Chapter 6

Forced-choice Perceptual Dialect Categorization

6.1 Introduction

The ability of listeners to categorize talkers by dialect has been studied with some interest in recent decades. While studies of variation in *production* dominated the fields of dialectology and sociolinguistics from the late 19th century in Germany (by Georg Wenker) and France (by Jules Gilliéron), spreading to Switzerland and Italy (Jaberg & Jud, 1928–1940), the British Isles (Orton, 1962), and the United States (e.g., Kurath, 1949), interest in *perception* of variation took off much more recently, coinciding with the rise of the fields of perceptual dialectology and sociophonetics in the 1980s and 90s (see Preston, 1999a; Thomas, 2002).

A number of studies involving various languages have considered how listeners explicitly categorize talkers into discrete dialect regions. In an early study of dialect identification, Preston (1993) played excepts of interviews from talkers from nine different equidistant locations between Saginaw, Michigan and Dothan, Alabama. Listeners in Michigan and southern Indiana responded which of the nine locations they believed each talker was from. Response patterns showed that listeners distinguished broadly between two groups of talkers, but finer distinctions were less obvious. While listeners in both study sites distinguished the two groups, the particular location of the boundary between the two was further south, on average, for the Indiana listeners.

Clopper and Pisoni (2004b) expanded the scope of their study to dialects across the entire United States. Using sentence-length recordings from the TIMIT corpus (Zue, S., & Glass, 1990), they played recordings from six dialect regions spoken by males and females with a range of ages, ethnicities, and social backgrounds. In a six-alternative forced-choice task, listeners selected which of six geographic regions they believed each talker was from. Overall accuracy was 31%, statistically above chance (16.7%). However, systematic confusions between phonologically similar dialects were found. Three perceived dialect regions were found, a Northeast, a Southern, and a Western.

Clopper (2004) used a nearly identical protocol as Clopper and Pisoni (2004b) except talkers came from the Nationwide Speech Project (NSP) corpus (Clopper & Pisoni, 2006b). The NSP corpus contains talkers from a much narrower range of ages and backgrounds; all of them were undergraduate students at Indiana University, the same university where participants were recruited. These stimuli, likewise, presented a narrower range of linguistic variation than the TIMIT stimuli. Listeners were also either mobile, having lived in at least three different places, or non-mobile, having lived in only one place. In this variation of the protocol, participants had an overall accuracy of 26%, which was still statistically above chance for a six-alternative task. The same perceived dialect groups emerged as in the TIMIT-based study: Northeast, South, and Midwest/West. Non-mobile listeners from the North dialect region, though, perceived the Northern and Midland as more similar than did the mobile listeners. That is, the listeners who had greater experience with dialectal variation in the United States recognized a distinction between the North and Midlands dialects that the people who grew up and remained in the North did not recognize as clearly.

Perceptual identification of other dialects of English have also been studied. Van Bezooijen and Gooskens (1999) played brief narratives from talkers from five regions of the United Kingdom to UK listeners. The accuracy rates for country, region, and area of origin were quite high at 92%, 88%, and 52%, respectively.

Beyond the realm of English dialect perception, Van Bezooijen and Gooskens (1999) also used a similar protocol in the Netherlands. Talkers from four regions of the Netherlands and Dutch-speaking Belgium were played to Dutch listeners. Accuracy rates were also high in this study compared to the American results with listeners accurately identifying the country, region, and province of the talkers 90%, 60%, and 40% of the time, respectively.

German-speakers' perceptions were studied by Burger and Draxler (1998). Strings of digits (telephone numbers) were recorded by male and female talkers from 23 German dialect regions. Listeners selected from among these regions where they believed each talker was from. Remarkably, the accuracy rate for the exact region was 23% and 49% for the broad region. Swiss, Austrian, and Saxonian varieties were identified the most accurately at 83%, 58%, and 42%, respectively. The region of Hanover was the default region for varieties of High German perceived as unmarked.

In the European Francophone world, Woehrling and Mareüil (2006) presented listeners in Paris and Marseilles (the second largest city in France, in the south) with talkers from six regions across France. Listeners in both research sites had similar accuracy rates around 43%. The highest accuracy was for the Swiss accent. Considering the patterns of confusions, it was clear that Normandy in the north and Marseilles in the south serve as the mental references for stereotypical representations of their regions. Thus, three perceived dialect groups emerged in their analysis: a north, a south, and Swiss. The Basque Country was included in the southern group, but its status is more complicated. It was confused with the other non-Marseilles set, Languedoc, but much less so with Marseilles. There appears to be a continuum in the southern region with the Basque Country and Marseilles on the ends and Languedoc in between, a pattern that matches the geographic locations of these regions.

Boughton (2006) considered the role of social class in the identification of regional varieties of French. Listeners in eastern France from the the Pays de la Loire region listened to 68-word extracts of scripted speech of talkers from their own region and from the western city of Nancy who were from the working and middle classes and were either 'younger' or 'older'. Listeners indicated where they believed the talkers were from in a free-response identification task. The results showed that regional identification accuracies was fairly low for talkers from both regions at around 25%. However, when considering the social identities of the talkers, it appeared that listeners were treating the task as a standard/non-standard dialect identification task with the geographic labels for 'North' and 'East' corresponded to classifications of speech perceived as non-standard and 'West', 'Center', and 'Paris' corresponding to perceived standardness. Thus, even though the task was explicitly about the identification of regional speech, participants implicitly treated it as a social evaluation task.

In a broader study of French perceptions including Belgium and Switzerland in addition to France, Avanzi and de Mareüil (2017) conducted two studies. In the first, listeners in all three countries were asked to identify the country of origin of talkers from the three countries. The overall accuracy by country was consistent at around 60% (where chance is 33%). There was an own-country-accuracy-advantage for the Belgian and Swiss listeners and a general bias to respond France. The French listeners had difficulty discriminating between Belgian and Swiss talkers, suggesting only a two-way distinction, France and "other".

In a second study, Avanzi and de Mareüil (2017) considered French, Belgian, and Swiss listeners' ability to identify five regional variants within their own countries. The overall accuracies were fairly low within each country at around 30% but still above chance (20%).

The general observation across these studies is that listeners are generally better at identifying dialects in broader geographic regions than narrower ones. Dialects associated with regions that also have strong cultural identities, such as individual nations or important cultural regions within a nation (e.g., Normandy and Provence in France, or the North and the South in the US) are best identified and demonstrate the most confusion between dialects in those regions.

The maximum number of reliable perceptual groupings appears to be three. The general status of these groups is a local two-way distinction plus an additional "other" group. This pattern is seen most clearly in the US and French studies. In the US there is a South and Midwest/West local distinction plus the Northeast "other". In France there is a North and South local distinction plus a Swiss/Belgian "other". The German study suggests a similar pattern as well, but it does not emerge as clearly; although, upon further inspection it may. The UK results suggest that people in the UK are able to identify at least five regions specifically and fairly reliably. This could suggest that the two local groups plus one "other" is not necessarily a maximum but tendency. In any case, the UK would seem to be unique in this regard based on the studies reviewed here. Perhaps other countries with long linguistic and dialectal histories will also show a specific distinction between several regional varieties, for example, Korea or Japan.

6.2 Methods

6.2.1 Listeners

Participants in this experiment come from the statewide study. 6 participants were unable to complete the task due to technical difficulties,[1] so the total number of participants included in this experiment is 103. No participants were excluded due to poor performance as the accuracies of individuals are normally distributed as shown in Figure 6.1 and an Anderson-Darling test was not significant (A $= 0.48$, $p = 0.22$). Furthermore, none of the participants can be considered outliers based on their divergence from predicted values.

[1] The program used to present the experiment would spontaneously crash. I was unable to diagnose the source of the problem. I suspect the issue had to do with the way memory was allocated by the audio module.

