40.41) = -2.80$ $p =$ **.008**, $d = -.88$	8.00 (1.14)	9.15 (0.95)	$t(50.29) = -4.02$, $p <$ **.001**, $d = -1.13$
LX skill *self-rating*	7.57 (2.58)	4.53 (3.26)	$t(24.50) = 3.43$ $p =$ **.002**, $d = 1.39$	7.19 (3.26)	5.89 (2.99)	$t(51.62) = 1.52$ $p = .13$, $d = 0.42$
PPVT *age-normed*	90.69 (10.4)	103.12 (9.39)	$t(33.76) = -4.40$ $p <$ **.001**, $d = -1.51$	84.86 (6.6)	103.96 (7.1)	$t(52.38) = -10.33$ $p <$ **.001**, $d = -2.85$

Notes. Standard deviations are reported in parentheses. Comparisons assume unequal variances.

As in experiment 1, the data once again indicate that using NS status as a grouping criterion results in more cleanly differentiated groups compared to PPVT. However, the difference in NS group sizes highlights an important drawback of using this variable as a criterion for creating subgroups, namely that it can be difficult to determine a priori what participants will answer, which can result in unequal sample sizes across subgroups. In experiment 1 we ended up with nearly equal subgroups, but in experiment 2 we have over twice as many participants in the NS-No compared to the NS-Yes group. In contrast, deciding groups by performing a median split on PPVT ensures equal group sizes irrespective of the peculiarities of the sample, somewhat compensating for the weakness of being arbitrarily determined by the given data.

Behavioral Accuracy

The accuracy with which participants selected the correct target picture in the experimental task was checked to make certain that they understood the sentences and the task. Accuracy was high, though not as high as in experiment 1, with only 31 incorrect responses out of 1008 trials (96.92% correct). However, as previously mentioned, two participants made 11 mistakes each (31.25% correct) and were removed from all analyses. The remaining incorrect responses were spread across 6 participants, with one participant making 4 mistakes and the rest making 1. Only accurate trials were included in all subsequent analyses.

Eye-movement Analyses

We carried out analyses that address two main questions: (1) as in the second question from Experiment 1, do patterns of fixations to the locally coherent referent vary between groups, and (2) do lower and higher skill participants differ in their pattern of fixation proportions to the theme *when sentences end with the locally-coherent action-related item*? These two questions are related since the locally coherent referent *is the theme* in the Action-Related as target condition. However, in the Analysis of Locally Coherent Fixations that addresses question 1 we look at fixations occurring *before the theme*, *across conditions*. Looking at this time-window allows us to replicate a portion of the analysis in experiment 1. In contrast, in the Analysis of Fixations to Action-Related as Target vs Unrelated as Target, which addresses question 2, we compare patterns of fixations occurring during a time window *including the theme*, *between conditions*. Thus, while the first set of analyses focus on individual differences in initial fixations to locally-coherent targets, the second set of analyses focus on the possible effects of such differences in locally-coherent processing on subsequent processing when the locally-coherent target actually becomes the theme.

Analysis of Locally Coherent Fixations. We restricted these analyses to the anticipatory time window (going from verb onset to theme onset) so that we could collapse across the Action-Related-as-Target and Unrelated-as-Target conditions, which are equivalent in auditory and visual presentation up to this point. We calculated the mean log-gaze ratio of looks to the Action-Related vs Unrelated target areas (Figure 4) at the trial level. As in experiment 1, we then entered group status as a categorical predictor of trial level log-gaze ratios in a linear mixed effects model, with random intercepts for subjects and items. For both analyses, the lower skill group (NS-No, PPVT-Lower) was set as the baseline, and thus the value presented as the mean for the lower skill group is the intercept of the model and the mean of the higher skill group is calculated by adding the coefficient for the fixed effect of group to the intercept. The presented t-tests, calculated using the Satterthwaite approximations to degrees of freedom using lmerTest (Kuznetsova, Brockhoff, & Christensen, 2016), measure whether the coefficient of the fixed effect of group is significantly different from zero, and thus whether the two groups are significantly different. This analysis indicated a marginally significant difference with a moderate effect size between the NS-No (*M* = .14, *SD* = .32) and the NS-Yes group (*M* = .28, *SD* = .32), *t* (59.09) = 1.75, *p* = .09, *d* = 0.43. Over the same time window, there was also a marginally significant difference with a small effect size between the PPVT-Lower (*M* = .11, *SD* = .32) and PPVT-Higher group (*M* = .25, *SD* = .42), *t* (53.45) = 1.88, *p* = .07, *d* = 0.38, indicating that the higher skill groups were tentatively *more* likely than the lower skill groups to look at the Action-Related vs the Unrelated target.[6]

