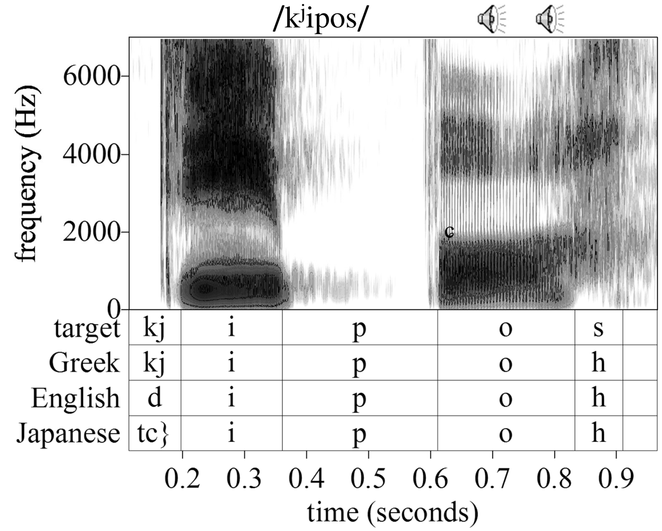
Listener Bias in Categorical and Continuous Measures of Children's Fricatives

Benjamin Munson
University of Minnesota, Minneapolis

Measuring Children's Productions



/k/ before front vowels in Greek:

- Greek speakers mostly hear correct /k/
- English speakers mostly hear [t] substitution
- Japanese mostly hear [tɕ] substitution

Abstractness and Specificity

- The sound structure of language encompasses representations in multiple sensory domains and at multiple levels of abstraction away from raw sensory experiences
- Acquisition happens in all of these domains and processes

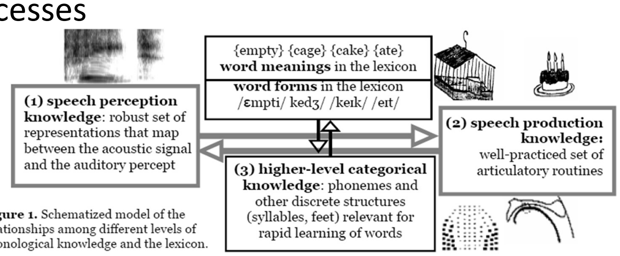
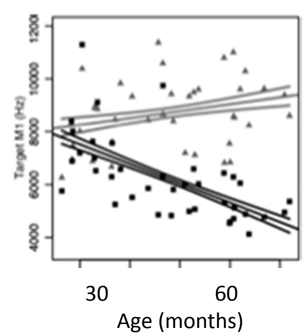


Figure 1. Schematized model of the relationships among different levels of phonological knowledge and the lexicon.

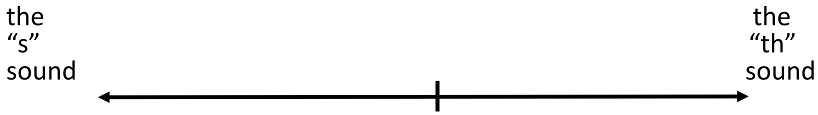
Continuous Differentiation in Development



- Li (2012): the differentiation of /s/ from /ʃ/ centroids between 30 and 60 months is continuous, and extends beyond the point at which these sounds are transcribed to be correct

Visual Analog Scaling

- An alternative: Visual Analog Scaling (VAS) used at least as early as Massaro & Cohen, 1983) to probe adult perception continuously.