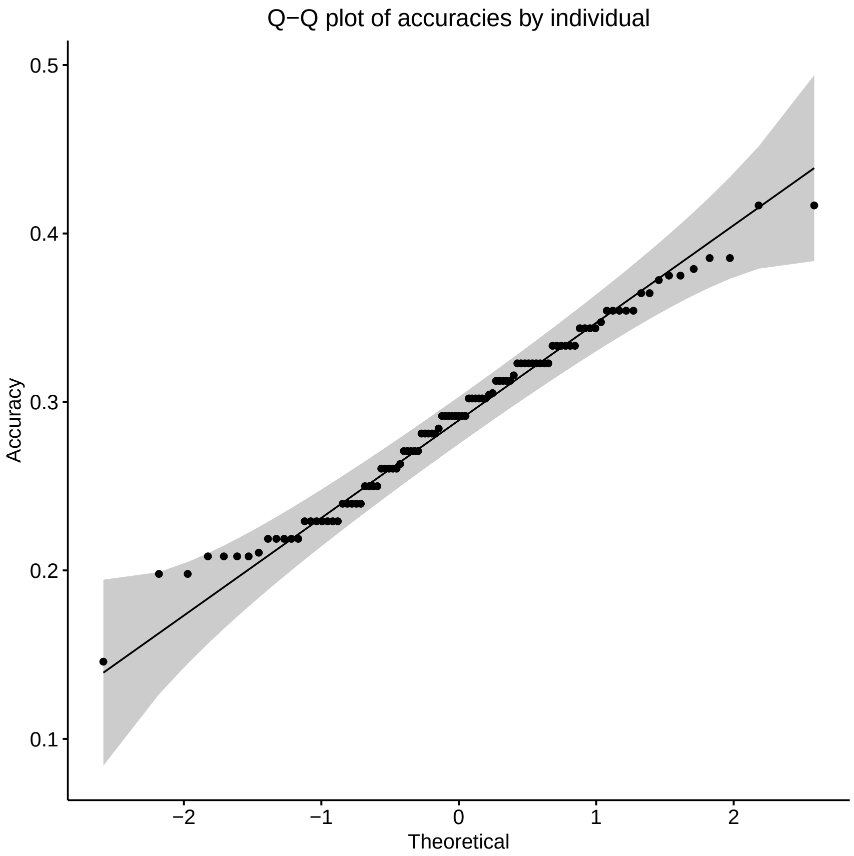

Figure 6.1: Q-Q plot of accuracies for individual participants

6.2.2 Talkers

Twenty-four female talkers were selected from the Indiana Speech Project Corpus (Clopper et al., 2002). The talkers included four talkers from each of six geographic regions in Indiana included in the corpus: near Chicago, Fort Wayne, Indianapolis, Bloomington, near Louisville, and Evansville. All talkers were in their early 20s at the time of the recording. The corpus includes only female talkers. The first four talkers from each region were selected.

The six geographic regions lie within the Inland (near Chicago), North (Fort Wayne), Midland (Indianapolis and Bloomington), and South (near Louisville and Evansville) dialect regions as described in the *Atlas of North American English* (Labov et al., 2005). The characteristics of these dialects are described in detail in Section 2.3.

6.2.3 Stimulus Materials

A list of stimuli were selected from the Indiana Speech Project Corpus (Clopper et al., 2002). Sixteen sentences were chosen so that each of the four talkers in the six talker groups would read four sentences. The same sixteen sentences were repeated for each talker group for a total of 96 unique tokens.

The sixteen sentences are of two general types: 1) those containing a phoneme that is stereotypical of a dialect region, and 2) those containing no phonemes that are stereotypical of a dialect region. The stereotype-containing sentences are of three types: 1) containing the Southern stereotype phoneme /ay/ before voiced segments; 2) containing the Northern stereotype phoneme /ae/, without specifications on following phonological environment; and 3) containing both stereotype phonemes /ay/ and /ae/. The non-stereotype sentences do not contain either /ay/ or /ae/ in content words; function words such as "an" and "by" were allowed as they were unavoidable.

All of the sentences are drawn from the high probability sentences of the Speech Perception in Noise (SPIN) test (Kalikow, Stevens, & Elliott, 1977). They are meaningful English sentences in which the final word is highly predictable based on the semantic context of the sentence. Some examples of these sentences are shown below. The full list of stimuli can be found in the Appendix at the end of this chapter.

/ae/ only	Paul hit the water with a splash.
/ay/ only	Her entry should win first prize.
/ae/ and /ay/	The flashlight casts a bright beam.
Non-stereotype	A round hole won't take a square peg.

The original sound files were cropped so that they only contained relevant speech material. The files were in 16-bit encoding with a sampling rate of 44,100 Hz. All the files were leveled to have the same average intensity using Praat's built-in `scale intensity` function (Boersma & Weenink, 2019).

6.2.4 Procedure

Participants were seated at a computer equipped with a mouse, trackpad, and circumaural headphones. The experiment was written in Python using PsychoPy2 version 1.85.4. (Peirce et al., 2019). Participants were first guided through a practice block of four trials where they were able to become familiar with the format of the experiment. The experimental block consisted of 96 trials. The stimuli were presented in a new random order for each participant. On each trial participants were asked to listen to the stimulus and select the region of Indiana where they believed the talker was from. The response alternatives were displayed on a multi-colored map of Indiana as shown in Figure 6.2. The alternatives and their placement on the map were based on the analysis of perceptual dialect maps presented in Chapter 4. Participants were only able to listen to each stimulus a single time.

6.2.5 Analysis Methods

This section describes methods used to analyze the four-alternative forced-choice categorization data.

6.2.6 Structure of the Classification Data

The basic structure of the data for this experiment involves a label for the stimulus and a label for the response. The data can analyzed as a confusion matrix showing the region of origin of the talker and the response of the participant.

The patterns of responses to stimuli can be depicted as a network graph. The stimuli categories are represented as the nodes while the response categories are represented as the lines connecting the nodes (these are called *vertices* and *edges*, respectively, in graph theory).

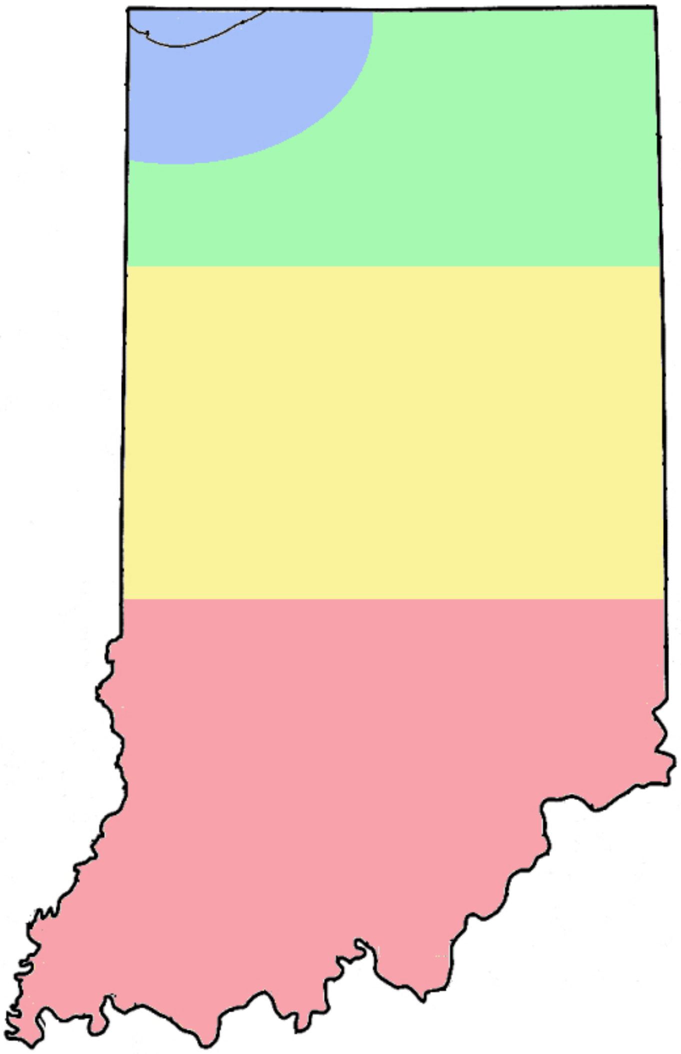

Figure 6.2: Response alternatives in the four-alternative forced-choice categorization task

Types of Classification Patterns

The analysis of forced-choice classification data is not so much about the accuracy of the classification but about the patterns of errors. Certainly we are interested in the accuracy, but we are also interested in the ways that the stimuli are misclassified. Therefore, in this section we will consider some types of patterns that may logically be present in forced-choice classification data.

Some concepts from signal detection theory are relevant here. In its most basic form, a signal detection experiment asks participants to indicate if a stimulus is present or not. For example, the participant may be attempting to detect a "beep". The researcher may manipulate some parameter of the sound such as its loudness. Loud beeps will almost always be detected; this is called a "hit" or a "true positive". When a stimulus is not presented during a trial, the participant will almost always indicate such; this is called a "correct rejection". Both of these are correct responses. There are, likewise, two types of incorrect responses. A "false alarm" is when the participant indicates a stimulus was present when in fact is was not. A "miss" is when the participant fails to detect a stimulus that was present. By examining these patterns of correct and incorrect responses, it is possible to determine a threshold for detecting a stimulus as well as biases in the responses.