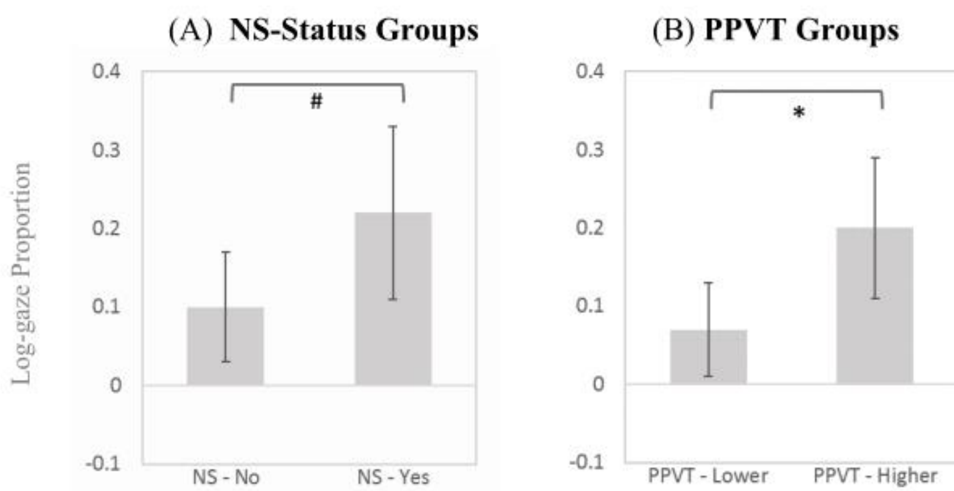

Figure 4. Between group comparisons of mean log-gaze proportion of Action-Related vs. Unrelated item in Anticipatory Time Window for (A) self-determined native speaker status groups and (B) PPVT median split groups. # *p* = .09. * *p* < .05. 95% Confidence Interval Error bars.

Analysis of Fixations to Action-related as Target vs Unrelated as Target. To explore the hypothesis that lower skill participants, relative to higher skill participants, would show more rapid and/or

[6] A linear mixed effects analysis of the relationship between log-gaze ratio of looks to the Action-Related vs Unrelated in the anticipatory time window and PPVT (grand mean centered), with random intercepts for subjects and items, found PPVT did not significantly affect mean log-gaze ($\chi^2(1)=1.83$, p=.18):

$$LogGaze_{trial} = 0.18 + 0.004 * PPVT_i + r_{subject} + r_{item} + \varepsilon_{trial}$$

more robust fixations to the theme when sentences end with the locally-coherent action-related items compared to when sentences end with an unrelated item, we once again compared the groups over a broad time window. Whereas in the previous analysis we collapsed across the Action-Related as Target and Unrelated as Target conditions and compared looks to the locally coherent item, in this second analysis we are comparing looks to the target between conditions. The time course of fixations to sentence targets in the action-related and unrelated conditions was visualized by first calculating the mean proportion of time spent fixating the targets. Means were then averaged across participants in each of the groups and plotted against time from sentence onset in Figure 5.