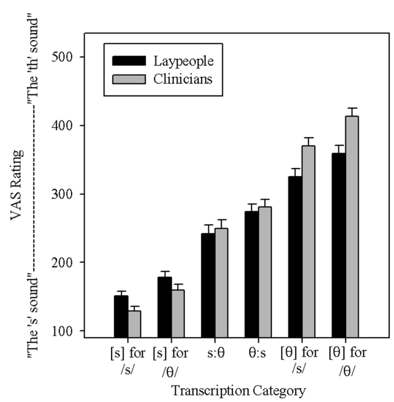
- Participant responds by clicking on the line (a visual analog to the number line, with a neutral midpoint and seemingly infinite variation to the idealized endpoints)
- Used in a variety of recent studies (Bernstein, Johnson, Beckman, Edwards, & Munson, 2015; Julien & Munson, 2012; Munson, Johnson, & Edwards, 2012; Munson, Schellinger, Edwards, Beckman, & Meyer, 2010; Schellinger, Munson, & Edwards, 2015)

Visual Analog Scaling

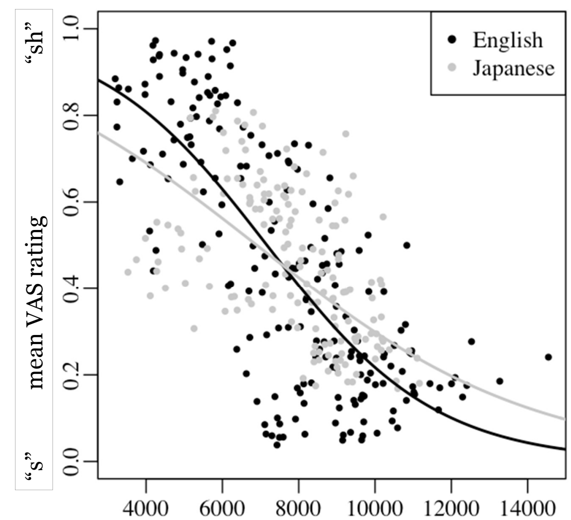
- 200 CV sequences from single-word productions of English-speaking children, aged 2 through 5 years.
  - correct /s/
  - [s] for /θ/
  - intermediate: [s:θ]
  - Intermediate: [θ:s]
  - [θ] for /s/
  - correct /θ/

Visual Analog Scaling

- VAS ratings differentiate among more transcription categories than traditional binary systems do.
- Provides evidence for covert contrasts



VAS Ratings are Related to Acoustics



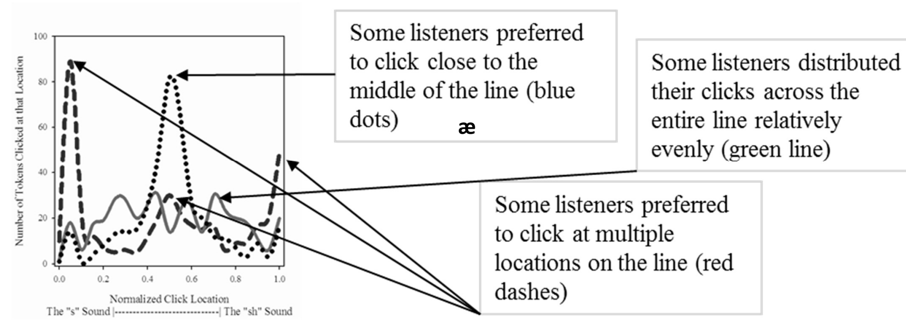
Beckman, Edwards, & Munson, 2011

You’ve Heard this all Before

- I have been presenting work on VAS ratings of children’s speech for the last seven years.
- Today’s new takes:
  - Why are some people more categorical than others?
  - Is it less susceptible to bias than binary measures are?

Degree of Categoricity

Data from Munson & Urberg-Carlson, in prep, data from an “s”-”sh” VAS



Maddeningly Hard to Measure

- Traditional measures of categoricity (like Probit analyses) don’t work, as the continua vary in multiple acoustic dimensions
- This is going to require some creativity with measurement.