An important point to note is the effect of a response bias on the results. If a participant always responds positively, that is, indicates that a stimulus is present, and there actually is a stimulus present half of the time, then the participant will have an accuracy of 50%. However, the participant also was incorrect half of the time.

Another key point is that participants may have relatively higher and lower thresholds for a wide range of reasons. For example, younger participants may have a lower threshold and be able to detect quieter sounds. Others may have higher thresholds because their personality makes them more reluctant to respond positively when they are uncertain a stimulus is present. If the pitch of the sound is manipulated as well at the intensity, maybe some participants will be more sensitive to certain pitches regardless of the intensity.

In a four-alternative forced-choice paradigm, there are additional types of response patterns that are logically possible. These are given below:

correct: the classification is correct

guess: a random category is chosen

default category bias: one category is chosen by default

specific sensitivity bias: the classifier is more sensitive to one category than the others

category confusion: two categories are confused for one another

These types of patterns may also be combined. For example, there can be multiple types of category confusion. If someone cannot distinguish between two categories, the response might involve guessing (randomly choosing either category) or a default category bias (when a stimulus of either of the confused categories is encountered, only one of the categories is selected). Also, a correct response might be the result of a guess.

Of course, in a paradigm with four categories, responses may include several of these response patterns. A single participant may confuse two categories and have a sensitivity bias for a one particular category. In fact, such a set of patterns could be quite common. It is not difficult to imagine that someone from southern Indiana can easily identify another talker from southern Indiana while not being able to distinguish if someone is from near Chicago or from another part of northern Indiana.

Definitions of Measurements

As mentioned earlier, forced-choice responses can be mined for more interesting insights than mere accuracy. While accuracy is an important measure of participants' response patterns, we are also interested in sensitivity (recall), specificity, precision, detection prevalence, and balanced accuracy. The following definitions are provided in the Caret package for R (Kuhn, 2008). To help define the measurements, we can consider the four possible responses described in the previous section.

The columns represent status of the stimulus, whether the event occurred or not. The rows represent the response, whether an event was reported as occurring or not. Because

Response	Stimulus	
	Event	**No Event**
Event	A	B
No Event	C	D

the measures described in this section only take into account whether a specific stimulus was categorized correctly or not, this 2x2 matrix is sufficient to summarize the response patterns. Another method involving network graphs will be used to further summarize the between-category response patterns.

A positive correct response is when an event occurred, and the response was that the event occurred, A in the table. As an example in the context of the experiment, this would be when the stimulus token was spoken by someone from the North region and the participant responded that the talker was from the North region. A positive incorrect response, shown at B in the table, would be when the participant responded North but the talker was actually from one of the other three regions. A negative incorrect response, shown at C in the table, is when the talker was from the North region but the participant chose another region. Finally, a negative correct response, shown at D, is when the talker was not from the North and the participant chose a region that was not the North.

Accuracy is presented as the overall proportion of correct positive responses across all four categories of stimuli. The other measurements are calculated and presented according to the stimulus categories. The formulas for calculating the measurements are given below (Kuhn, 2008):

Sensitivity = A/(A + C)

Specificity = D/(B + D)

Precision = A/(A + B)

Detection Prevalence = (A + B)/(A + B + C + D)

Balanced Accuracy = (sensitivity + specificity)/2

Sensitivity, sometimes called *recall*, is proportion of positive correct responses to total number of times an event occurred. That is, when a stimulus was spoken by someone from, say, the South, how often did the participant respond South. The meaning of this measurement should be fairly intuitive; it is what is normally thought of as accuracy. *Specificity* measures another type of correct response except it is the negative response accuracy. In other words, when the talker was not from the South region, how often did the participant not respond South. This measurement is a less intuitive form of accuracy; it shows how well people can actually avoid misclassifying stimuli. *Precision* measures how often a response is correct when a particular category is chosen. A relatively low value for precision indicates that when a category is selected, the response is incorrect. A low precision value might occur when a single category is selected by default or when there is a confusion between two categories that results in guessing. A high precision value might occur when tokens of a particular category are highly salient such that they are easily recognizable as being a member of the category, and the category is not easily confusable with another. *Detection Prevalence* measures how frequently each category was selected. It indicates if any categories were selected more or less frequently than others. *Balanced Accuracy* is the average of both measures of accuracy, positive and negative. Using the terminology of signal detection theory, it is the average of the proportions of *hits* and *correct rejections*. It gives a sense of how accurate responses are to each category overall.

Network Graphs

Network graphs are a way to show connections and relationships between members in a system. A network graph consists of *nodes* and *edges*. In a social network, we might be interested in showing who knows whom. The individual people in the network are represented by nodes. The dimension indicating that two people know each other can be represented by an edge (line) connecting the nodes. We might also be interested in how frequently the various members of the community interact with each other. There are several ways this dimension could be displayed, such as position in the graph or color of the edge, but one of the most straightforward ways to represent a single numerical dimension is by varying the width of the edge. People who interact a lot are connected by a broad edge while people who interact infrequently are connected by a narrow edge.

In a forced-choice categorization paradigm, a network graph can neatly summarize patterns of confusions between stimulus and response. Such a graph represents the stimulus categories as nodes and the responses as edges pointing to the node associated with the response. The width of the edge represents the frequency that a each response was chosen. Because it is likely that each pair of categories is not equally confusable—and we are, in fact, particularly interested in such inequalities—each pair of nodes are connected by two directional edges. There are also self loops, that is, edges pointing from each node back to itself, indicating how often the stimulus and response categories were the same. The self loop can be thought of as representing the “correct” response.

Simulations of Response Patterns

In anticipation of the response patterns of the participants’ being less clear and involving multiple combinations of patterns, we can simulate the response patterns described above to see how the patterns look in an ideal form. We will consider patterns where responses are: 1) completely random, 2) mostly random, 3) mostly correct, 4) biased toward a single category, 5) selectively biased toward a particular category, 6) the result of a confusion between two categories with guessing, and 7) the result of a confusion between two categories with a default category bias. The analysis of these simulated results will include the resulting confusion matrices; relevant measures such as accuracy, sensitivity, specificity, These results are based on simulations of 20 iterations of 96 responses (as if 20 participants completed the task, roughly the number of participants in each experimental group). Random selections were made from a set of relevant options with replacement using the `sample` function in the Base package in R (R Core Team, 2018). The four stimulus/response categories are referenced by alphabetical labels: A, B, C, D. Stimuli were simulated with the same frequency by category as in the actual task completed by participants. Two categories had 16 occurrences each (A and B), and two had 32 each (C and D). A vector of "correct" responses was created, and all of the simulations are based on probabilistic changes to this response vector. All of the simulated data presented below were run through the same analysis process as the actual participant data presented later.

The results of the random response simulation are shown in Table 6.1 and Figure 6.3. In this simulation all of the responses were randomly selected from among the four response categories.

The first thing to notice about the randomly generated responses is that the overall accuracy is 25%, exactly what would be expected from randomly selecting from four categories. Of course, this is only one simulation, and it is certainly within the realm of probability that the accuracy could have been higher or lower. However, the range that would be expected by chance with 96 responses x 20 iterations = 1920 is fairly narrow. With this many responses, it would be reasonable to expect that results are still random if the accuracy were between 23% and 27%, the theoretical range of a 95% confidence interval (CI) for this amount of data.

Sensitivity is essentially the accuracy for each category. B, C, D are both close to the predicted sensitivity level. A, however, is further away from the expected average. This greater divergence from the overall mean is due to the lower frequency of the A and B categories, which each account for only 1/6 of the total data, or 320 responses each. In this

Table 6.1: Simulation of completely random responses

Overall Accuracy	0.25			
	A	B	C	D
Sensitivity	0.30	0.27	0.25	0.22
Specificity	0.74	0.75	0.75	0.77
Precision	0.19	0.18	0.33	0.33
Detection Prevalence	0.26	0.25	0.25	0.23
Balanced Accuracy	0.52	0.51	0.50	0.50