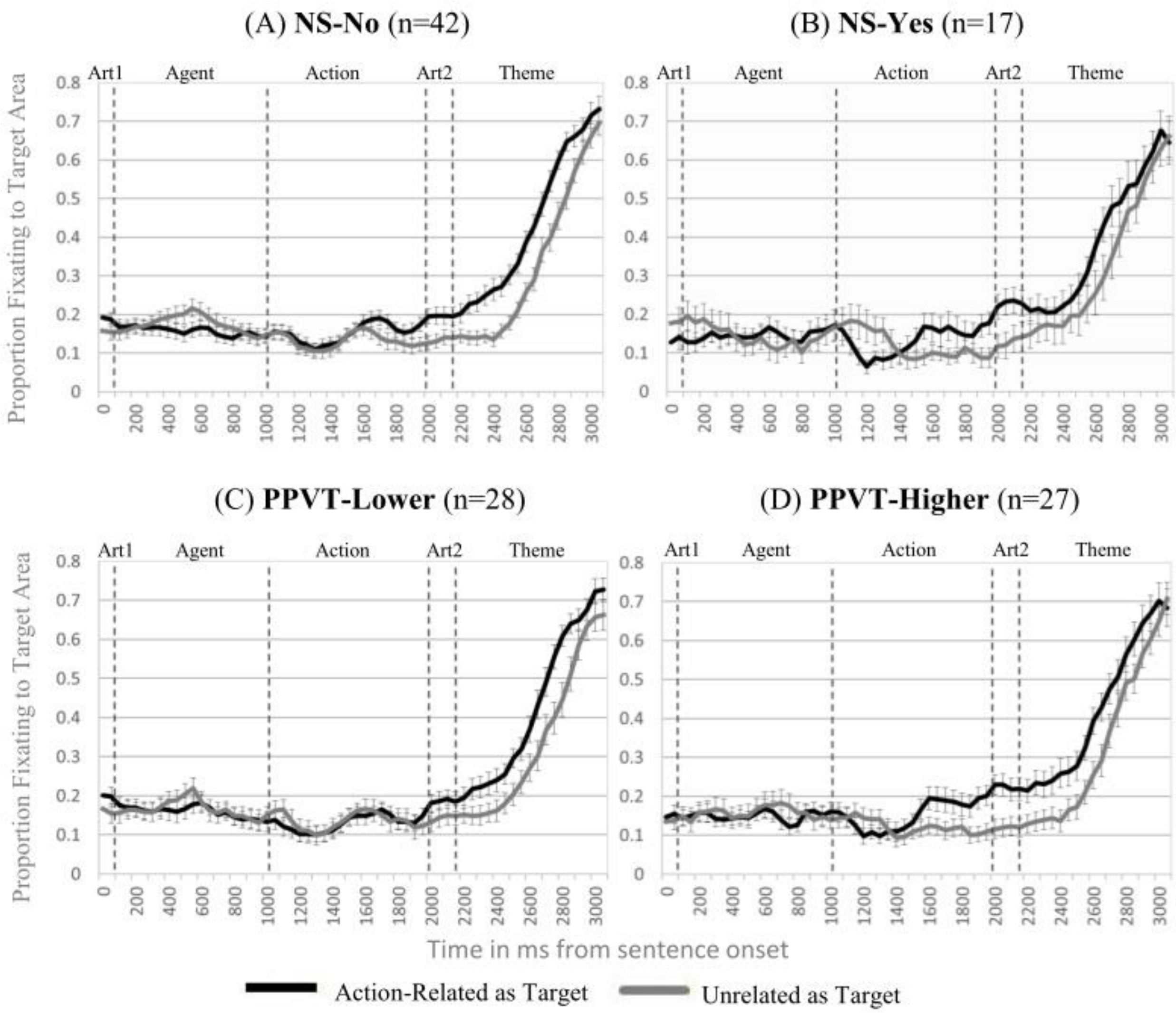

Figure 5. Fixation Proportions to Target in Action-related and Unrelated Conditions by NS Group (A, B) and PPVT Group (C, D), calculated over 50-ms bins (with SE bars).

