The effect of isolated words on segmenting noise-band vocoded speech

Katherine M. Simeon[1], Hillary E. Snyder[1], Casey Lew-Williams[2] & Tina M. Grieco-Calub[1]

[1]The Roxelyn & Richard Pepper Department of Communication Sciences & Disorders, Northwestern University, Evanston, IL
[2]Department of Psychology, Princeton University, Princeton, NJ

Background

According to the theory of statistical language learning, humans segment speech and detect word boundaries by tracking transitional probabilities (TP) of different syllable combinations (Saffran, 2003; Singh et al., 2012). TP refers to the likelihood that syllable Y will follow syllable X. Successful segmentation of spoken language assumes that listeners have full access to the spectral and temporal components of the speech signal. The robustness of statistical learning under conditions of degraded speech, however, has yet to be tested. The purpose of this study is test the ability of adults to segment speech that has been spectrally degraded by noise-band vocoding. **Experiment 1 (Exp. 1) tests the hypothesis that statistical language learning is dependent on the spectral fidelity of the speech signal by exposing adults to an artificial language that was either unprocessed or vocoded into 8 or 16 spectral channels. Experiment 2 (Exp. 2) tests the hypothesis that statistical language learning of degraded speech can be facilitated with temporal cues (i.e., silence) around target words.**

Methods

Participants: Exp. 1: 48 young adults (32 female, 18-33 years old); Exp. 2: 42 adults (31 female, 18-34 years old). All participants were native English speakers, with no prior history of hearing loss or speech and language services. All procedures were approved by the Institutional Review Board at Northwestern University.

Stimuli

Artificial language: Four nonsense trisyllabic words were concatenated to create a monotone (female), pause-free speech stream (Duration: 3.25 minutes for Exp. 1; 3.52 minutes for Exp. 2).

Rules of the language:

1) Two words occurred at a **low frequency** (35 times);
2) Two words occurred at a **high frequency** (70 times);
3) The high frequency words combined to generate **two frequency-matched part-words** (35 times);
4) Transitional probability (TP) is the likelihood that **syllable Y** will follow **syllable X**. The **low frequency words** had a TP = 1. The **frequency-matched part-words** had a TP = 0.5.
5) Two languages were used to counterbalance words versus part-words.

	Language 1 (L1)	Language 2 (L2)
Words (TP = 1.0)	pabiku, tibudo	tudaro, pigola
Part-Words (TP = 0.5)	tudaro, pigola	pabiku, tibudo
Non-Words (did not occur in speech stream)	robaku, dolati	robaku, dolati

Listening conditions

The artificial language and test words were either:

1) **unprocessed** (full spectral resolution);
2) noise-band vocoded with **16 frequency channels (16-Ch)**;
3) noise-band vocoded with **8 frequency channels (8-Ch)**.

To create noise-band vocoded conditions, stimuli were bandpass filtered into the relative number of frequency bands (16 or 8, between 200-7000 Hz) using the Greenwood function (Greenwood, 1990). The temporal envelope in each frequency channel was extracted through half-wave rectification and low-pass filtered at 400 Hz (24 dB/octave slope). Bandpass noise served as the carrier for the temporal envelope in each channel, which were then summed.

Temporal gaps (in Exp. 2 only): 10% of the trisyllabic sequences were preceded and followed by a 500 ms interval of silence, to isolate these sequences in time. The isolated trisyllabic sequences were randomly selected from the designated words (TP = 1.0) of the artificial language.

Procedure

Familiarization: Participants were seated 1.5 m away from a Genelec loudspeaker and exposed to L1 or L2 for 6 minutes (Exp. 1) or 7 minutes (Exp. 2) in unprocessed, 16-channel vocoding or 8-channel vocoding. For Exp. 1, N = 16 per listening condition. For Exp. 2, N = 14 per listening condition.