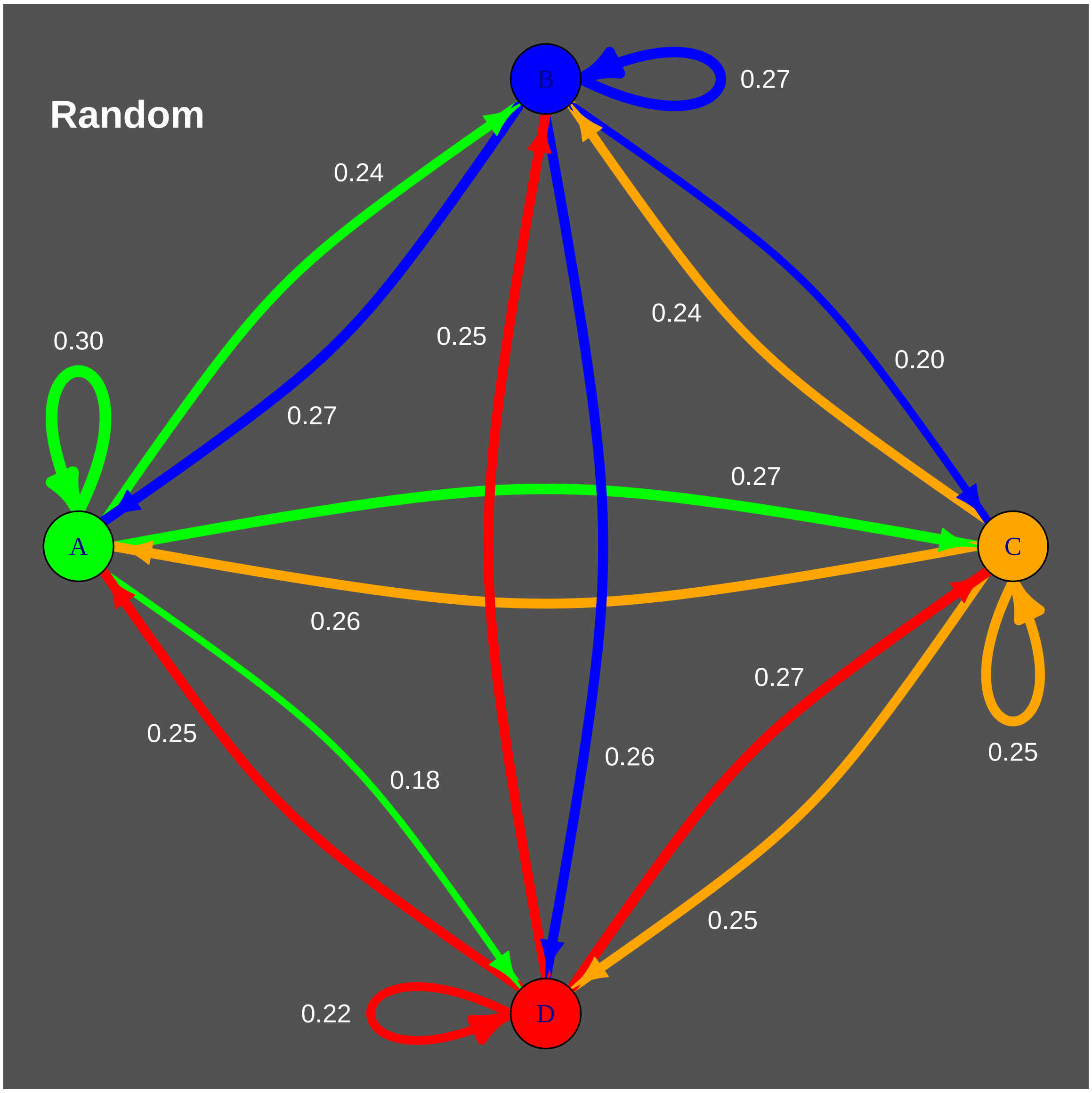

Figure 6.3: Simulation of completely random responses

case the 95% CI spans from 20% to 30%. The 95% CI for the categories with 32 stimuli, C and D, is narrower at 21.5% to 28.5%.

Specificity is the rate at which the incorrect responses were avoided. These values are essentially independent of sensitivity. The categories with higher sensitivity values do not have lower specificity values.

Because precision is calculated based on response frequencies rather than stimulus prevalence, the categories with fewer stimuli have lower precision values than those with more stimuli when responses are random. In fact, it would take a rather large response bias for one of the categories with few stimuli to overcome the stimulus prevalence bias.

Detection prevalence only takes into account the response frequency for each category and the total number of stimuli. In the actual experimental data, this would be the same as the proportion of times participants selected each region category. In this simulation, the values should be close to 25%. We can observe that the detection prevalence for category A is only 26% while its sensitivity is quite high at 30%. This means that even though category A was selected at the same frequency as the the others, it happened to also be the correct response more frequently than with the other categories. Nonetheless, this pattern is within the range of what would be expected if the results were random, which is the case.

Finally, the balanced accuracy presents the average of the hits and correct rejections for each category. All of the values are close to 50%. The value for category A is slightly higher at 52%, which represents the fact that its sensitivity value was relatively higher than the group average compared to its specificity value and the group average.

Moving on to consider the data summarized in Figure 6.3, the width of the arrows corresponds with the proportion of responses given for each stimulus category. Thus, the proportion represented by the self-loop for each category is identical to the sensitivity value for that category. The other arrows pointing between categories do not directly correspond to any of the other measures we have considered until now. Rather, they help to reveal response patterns beyond accuracy. We will now consider the influence of introducing specific biases into the data simulation process.

Table 6.2 and Figure 6.4 show the results of randomly selecting a category response 70% of the time while 30% are correct responses. The overall accuracy has increased from the random choice accuracy to 48%. This accuracy is substantially above the 30% suggested by the description. This is because one quarter of the random responses happened to be correct by chance. The theoretical mean accuracy, then is actually 47.5%.

Sensitivity and specificity are fairly consistent across all categories, which is expected since the changes where applied uniformly to all categories. No values are beyond the range that would be considered most reasonable. The precision value follows the same relative pattern as with the purely random results—lower values for the two categories with fewer stimuli. However, within their same-stimuli-count groups, the values are consistent with what is expected in uniformly random data.

The detection prevalence gives a hint as to why sensitivity for categories A and B was relatively weaker than C and D. Both of the former categories were selected less frequently overall and below the theoretical mean. This reduced accuracy is reflected in the balanced accuracies as well. These patterns, of course, are not outside of what would be expected by chance, but they do represent a certain degree of inter-relatedness between some of the measures.

Figure 6.4 shows the response patterns for this data where responses are mostly random. The first thing to note is that even though most of the responses are guesses, the most frequent response category is the correct one. This pattern is initially surprising, but we can consider why the responses come out the way they do. Initially, the correct rate is 30%. Then an extra 17.5% correct responses are added due to chance. After accounting for these

Table 6.2: Simulation of 30% correct responses

Overall Accuracy	0.48			
	A	B	C	D
Sensitivity	0.46	0.47	0.50	0.49
Specificity	0.83	0.81	0.83	0.84
Precision	0.35	0.33	0.60	0.61
Detection Prevalence	0.22	0.23	0.28	0.27
Balanced Accuracy	0.64	0.64	0.66	0.66