For the analysis, we calculated the mean difference in fixation proportions between the target items in the Action-Related as Target and Unrelated as Target conditions over the verb-phrase time window (going from Action onset until Theme offset) for each participant. There was no significant difference between the NS-No (M = .06, SD = .10) and NS-Yes group (M = .06, SD = .15), t (22.25) = .08, p = .94, d = 0.03. Likewise, there was no significant difference between the PPVT-Lower (M = .04, SD = .11) and PPVT-Higher group (M = .08, SD = .13), t (50.72) = -.96, p =.34, d = -0.27.[7]

Experiment 2 Discussion

In experiment 2 we set out to address the question: How do vocabulary size and self-ascribed NS status influence the timing and degree of lexical activation during comprehension of spoken sentences with *unexpected endings*?

Following our findings from Experiment 1 where lower skill (PPVT-Lower, NS-No) participants showed relatively greater activation of unexpected, but locally-coherent sentence outcomes, we developed two contrasting hypotheses. Our primary hypothesis was that lower skill participants would show reduced recovery costs compared to higher skill (PPVT-Higher, NS-Yes) participants for the processing of sentences ending with an Action-Related item compared to those ending with an Unrelated item. Our second hypothesis was that higher skill participants would show relatively greater flexibility in responding to the presence of uncertain sentential outcomes in the task, resulting in overall faster recovery for unexpected outcomes than lower skill participants. While the analysis of fixations to the Action-Related as Target vs Unrelated as Target showed that all participants were faster to recognize unexpected endings that were locally-coherent with the sentential action, in the analysis of locally-coherent fixations we observed differences as a function of English skill. Namely, higher skill participants were marginally

[7] Visual inspection of the fixation timecourse indicates the possibility that these null results are due to group differences occurring over shorter time spans. However, analyses over shorter time windows (verb+art and theme only) also did not reveal significant group effects.

more likely to consider the locally coherent option than lower skill participants. Our results therefore did not support the first hypothesis, and were more consistent with the second hypothesis, which, we argue below, is consistent with Kuperberg and Jaeger's (2016) expected utility function of prediction.

General Discussion

Adults comprehend spoken sentences rapidly by (pre)activating a host of potential outcomes, with the dynamics of this process varying according to numerous contextual and individual differences. Kuperberg and Jaeger recently proposed that such predictive pre-activation is a function of its expected utility to a given processing goal, which in turn depends on the comprehender's "estimates of the relative reliability of their prior knowledge and the bottom-up input" (p. 32). We proposed that such *estimates* are likely to be contingent on both the language knowledge, operationalized via a measure of vocabulary size, and confidence, operationalized by self-ascribed Native Speaker (NS) status, of the comprehender. The current study sought to disentangle how knowledge and confidence with a specific language, as opposed to with (any) language more generally, contribute to predictive linguistic processing. We did so by conducting two visual-world experiments looking at a heterogeneous population of adult speakers including simultaneous and sequential bilinguals, who varied substantially regarding their vocabulary size and confidence with English, but who were relatively more equivalent in terms of their age and overall world knowledge. In both experiments, we divided participants into 'higher' and 'lower' skill subgroups by two different criteria: (1) according to a standardized test of vocabulary size, the Peabody Picture Vocabulary Test-Version 4 (PPVT; Dunn & Dunn, 2007), following the procedure that Borovsky et al. (2012) used with monolingual children and adults, and (2) based on a simple dichotomous split according to participants' answer to the question "Do you consider yourself a native speaker of English?" (yes/no)

The two experiments reported in this paper sought to capture how higher and lower skill listeners activate and accommodate both highly-expected and less-expected outcomes by measuring how listeners looked to likely and less-likely sentence outcomes in either typical sentences ending with the highly-expected outcome (Experiment 1), and atypical sentences ending in either less-likely locally-coherent outcomes or much less-likely unrelated outcomes (Experiment 2).