Two-Alternative Forced Choice (2-AFC): Participants heard a word pair and were instructed to select the word that sounded more "familiar." There were 24 trials and trial order was counterbalanced for each participant. The measurement of interest was task accuracy

Experiment 1
Noise-band vocoding disrupts speech segmentation

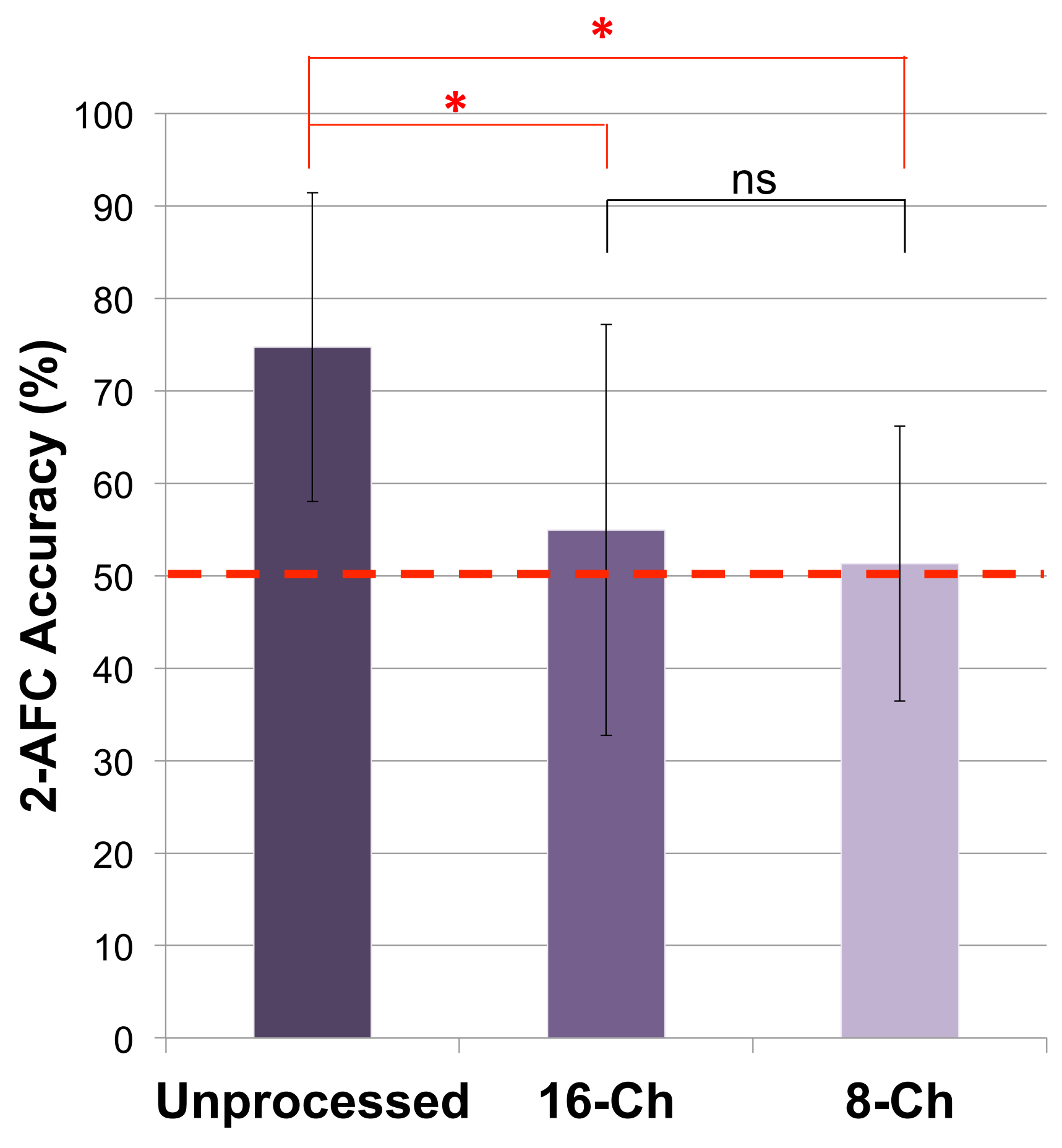

Figure 1: Accuracy (mean + SD) on the 2-AFC task by listening condition. Participants segmented words significantly above chance performance (50%, *red dotted line*) in the unprocessed listening condition. Performance was not statistically different from chance for the 16-Ch and 8-Ch noise-band vocoded conditions. A one-way ANOVA showed a main effect of condition (F (2,45) = 7.7, *p* < 0.05). Post-hoc comparisons revealed statistically significant differences between unprocessed and 16-Ch conditions (*p* < 0.05) as well as the unprocessed and 8-Ch conditions (*p* < 0.05).