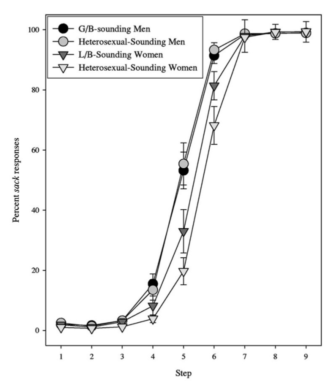
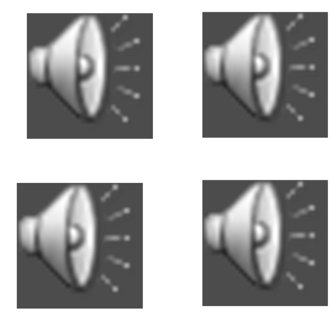


FIG. 1. Percent sack responses for GLB- and heterosexual-sounding men and women on the nine-step sack-shack continuum. Error bars represent one standard error of measurement.

From Munson, Jefferson, & McDonald, 2006

Attention to continuous vs categorical detail

- Can we change the categoricity of someone’s VAS responses by constructing an experiment that draws attention either to categorical detail or to continuous variation?

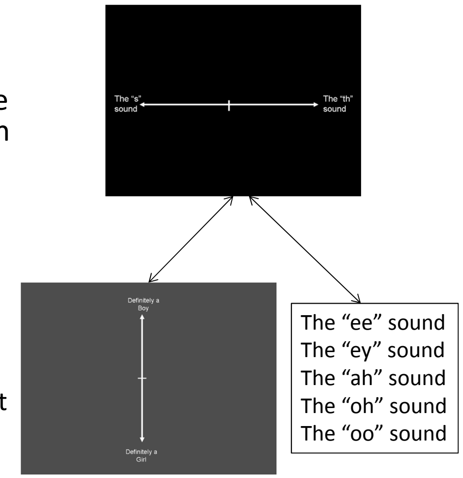


Drawing Attention

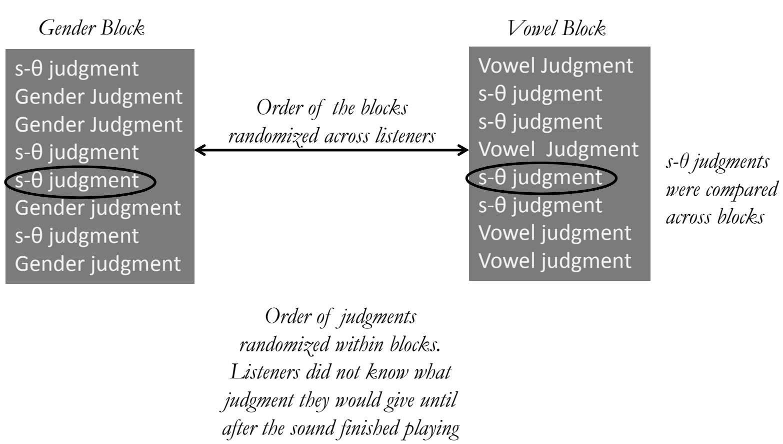
- The general design: interleave VAS ratings of children's /s/ and /θ/ productions on a "s" to "th" scale with judgments of a continuous variable (gender typicality) or a categorical variable (what vowel the child produced).

Experiment Design

- Two blocks: one randomly interleaved /s/-/θ/ ratings with gender judgments, one randomly interleaved it with vowel judgments.
  - Order of the blocks randomized.
  - Listeners never knew what ratings they were making until the stimulus was done playing
  - Each block had 200 stimuli, the same as in Schellinger et al. (2015)