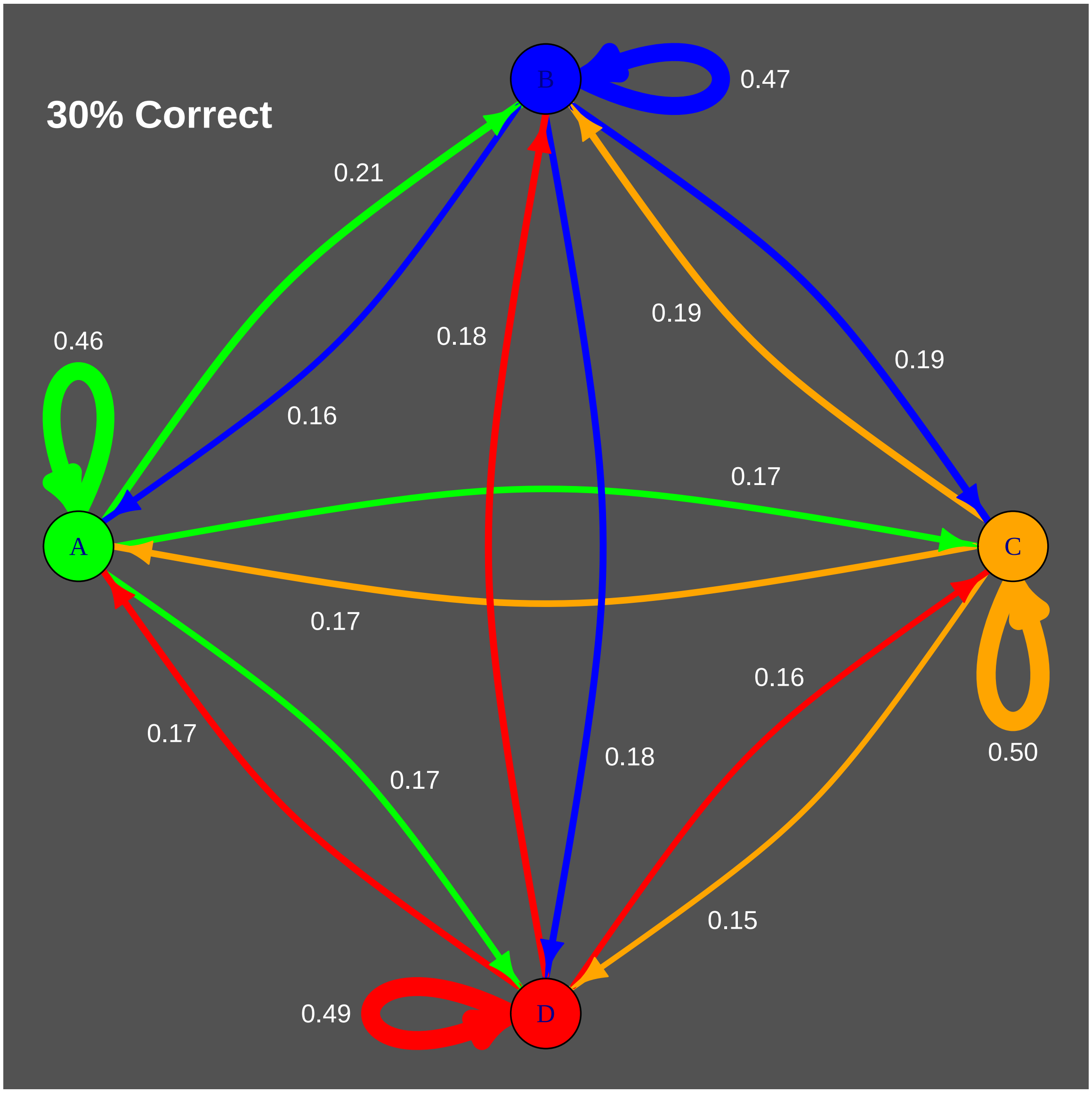

Figure 6.4: Simulation of 30% correct responses

47.5% correct responses, the remaining 52.5% of responses must be divided between the other three categories. Doing so yields a theoretical mean response per category of 17.5%

I will not describe the results presented in Table 6.3 and Figure 6.5 because they simply demonstrate a more extreme example of the previous results. This simulation shows the results of simulating 50% correct responses.

The single-category bias (or default-response bias) condition is shown in Table 6.4 and Figure 6.6. This simulation selects category C 30% of the time. The remaining responses are simulated at the 30% correct rate.

The overall accuracy is 43%. The sensitivity measures show that the unbiased categories are all below the overall accuracy while the biased category C is the only one above the overall accuracy. Category C has a corresponding lower specificity score since the higher response rate was due to a default bias toward C rather than a principled sensitivity to the category. Precision is harmed for category C because the increased response rate does not lead to a greater accuracy. Note, though, that the precision score for category C does not dip into the range of the categories with a lower stimulus prevalence. The bias for category C shows up clearly in the detection prevalence for the category, which is 30% greater than for the other categories. The balanced accuracies do not play a role in signaling the bias. The increase in sensitivity is balanced out by the decrease in specificity.

Figure 6.6 clearly shows the result of a default-response bias. The unbiased categories show a moderate specificity around 35%; however, the response rates for category C are much greater than for any other category, around 40%. Meanwhile, the other categories have a non-self response rate around 12%. Category C has the largest response rate for itself at 63%

The patterns observed when responses are selectively biased toward a single category are shown in Table 6.5 and Figure 6.7. In this simulation, when category B occurs, category

Table 6.3: Simulation of 50% correct responses

Overall Accuracy	0.64			
	A	B	C	D
Sensitivity	0.71	0.59	0.62	0.66
Specificity	0.87	0.91	0.88	0.87
Precision	0.53	0.55	0.72	0.71
Detection Prevalence	0.23	0.18	0.29	0.31
Balanced Accuracy	0.79	0.75	0.75	0.76

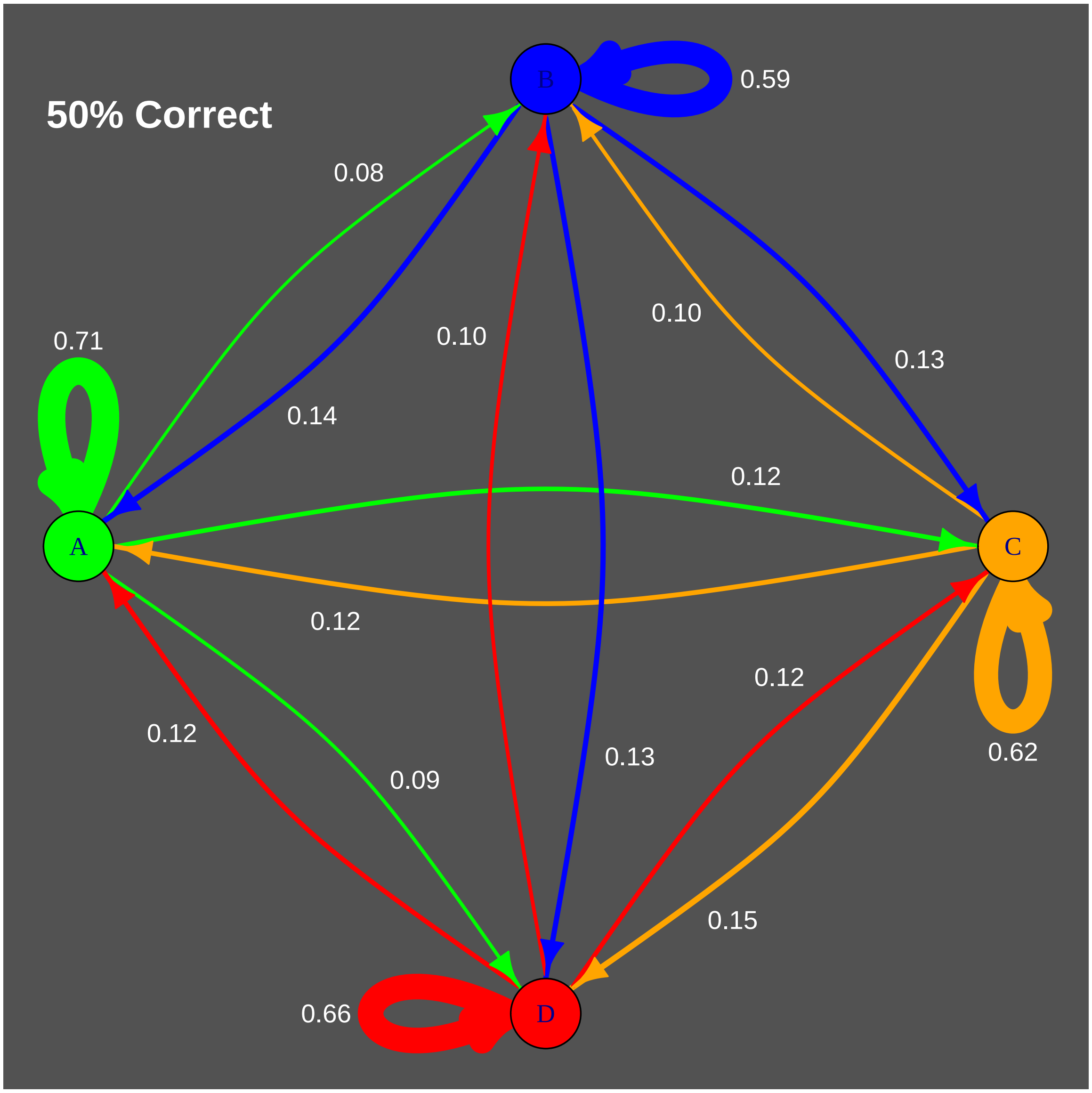

Figure 6.5: Simulation of 50% correct responses

Table 6.4: Simulation of selecting category C by default 30% of the time

Overall Accuracy	0.43			
	A	B	C	D
Sensitivity	0.36	0.36	0.63	0.32
Specificity	0.87	0.87	0.60	0.88
Precision	0.36	0.35	0.44	0.57
Detection Prevalence	0.17	0.17	0.47	0.19
Balanced Accuracy	0.61	0.61	0.61	0.60