A number of theoretically informative relations between language skill and predictive processing were possible in the current study. One broad potential outcome consistent with prior work on individual differences in lexical and sentential processing (e.g. Borovsky et al., 2012; Fernald, Perfors & Marchman, 2006; Mani & Huettig, 2012), was that higher-skill participants would more quickly interpret spoken input and more robustly generate anticipatory fixations for highly-expected sentential themes than lower-skill listeners. Yet there have been inconsistent findings in the bilingual processing literature as to whether native and non-native speakers differ in speed of language processing (e.g., Kilborn, 1992; Kaan et al., 2015) and whether skill with a specific language affects the degree of anticipatory processing for bi- and multilinguals (e.g., Dijkgraaf, et al., 2016; Dussias et al., 2013; Leal et al., 2016). An alternative possibility was that higher- and lower-skill participants would generate similarly robust predictions for highly-expected sentential outcomes, but vary in their activation of less-likely alternatives, such as items that are "locally coherent" with a recently encountered word (e.g. Kukona et al., 2011). This outcome would be consistent with studies that have reported individual differences in activation of locally-coherent (less-likely) outcomes that vary with language skill and other domain-general cognitive abilities, such as cognitive control (Nozari, Trueswell & Thompson-Schill, 2016; Woodard, Pozzan & Trueswell, 2016). We set out to explore these two potential hypotheses in a series of two studies.

In experiment 1, we asked how listeners in the higher versus lower skill groups predicted both a likely and less-likely locally-coherent outcome during an eye-tracked simple spoken sentence comprehension task where the highly-expected outcome was always mentioned (The pirate chases the *ship*). While there were no significant group differences in the timing of anticipatory fixations towards the highly-likely outcome (SHIP), there were group differences in how listeners considered a less-likely, but locally-coherent sentential outcome that was not mentioned (CAT). Namely, lower-skill participants activated locally coherent lexical outcomes to a greater extent than higher-skill participants did. This result held true regardless of whether the group was split by PPVT score or by self-perceived Native Speakers status. To our knowledge, this study is the first to show such an association between language skill and locally coherent lexical activation in a bi- or multilingual population. One possible explanation for this finding is that lower-skill individuals are more likely to activate a wider range of semantic options during sentence processing, potentially due to greater uncertainty regarding their own skill and understanding. If so, lower-skill participants should show processing advantages in sentence contexts that contain less-expected, though still locally-coherent outcomes. This possible explanation of the findings from experiment 1 motivated the design of experiment 2, which included sentences containing outcomes that were either locally-coherent with the verb (The pirate chases the *cat*) or completely unrelated (The pirate chases the *bone*). Contrary to the results of experiment 1, in experiment 2 we found that higher-skill participants tentatively showed relatively greater facilitation for locally-coherent unexpected sentence outcomes compared to lower-skill participants. While these results were only marginally significant, the moderate/small effect sizes (d = 0.43 and d = 0.38, for NS-No/Yes and PPVT-Lower/Higher group comparisons, respectively) were of similar magnitudes to those seen in experiment 1 (d = -0.45 and d = -0.26), and thus, given recent work emphasizing consideration of effect sizes in addition to p-values (Norris, 2015), we decided to include them in our interpretation.

These findings align with recent work demonstrating that native speakers rapidly adapt their patterns of prediction based on changes in the reliability of cues that might enable prediction (Hopp, 2016), and are most consistent with the hypothesis that higher-skill participants are better able than lower-skill individuals to flexibly adapt to the demands of the current situation. This pattern supports Kuperberg and Jaeger's (2016) assertion that a listener's confidence in her language abilities plays an important role in determining the degree of predictive pre-activation. Indeed, Kuperberg and Jaeger (2016) explicitly link confidence, described in terms of reliability estimates of both bottom-up input and one's own prior knowledge, to the ability to "flexibly adapt comprehension to the demands of a given situation" (p. 45). While Kuperberg and Jaeger did not explicitly mention self-determined native speaker status as a potential variable that could modulate the utility of predictive processing, it seems likely that this variable taps into the same construct. In the context of prediction, prior experience with a specific language can be conceived of as generating probability distributions for potential sentential outcomes. Participants with higher confidence can then be conceived of as generating narrow peaky distributions, equivalent to making a smaller number of strong predictions, which may allow for increased *sensitivity*[8] to either similar (experiment 1) or different (experiment 2) probability distributions in the actual input. In contrast, participants with lower confidence can be conceived of as generating wide flat probability distributions, equivalent to making a larger number of relatively weak predictions, which may decrease sensitivity to the local probability distribution of the actual input. This flatter adaptation pattern is consistent with the lower-skill individuals' performance across both experiments, which suggests that these participants exhibited a general pattern of moderate activation for unlikely sentential outcomes, irrespective of the differential distribution of unlikely outcomes in the local context across the experiments.