Experiment 2
Temporal cues facilitate segmentation of spectrally degraded speech

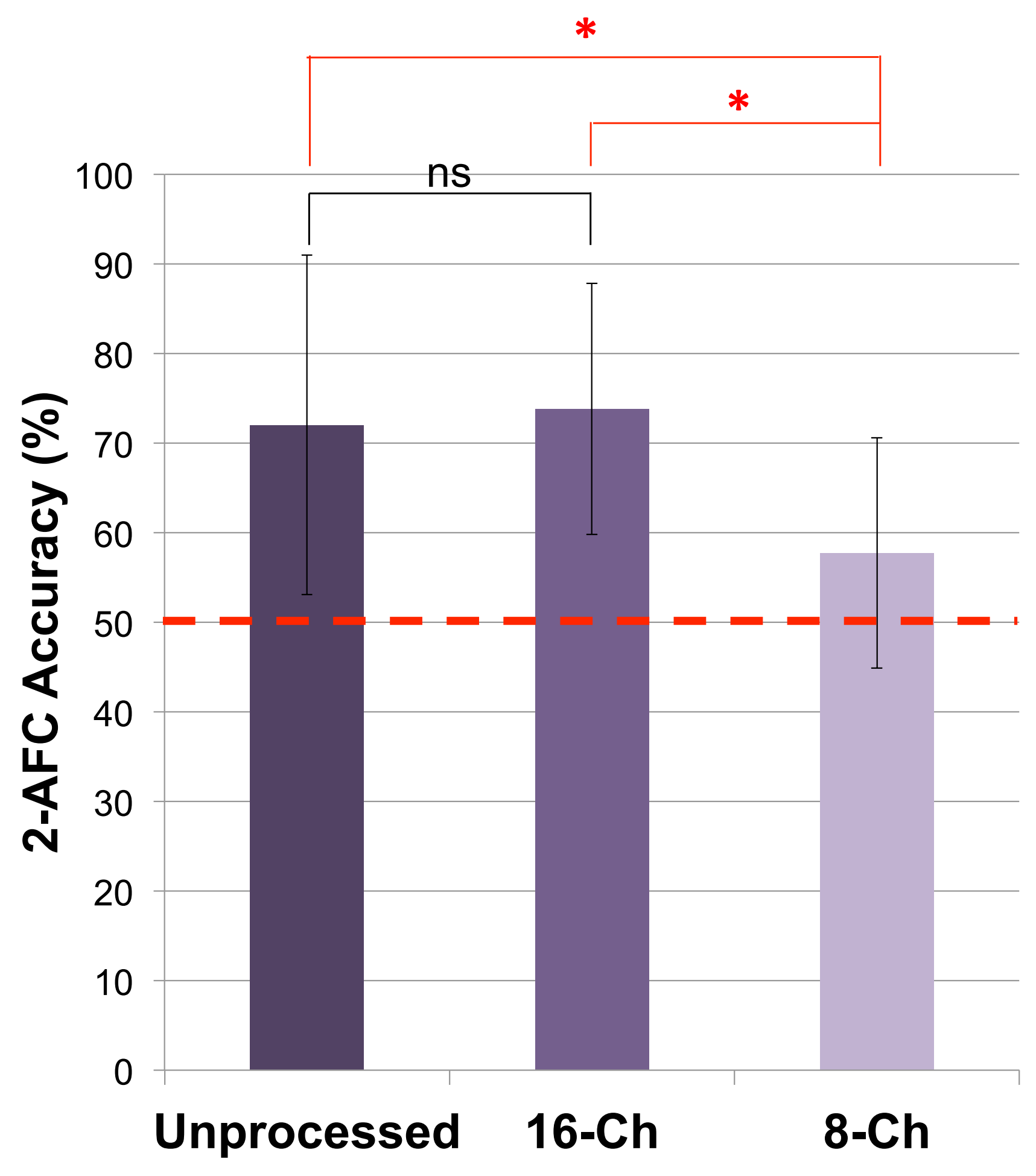

Figure 2: Accuracy (mean + SD) on the 2-AFC task by listening condition. Adults segmented words significantly above chance performance (50%, *red dotted line*) across all listening conditions. A one-way ANOVA revealed a main effect of listening condition (F (2,39) = 4.524, p < 0.05). Post-hoc comparisons revealed a statistically significant difference between 8-Ch and 16-Ch conditions (p < 0.05) and a marginal difference between 8-Ch and unprocessed conditions (p = 0.058).

Comparing the results of Experiments 1 & 2

A 2 x 3 ANOVA was used to test for main effects of experiment (No temporal gap, temporal gap) and listening condition (unprocessed, 16-ch, 8-ch). Results revealed a main effect of experiment (F (1,84) = 4.39, p < 0.05) and a main effect of listening condition (F (2,84) = 9.19, p < 0.05). These results show that, on average, speech segmentation improves with greater spectral detail. Additionally, the results suggest that temporal gaps improve performance on the speech segmentation task. Moreover, the interaction between experiment and condition trended towards significance (F (2,84) = 3.02, p = 0.054). This suggests that the effectiveness of temporal gaps varies based on listening condition. The biggest improvement appeared to occur in the 16-ch noise-band vocoded condition.

Summary & Conclusions

1. Noise-band vocoding disrupts speech segmentation (**Figure 1**). This suggests that listeners who have limited access to spectral fidelity of speech, like those with hearing loss or who use cochlear implants, may have impaired speech segmentation. Because speech segmentation has been promoted as a strategy used by infants in the early stages of language development, these data suggest that infants who hear with a cochlear implant may be particularly at risk for language delays.
2. Although there continued to be a main effect of listening condition, temporal gaps improved speech segmentation abilities (**Figure 2**). The benefit of temporal gaps appeared to be largest for the 16-ch condition. This finding suggests that when spectral fidelity is high (unprocessed condition), temporal gaps do not add benefit. Conversely, when spectral fidelity is very poor (8-ch condition), temporal cues may be limited in their ability to facilitate performance. The 16-ch might represent a particularly good balance of spectral and temporal cues that results in fairly good speech segmentation abilities.