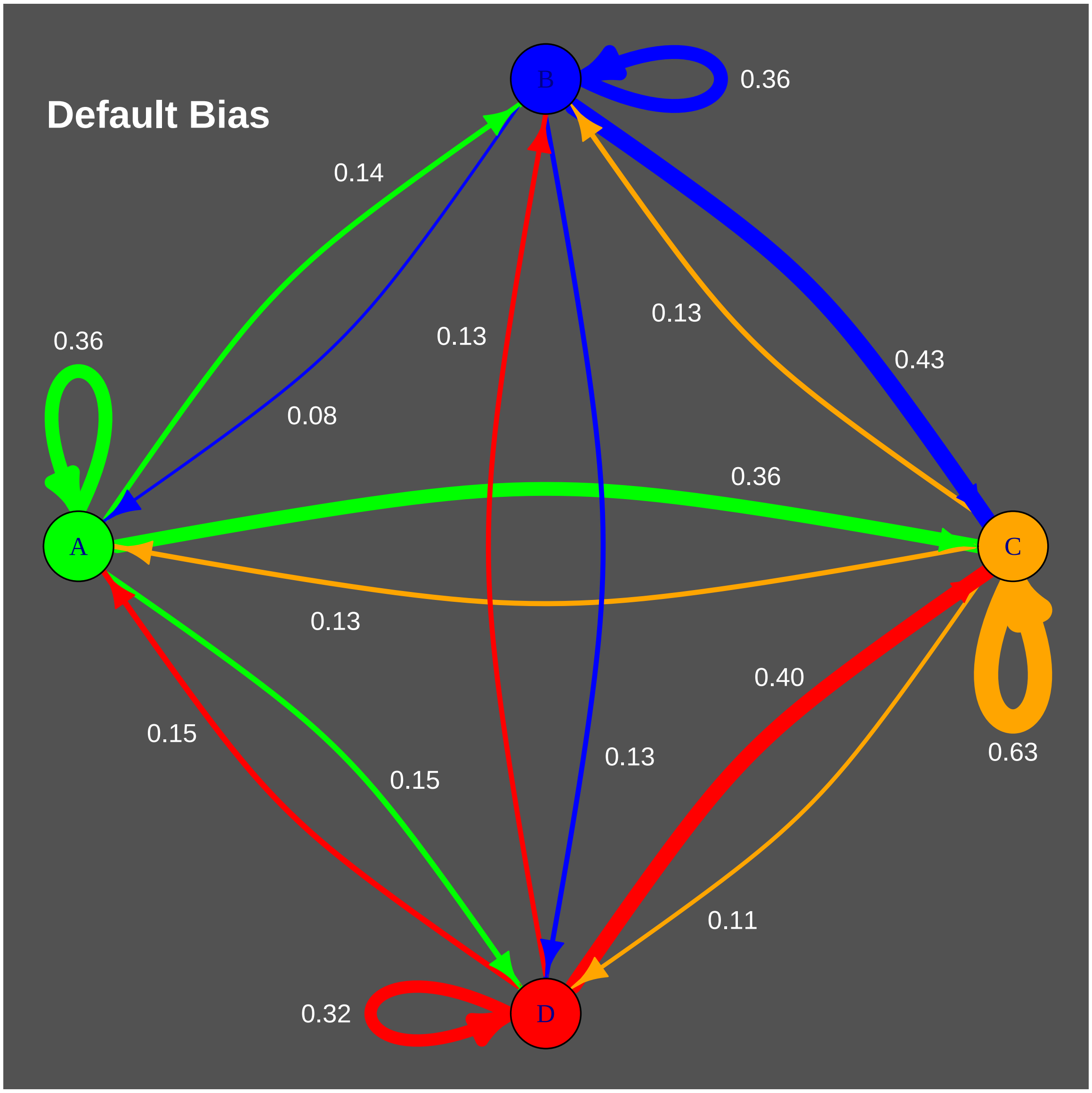

Figure 6.6: Simulation of selecting category C by default 30% of the time

B is selected 50% of the time. For the rest of the categories, the responses are 15% correct. These parameters were chosen because they approximate the response proportions found in the actual participant data.

Unsurprisingly, the sensitivity score for category B far exceeds those of the other categories. Specificity scores are uniform for all categories because the bias is a positive bias for the category rather than a negative bias against the category. The precision score for category B is above the neutral value for the low-prevalence categories because the responses for the category are based on a specific sensitivity rather than a default-response bias. The detection prevalence for category B is greater than the other categories since the other response options are ruled out when B occurs. The overall increased sensitivity for category B is reflected in its relatively higher balanced accuracy.

The selectively-sensitivity bias pattern is easily discernible in Figure 6.7. The self-loop for B is much larger than the other categories at 73%. Accordingly, responses for other categories are less frequent when B occurs. The other categories show response patterns similar to what we have observed in previous random simulations.

When two categories are easily confusable the responses may look like those shown in Table 6.6. Here, when the stimulus is from either category A or B, 50% of the time the response is randomly selected from among A and B. The remaining responses are correct 15% of the time. This simulation also implicitly includes a selective-sensitivity bias for both of these categories.

The overall accuracy is the same as with the previous two simulations. We can see a slightly increased sensitivity for categories A and B, but the increase is weak because the random selection dilutes the effect. Specificity scores are lower for the biased categories than the unbiased ones as are the precision scores, where the baseline is around 35%. The

Table 6.5: Simulation of choosing category B correctly 50% of the time

Overall Accuracy	0.43			
	A	B	C	D
Sensitivity	0.36	0.73	0.39	0.37
Specificity	0.81	0.79	0.82	0.84
Precision	0.28	0.41	0.52	0.54
Detection Prevalence	0.22	0.30	0.25	0.23
Balanced Accuracy	0.59	0.76	0.61	0.60

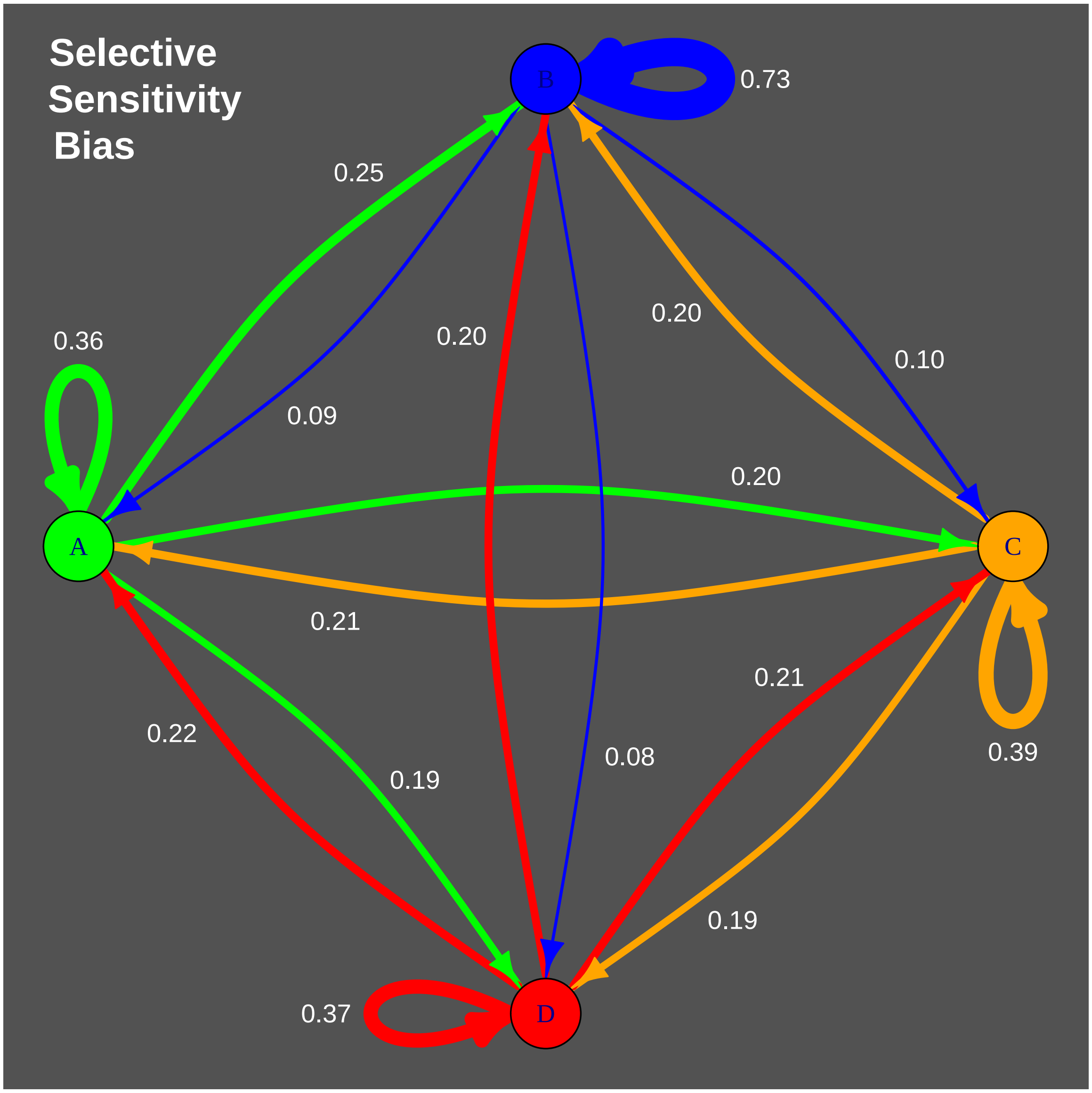

Figure 6.7: Simulation of choosing category B correctly 50% of the time

detection prevalence scores only subtly indicate the specific-response bias to A and B. The balanced accuracies are uniform for all categories.