Another way of characterizing a wide, flat probability distribution of potential sentential outcomes is as a noisy distribution that is easily affected by interference. This distribution description would be consistent with predictions generated by cue-based models that posit a relationship between individual differences in sentence comprehension and susceptibility to interference (Cunnings, 2016; Van Dyke & Johns, 2012; Van Dyke & McElree, 2006; 2011). While most work in this line of research has focused on written sentence comprehension, a recent study by Sekerina and colleagues (Sekerina, Campanelli, & Van Dyke, 2016) explored the issue using the Visual World Paradigm and presented evidence of locally coherent fixations to extra-sentential competitors as being supportive of interference accounts. This perspective suggests that one promising avenue for future work lies in integrating various

---

[8] Differences in sensitivity could be due to differences in the ability to perceive the probability distribution of the input, responsiveness to such perceptions, or a combination of the two.

theoretical perspectives of language processing, such as the utility account posited by Kuperberg and Jaeger (2016) with that of interference accounts of language processing.

We also consider a number of limitations that constrain the scope of our findings. First, there was a large degree of variability in terms of skill and experience with English within our participant group, much more than the variability along these lines that one might expect to find among monolingual speakers of English. This variability allowed us to create subgroups that can reasonably be referred to as 'higher' and 'lower skill' participants, which was critical for addressing our research questions. Furthermore, there was also large variability in patterns of English-LX relations. While there are often good reasons for the default position in bilingual research of holding such relations constant within a participant population, the variability in LX backgrounds of our sample was advantageous to our specific research questions by introducing substantial variation among participants regarding self-identification as a native speaker, while simultaneously randomizing out the complex effects of English-LX relations. Importantly, this allowed us to include speakers who are often excluded from predefined participant groups in language learning research and are thus likely to be underrepresented in the literature. However, it must be acknowledged that our use of such a heterogeneous population is in and of itself exploratory, and there could be various unknown issues that constrain the generalizability of our results. We also would like to note that all our participants were completing college-level coursework in English, and their vocabulary scores indicated overall good English proficiency. As such, they were all highly skilled in English. Indeed, the fact that participants' English proficiency levels varied within a relatively narrow range could potentially explain the lack of significant differences between groups regarding the timing of anticipatory fixations towards highly-likely outcomes in experiment 1. Thus, future work will be needed to examine whether our findings generalize to participant groups with an even wider range of skill with a specific language.

Furthermore, such future work could benefit from including more comprehensive measures of language skill than we were able to do in the present study. We included the PPVT, a widely used measure of vocabulary size, in part to allow for comparisons with previous work in this line of research with monolingual English speakers (Borovsky et al., 2012). While the PPVT has been used in previous research with bilinguals (Bialystok & Luk, 2012) and PPVT scores have been shown to correlate with TOEFL scores among adult L2 learners of English (Kharkhurin, 2012), we acknowledge that vocabulary size is only one of many aspects of language skill (Grüter, 2017). The inclusion of more comprehensive measures of English proficiency, assessing components such as grammatical competence and fluency, in addition to lexical knowledge, would serve to further elucidate how a wider range of language skills influence bilingual listeners engagement in prediction during language comprehension. Likewise, given the significant differences in self-ratings of LX skill between NS-Yes and -No participants – and in contrast to the relatively small (Experiment 2) or non-existent differences (Experiment 1) between PPVT-Lower and -Higher participants – future work may also benefit from similarly comprehensive measures of LX skill.