Future Directions

Data collection for Exp. 2 is still in progress in order to achieve equal sample sizes across both experiments; however, the preliminary results of Exp. 1 & 2 motivate potential follow up studies. One question is how do degraded listening conditions contribute to speech segmentation in individuals with hearing loss? While noise-band vocoding simulates how a cochlear implant processes the signal, we are unsure of how a poor signal sounds to a damaged auditory system. A logical next step would be to test cochlear implant users in segmenting speech, without noise-band vocoding, to see how they process language through their implant. Additionally, other cues, such as prosody and speech rate, may contribute to speech segmentation. Modifying the speech stream to mimic cues that are present in natural speech will contribute to current knowledge of how information in the speech signal can compensate for a lack of spectral resolution.

References

Greenwood, D. D. (1990). A cochlear frequency position function for several species—29 years later. *The Journal of the Acoustical Society of America*, 87(6), 2592-2605.

Grieco-Calub, T.M., Snyder, H.E., Reinhart, P.N., & Lew-Williams, C. (submitted). Noise-band vocoding disrupts statistical learning in adults.

Lew-Williams C, Pelucchi B, Saffran JR. Isolated words enhance statistical language learning in infancy. *Developmental Science*. 2011; 14:1323–1329.

Saffran JR. Statistical Language Learning: Mechanisms and Constraints. *Current Directions in Psychological Science*. 2003; 12:110–114.

Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. *Science* 1995;270:303–304.

Singh, L., Steven Reznick, J., Xuehua, L. Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental science*. 2012; 15:482-495.

Acknowledgements

We would like to thank Sarah Schwartz for assistance in data collection. Additionally, we are grateful to David Dorsch, Silvia Lam, and Kristi Ward for comments on this poster.