This particular pattern of confusable categories is difficult to observe in the standard measures because the pattern is found in the kinds of errors that occur. While the sensitivities for A and B individually are poor, the sensitivity for both as a combined category is quite good. If A and B were combined, the accuracy in this simulation would be around 62.5%, far greater than the other categories.

Figure 6.8 summarizes the response patterns unique to a two-category confusion. While the self-response loops for A and B are the same as C and D, an A stimulus is much more likely to receive a B response than any other category, likewise for B to A. This pattern of fewer responses for other categories is due to the selective-sensitivity bias also seen in Figure 6.7 with the modification that there is a strong connection with one other category instead of a stronger self-loop.

It may be that two categories are easily confusable, but one category label is more emblematic of the two. For example, both hamburgers and salads are foods, but signs are hung in public spaces to indicate that eating of food is prohibited, the signs often show a hamburger with a prohibited slash through it. Eating salads is still prohibited, but hamburgers are more emblematic of food and so serve to represent salads, too. This simulation shows the results of a two-category confusion bias with a default response bias. In this simulation in 6.7 and 6.9, when B is the stimulus, 50% of the time the response is A. The rest of the responses are correct 15% of the time.

The overall accuracy is lower at 35% than the previous simulation because only the responses for category B are modified, the A responses are treated the same as the other categories. As this is the case, the sensitivities for A, C, and D are the same while the score for B is substantially lower. The specificity for A, however, is relatively low because of the

Table 6.6: Simulation of randomly selecting either category A or B 50% of the time that either A or B occur

Overall Accuracy	0.40			
	A	B	C	D
Sensitivity	0.42	0.45	0.40	0.35
Specificity	0.76	0.76	0.84	0.85
Precision	0.26	0.27	0.57	0.55
Detection Prevalence	0.27	0.27	0.24	0.21
Balanced Accuracy	0.59	0.60	0.62	0.60

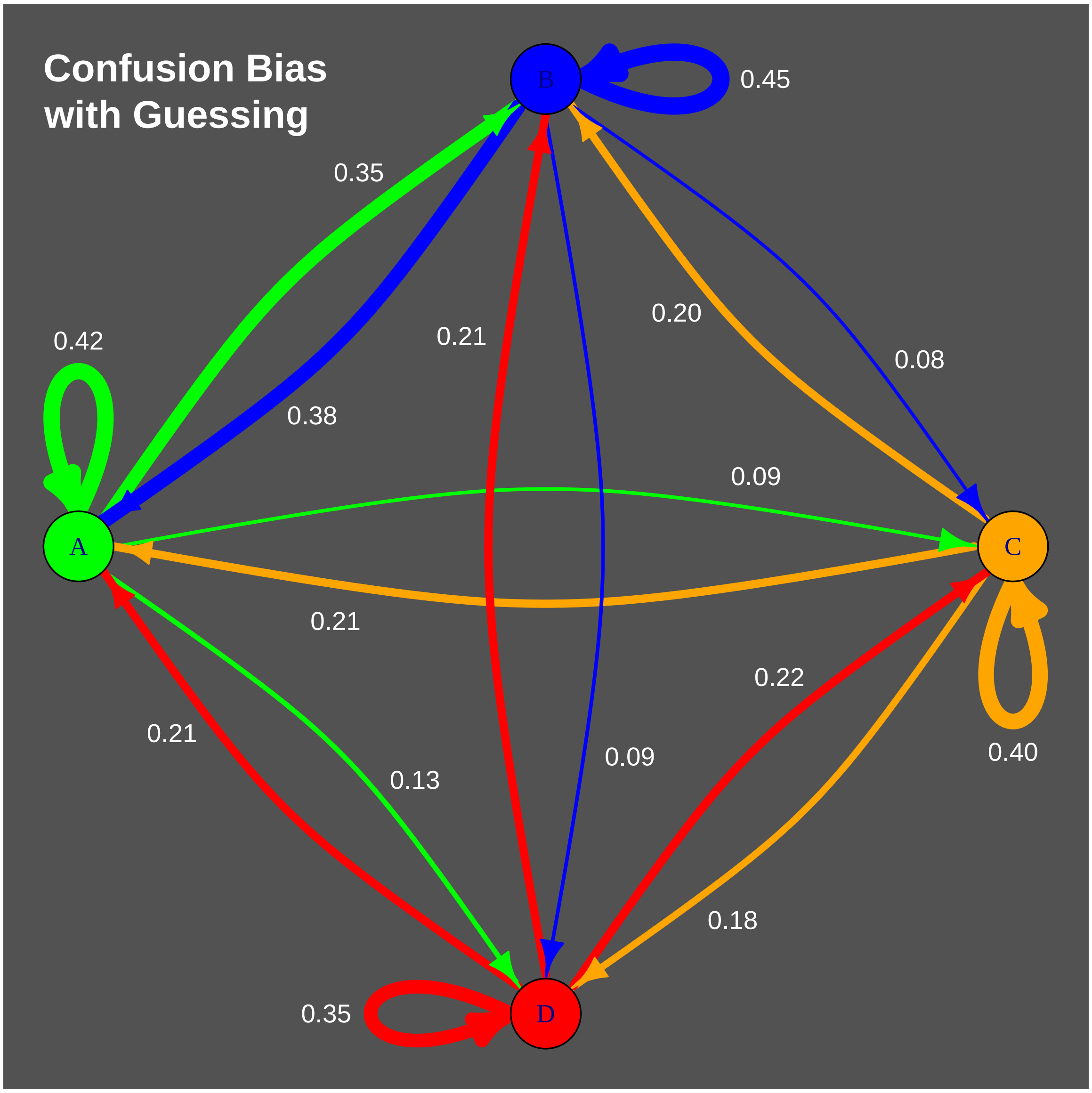

Figure 6.8: Simulation of randomly selecting either category A or B 50% of the time that either A or B occur

large numbers of false alarms due to B being classified as A. Precision for both A and B is lower than the predicted baseline for two different reasons. Precision for A is low for the same reason that its specificity is low. Precision for B is low because the responses that would normally point back to itself are given to A, so the responses for B stimuli occur at the same rate as those due to chance. The response bias for category A is shown by its 31% detection prevalence while the response bias against category B is shown by its 21% detection prevalence. Finally, the balanced accuracies for both categories A and B are lower than the unbiased categories C and D due to the confusion between A and B.

As with the previous simulation, the two-category confusion bias with a default response is difficult to see in the standard measures, but it is quite clear in Figure 6.9. In this figure, there is a broad line connecting B to A representing a 62% response frequency. As so many responses are given to A, there are few remaining for B's self-loop or the other two categories. In this regard, it also appears like the pattern in the selective-sensitivity bias condition depicted in Figure 6.7. Because none of the other categories were affected by the bias, including A, all of the other responses patterns are uniform (the 25% A to B response is due to chance).

Having seen some idealized examples of patterns that may logically be found in data with this structure, we will now consider the results from the actual experiment participants.