As discussed earlier, the inclusion of NS status as a grouping criterion has shown to be beneficial in that it led to more cleanly differentiated subgroups. At the same time, this procedure has limitations in that the sample size of subgroups becomes unpredictable. Thus, although this grouping criterion led to relatively even-sized subgroups in experiment 1, it did not result in even-sized subgroups in experiment 2. In experiment 2, although we tested, as planned, a similar number of participants in experiment 1, less than one-third (n=17) of the participants in experiment 2 identified as native speakers of English. This lop-sided division greatly constrained our ability to explore the differential impact of NS status and vocabulary size on lexical pre-activation. Based on the above sensitivity explanation for the results in these experiments, one might expect higher- and lower-skill listeners to differentially alter the dynamics of lexical activation as a function of the running proportion of non-anomalous/anomalous sentences over the course of a study. For example, one might hypothesize that while both higher-skill and lower-skill listeners would show a graded change in the degree of activation for locally-coherent referents, the slope of the change would be a function of language skill; with higher-skill equating to a steeper slope.

While additional work is necessary, the current results provide some promising insight into how listeners learn to predictively interpret spoken language in an uncertain world. Our research suggests that lexical activation for highly expected and less expected outcomes can be driven by a number of interactive processes that influence the listener over varying time scales that range from lifelong language experience, as reflected by one's knowledge and confidence regarding a language, and locally by the immediate demands of the task itself. For example, highly-skilled listeners' expectations for various sentential outcomes can be shaped by the specific demands/statistical regularities of the task itself, as we saw when they generated strong expectations for highly-expected outcomes in a task where only expected items were mentioned, but then modulated the strength of these expectations in experiment 2, when unexpected items were mentioned. To elucidate this pattern of results, we related our measure of fixation proportions to the probabilistic lexical activation output of Kuperberg & Jaeger's (2016) expected utility of prediction function, and related our measures of language knowledge and confidence, operationalized by a measure of vocabulary size and self-ascribed native speaker status, respectively, to estimates of reliability of prior knowledge and bottom-up input. The relationships between our measures provide support for the view put forward by Kuperberg & Jaeger (2016) that prediction can be understood as a generative probabilistic process operating as a function of its *expected utility* to some processing goal, with estimates dependent on comprehenders' perceptions of the reliability of both the bottom-up input (i.e. the immediate task demands/input) and their own prior knowledge.

To summarize, using eye-tracked measures of predictive processing of simple sentences we found between-group differences in how listeners generated expectancies for less-likely locally-coherent sentential outcomes. While lower- (vs. higher-) skill participants showed greater locally coherent processing for sentences ending with targets aligned with the cumulative set of cues, higher skill participants showed marginally greater locally coherent processing when the set of stimuli included sentences ending with targets that were aligned solely with the local but not cumulative cues. This pattern lends some support to Kuperberg and Jaeger's (2016) proposal that pre-activation in real-time language processing is a function of its utility, estimates of which can be driven by a number of factors that are relevant to the listener's changing goals and experiences, including the listener's knowledge and confidence with a specific language, as well as changes in the nature and demands of the task itself.

Acknowledgements

Financial support was provided by the National Institutes of Health: F32 DC010106, R03 DC013638 to A.B. There are no nonfinancial relationships to disclose.

Works Cited

Arai, M., Van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*(3), 218-250.

Bates, D., Maechler M., Bolker B., & Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Bialystok, E., & Luk, G. (2012). Receptive vocabulary differences in monolingual and bilingual adults. *Bilingualism: Language and Cognition*, *15*(02), 397-401.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417-436.

Borovsky, A., Burns, E., Elman, J. L., & Evans, J. L. (2013). Lexical activation during sentence comprehension in adolescents with history of specific language impairment. *Journal of Communication Disorders*, *46*(5), 413-427.

Borovsky, A., Sweeney, K., Elman, J. L., & Fernald, A. (2014). Real-time interpretation of novel events across childhood. *Journal of Memory and Language, 73*, 1-14.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal*, *64*(3), 311-317.

Chambers, C. G., & Cooke, H. (2009). Lexical competition during second-language listening: sentence context, but not proficiency, constrains interference from the native lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1029.

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, 33(2), 185-209.

Cunnings, I. (2016). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*.

Davies, A. (1991). *The native speaker in applied linguistics*. Edinburgh University Press.