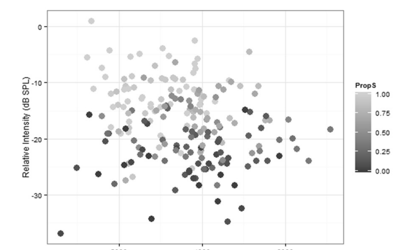


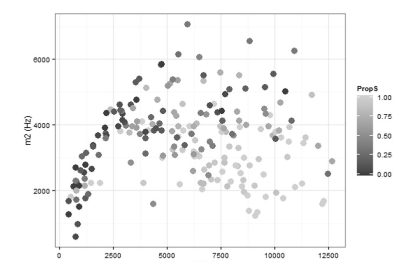
Schematic View of the Experiment



Stimuli

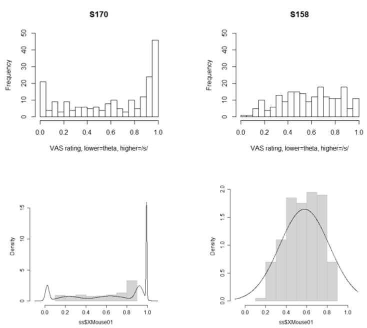
- Children aged 2-5 acquiring English
- Fricatives in initial position, either an /s/ or a /θ/ target
- Transcribed as either, [s], [θ], [s:θ], or [θ:s].
- Varied the acoustic parameters relevant for the /s/-/θ/ contrast

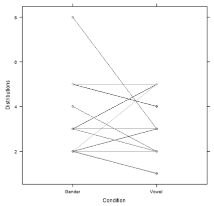




Analysis 1: Distributions

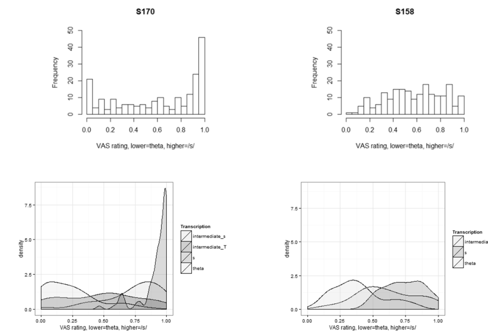
- Density Mixture Modeling in `mclust`
- Did the conditions differ in the shape of the response distributions?
- No

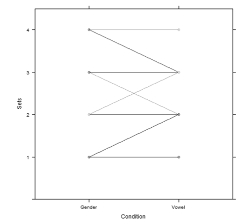




Analysis 2: Differentiation

- Did the conditions differ in how many transcription categories they differentiated among?
- No





Analysis 3: Acoustics

- Did the conditions differ in how strongly the responses were affected by the relevant acoustic characteristics of the fricatives (m1, m2, intensity relative to the following vowel)?
- No

	Estimate	Std.Error	df	t value	Pr(>|t|)	
(Intercept)	4.164e-01	2.032e-02	21	20.490	<0.00001	***
M1, middle 40 ms	-6.865e-02	6.207e-03	26	-11.060	<0.00001	***
M2, middle 40 ms	2.849e-02	4.757e-03	21	5.991	<0.00001	***
F2 at vowel onset	-7.669e-03	4.244e-03	21	-1.807	0.085	.
F0 at vowel midpoint	-2.156e-02	4.073e-03	23	-5.294	<0.00001	***
Duration	3.383e-02	3.965e-03	40.7	8.532	<0.00001	***
Relative Intensity	-1.084e-01	9.963e-03	22	-10.877	<0.00001	***

Interim Conclusion

- VAS ratings are impervious to whether they are paired with a task that asks people to rate gender or one that asks them to rate the vowel that they heard.

Is it VAS or the Experiment?

- Is VAS impervious to bias, or does this particular experimental manipulation simply not bias responses?

Redo the Experiment

- We redid the experiment with a new set of listeners. It was identical in all ways except one: listeners made a binary response of whether they heard “s” or “th” rather than a VAS judgment.
- Binary judgments in both conditions differentiated among all six transcription categories

Logit Mixed-Effects Model: Acoustics

- The two conditions differed in the weighting that listeners gave to the stimuli

	Estimate	Std.Error	z value	Pr(>|z|	
(Intercept)	0.451709	0.104600	-4.318	<0.0001	***
M1, Middle 40 ms	-0.985314	0.053929	-18.270	<0.0001	***
M2, Middle 40 ms	0.598601	0.046904	12.762	<0.0001	***
F2 at vowel onset	-0.150684	0.038941	-3.870	0.0001	***
F0 at vowel midpoint	-0.340250	0.040636	-8.373	<0.0001	***
Duration	0.373812	0.045587	8.200	<0.0001	***
Relative Intensity	-1.312905	0.063463	-20.688	<0.0001	***
M1 by Condition	**0.142988**	**0.048544**	**2.946**	**0.0032**	**
Midpoint F0 by Condition	**0.093393**	**0.040295**	**2.318**	**0.0204**	*

Logit Mixed-Effects Model: Acoustics

- If you convert the data from Experiment 1 to binary responses and do the same analysis, the acoustics do not differ as a function of experiment.