Davies, A. (2003). *The native speaker: Myth and reality* (Vol. 38). Clevedon: Multilingual Matters.

Davies, A. (2013). *Native speakers and native users: Loss and gain*. Cambridge University Press.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test: Manual*. Pearson.

Dijkgraaf, A., Hartsuiker, R., & Duyck, W. (2016). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*.

Dussias, P. E., Kroff, J. R. V., Tamargo, R. E. G., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition*, *35*(02), 353-387.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98.

Grüter, T. (2017). Vocabulary does not equal language, but neither does morphosyntax. *Bilingualism: Language and Cognition*, *20*(1), 17-18.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem?. *Second Language Research*, *28*(2), 191-215.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1352.

Hopp, H. (2013) Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Language Research, 29,* 33–56.

Hopp, H. (2015) Semantics and morphosyntax in L2 predictive sentence processing. *International Review of Applied Linguistics, 53,* 277–306.

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277-307.

Jaeger, B. C. (2017). r2glmm: Computes R Squared for Mixed (Multilevel) Models. R package version 0.1.2. https://CRAN.R-project.org/package=r2glmm.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R 2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086-1105.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?. *Linguistic Approaches to Bilingualism*, *4*(2), 257-282.

Kaan, E., Ballantyne, J. C., & Wijnen, F. (2015). Effects of reading speed on second-language sentence processing. *Applied Psycholinguistics*, *36*(04), 799-830.

Kail, R. V., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, 86, 199–225.

Kharkhurin, A. V. (2012). A preliminary version of an internet-based picture naming test. *Open Journal of Modern Linguistics*, *1*, 34-41. http://dx.doi.org/10.4236/ojml.2012.21005

Kidd, E., & Bavin, E. (2007). Lexical and referential influences on on-line spoken language comprehension: A comparison of adults and primary-school-age children. *First Language*, 27(1), 29-52.

Kilborn, K. (1992). On-line integration of grammatical information in a second language. *Advances in Psychology*, *83*, 337-350.

Knoeferle, P., & Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension?. *Frontiers in Psychology*, *3*(538), 2.

Kukona, A., Fang, S. Y., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*(1), 23-42.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. *Language, Cognition and Neuroscience*, *31*(1), 32-59.

Kuznetsova A., Brockhoff P. B., & Christensen R. H. B., (2016) lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-32. https://CRAN.R-project.org/package=lmerTest.

Leal, T., Slabakova, R., & Farmer, T. A. (2016). The fine-tuning of linguistic expectations over the course of L2 learning. *Studies in Second Language Acquisition*, 1-33.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325-343.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447-464.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, *47*(2), 265-287.

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940-967.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177-190.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86.

Mitsugi, S., & MacWhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition*, *19*(01), 19-35.

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(S1), 97-126.

Nozari, N., Trueswell, J. C., & Thompson-Schill, S. L. (2016). The interplay of local attraction, context and domain-general cognitive control in activation and suppression of semantic distractors during sentence comprehension. *Psychonomic Bulletin & Review*, 1-12.

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rampton, M. B. H. (1990). Displacing the ‘native speaker’: Expertise, affiliation, and inheritance. *ELT Journal*, *44*(2), 97-101.

Sekerina, I. A., Campanelli, L., & Van Dyke, J. A. (2016). Using the Visual World Paradigm to Study Retrieval Interference in Spoken Language Comprehension. *Frontiers in Psychology*, *7*, 873.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49, 238–299.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355-370.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Troyer, M., & Borovsky, A. (2017). Maternal socioeconomic status influences the range of expectations during language comprehension in adulthood. *Cognitive science*, *41*(S6), 1405-1433.

Van Dyke, J. A., & Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Language and linguistics compass*, *6*(4), 193-211.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157-166.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of memory and language*, *65*(3), 247-263.

Woodard, K., Pozzan, L., & Trueswell, J. C. (2016). Taking your own path: Individual differences in executive function and language processing skills in child learners. *Journal of Experimental Child Psychology*, *141*, 187-209.