	Estimate	Std.Error	z-value	Pr(>|z|)	
(Intercept)	-0.67810	0.18076	-3.751	0.000176	***
M1, middle 40 ms	-0.58186	0.06084	-9.564	<0.000001	***
M2, Middle 40 ms	0.30108	0.05209	5.780	<0.000001	***
F2 at Onset	-0.07477	0.04100	-1.824	0.068195	.
F0 at Midpoint	-0.27981	0.04143	-6.754	<0.000001	***
Duration	0.29393	0.04293	6.848	<0.000001	***
Relative Intensity	-1.14442	0.09497	-12.050	<0.000001	***

Logit Mixed-Effects Model: Acoustics

- In the gender condition, listeners weighted m1 more strongly than they weighted it in the vowel condition
- In the gender condition, listeners attended to f0 when making judgments; in the vowel condition, they did not

Conclusion

- VAS ratings are more stable than binary ratings to the bias introduced by mixing listener ratings with ratings of continuous or categorical responses.
- Bias is introduced at the decision stage, not in the encoding stage

Future Work

- Look at other, more conventional ways of biasing responses:
  - Long- versus short-lag responses (as in Babel & Johnson, 2010)
  - Priming a bimodal or unimodal distribution (as in Clayards et al., 2008)
    - Re-analyzing these data to determine whether the condition effects are really just response-latency effects.

Acknowledgments

This work was done in collaboration with the following people, but they have not reviewed the content of this presentation

Jan Edwards
University of Wisconsin, Madison
Sarah Schellinger
St. Xavier University, Chicago

- Thanks to our funding sources
  - NSF Grant BCS0729277 to Benjamin Munson
  - NSF Grant to Jan Edwards
  - NIH grant R01 DC002932 (originally awarded to Jan Edwards and Mary E. Beckman, later to Jan Edwards, Mary E. Beckman, and Benjamin Munson [2011-2016]).

References

Bernstein, S., Johnson, A., Beckman, M.E., Edwards, J., & Munson, B. (2015). *Child-Level Factors and the Acquisition of the /t/-/k/ Contrast: Perception.* Poster to be presented at the American Speech-Language-Hearing Association Annual Meeting, Denver, CO. See also http://learningtotalk.org/node/556

Julien, H., & Munson, B. (2012). Modifying Speech to Children based on their Perceived Phonetic Accuracy. *Journal of Speech, Language, and Hearing Research, 55*, 1836-1849.

Li, F. (2012). Language-specific development differences in speech production: a cross-language acoustic study. *Child Development, 83*, 1303-1315.

Massaro, D., & Cohen, M. (1983). Categorical or continuous speech perception: a new test. *Speech Communication, 2*, 15-35.

Munson, B., Johnson, J., & Edwards, J. (2012). The Role of Experience in the Perception of Phonetic Detail in Children's Speech: A Comparison of Speech-Language Pathologists with Clinically Untrained Listeners. *American Journal of Speech-Language Pathology, 24*, 124-139.

Munson, B., Edwards, J., Schellinger, S.K., Beckman, M.E., & Meyer, M.K. (2010). Deconstructing phonetic transcription: covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical Linguistics and Phonetics, 24*, 245-260.

Munson, B., & Urberg Carlson, K. (2015). *An Exploration of Methods for Rating Children's Productions of Sibilant Fricatives*. Manuscript under consideration.

Schellinger, S., Munson, B., & Edwards, J. (2015). *Gradient Perception of Children's Productions of /s/ and /θ/: A Comparative Study of Rating Methods*. Manuscript under consideration.