Table 6.7: Simulation of selecting category A 50% of the time that B occurs

Overall Accuracy	0.35			
	A	B	C	D
Sensitivity	0.36	0.20	0.39	0.37
Specificity	0.70	0.79	0.82	0.84
Precision	0.20	0.16	0.52	0.54
Detection Prevalence	0.31	0.21	0.25	0.23
Balanced Accuracy	0.53	0.49	0.61	0.60

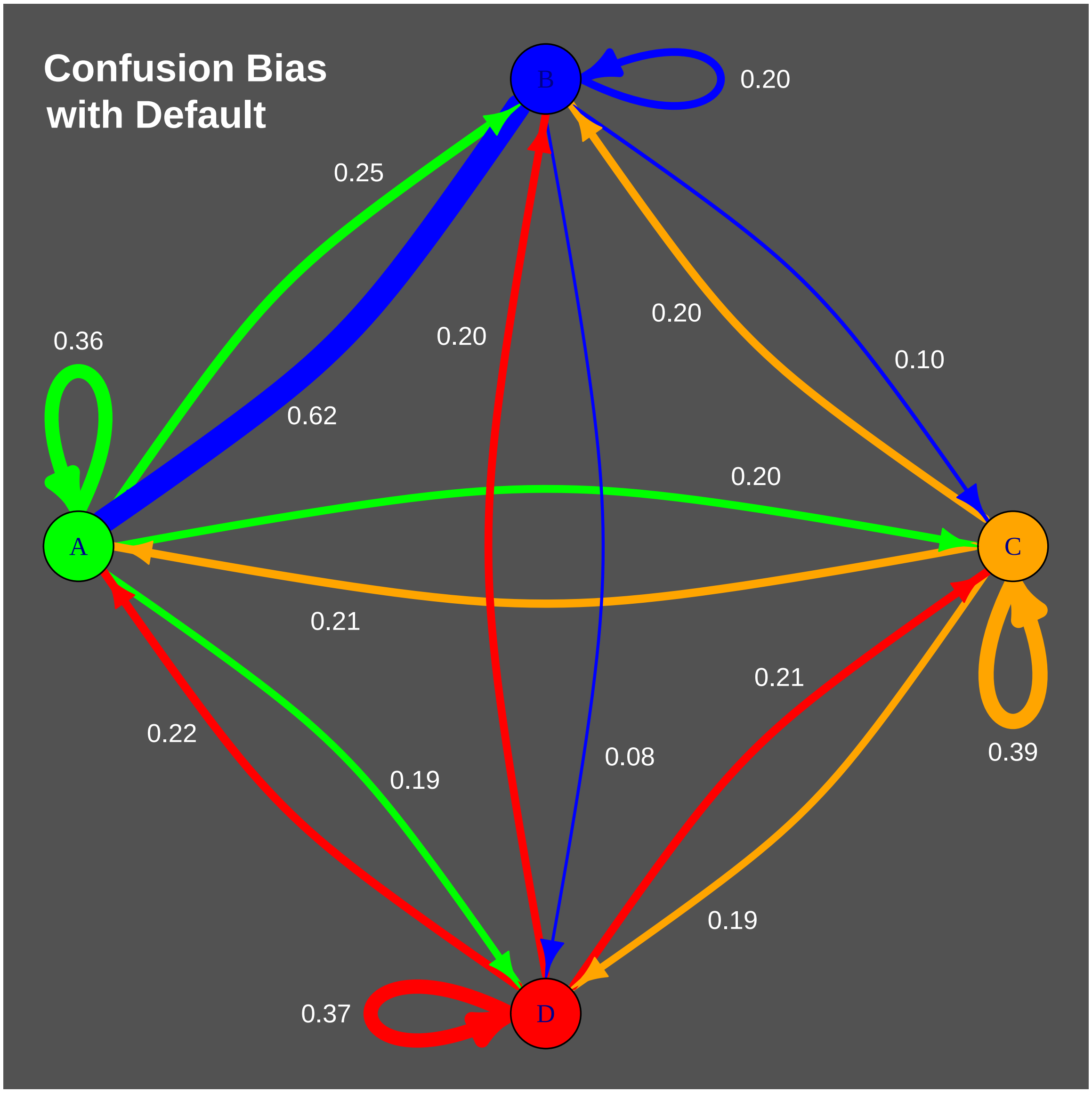

Figure 6.9: Simulation of selecting category A 50% of the time that B occurs

6.3 Results

This section presents the results of the four-alternative forced-choice categorization task.

6.3.1 Categorization Accuracy

The mean proportions of correct responses by each participant group are presented in Table 6.8. For participants from northwest Indiana, the results for all participants are presented as well as results for the subsets of participants who self-identified as "Black" and "White". There was not enough racial diversity among participants from the other regions to justify analysis by subsets. Chance performance in a four-alternative task is 25%. While the mean accuracies for the groups are not much above chance, most of the groups are statistically above chance according to a binomial test ($p < 0.05$). One group, NW-Black, was not significantly above chance; however, this can be attributed to the small size of the group (n = 5). The NW-Black group has a higher accuracy and narrower standard deviation than the NW-White group, whose accuracy rate was found to be above chance.

Table 6.8: Mean percent of correct responses (accuracy) for each group with standard deviation in parentheses

	Mean **Percent Correct**
NW-All	27.6 (5.4)
NW-Black	27.9 (5.3)
NW-White	27.7 (5.6)
Northeast	30.5 (5.3)
Central	28.9 (4.5)
South	30.7 (5.6)
Total	29.0 (5.4)

A one-way between subjects ANOVA with talker dialect (Inland, North, Midland, South) as the within-subject variable and listener group (northwest, northeast, central, south) as the between-subject variable revealed a significant main effect of talker dialect [$F(3, 297)$ =

28.5, $p < 0.001$] and a significant talker dialect by listener group interaction [F(9, 297) = 5.5, $p < 0.001$].

Figures 6.10 and 6.11 show the percent correct categorization for each of the four talker dialect groups. Pairwise post-hoc Tukey tests on talker dialect show the locus of the interactions and main effect of dialect. Overall, performance was best for Midlands talkers and worst for Inland talkers. In general, participants performed best on talkers from their own regions. Participants from northwest Indiana tended to perform better on Inland talkers (those from their own region). This interaction was only significant ($p < 0.05$) between the northeast and northwest participants, though.

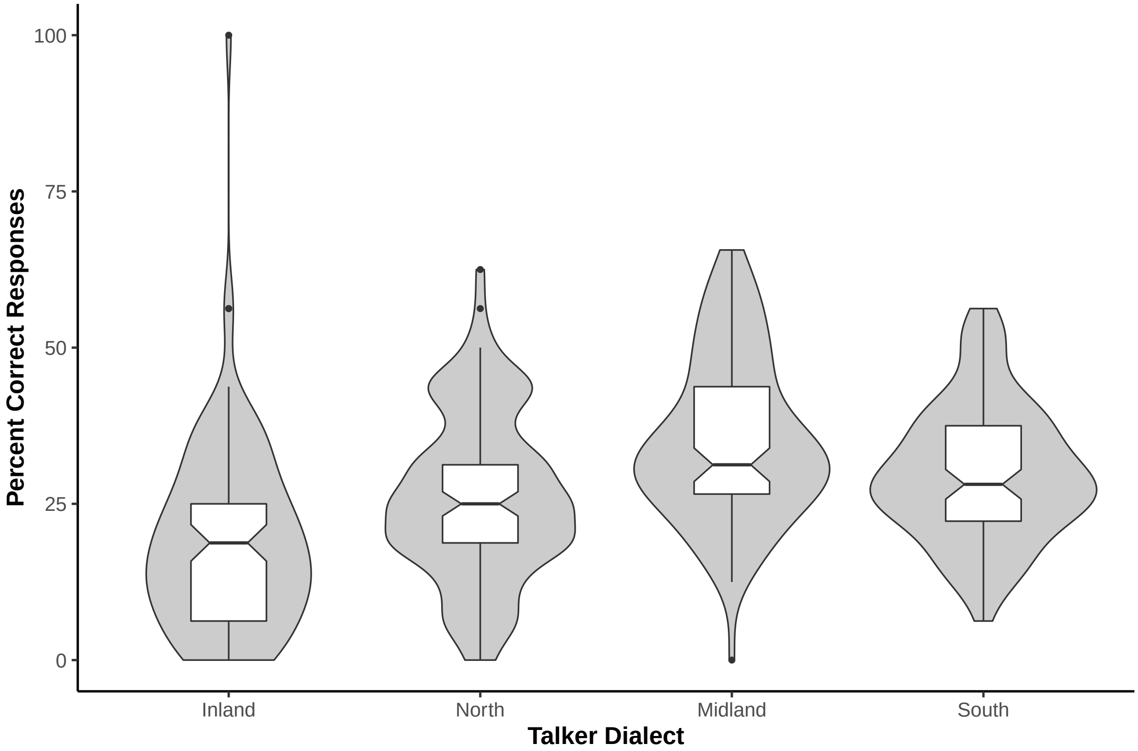

Figure 6.10: Percent correct categorization for each of the four talker dialect groups. The violin plot represents the relative density of participants along the correct response rate scale. The boxes span the interquartile range with the middle line showing the median. The whiskers extend no more than 1.5 times the interquartile range. Notches in the boxes span the 95% confidence interval for the median.

Table 6.9 presents the mean percent correct categorization for the listener groups for each of the talker dialects. In addition to the data depicted in Figure 6.11, this table separates

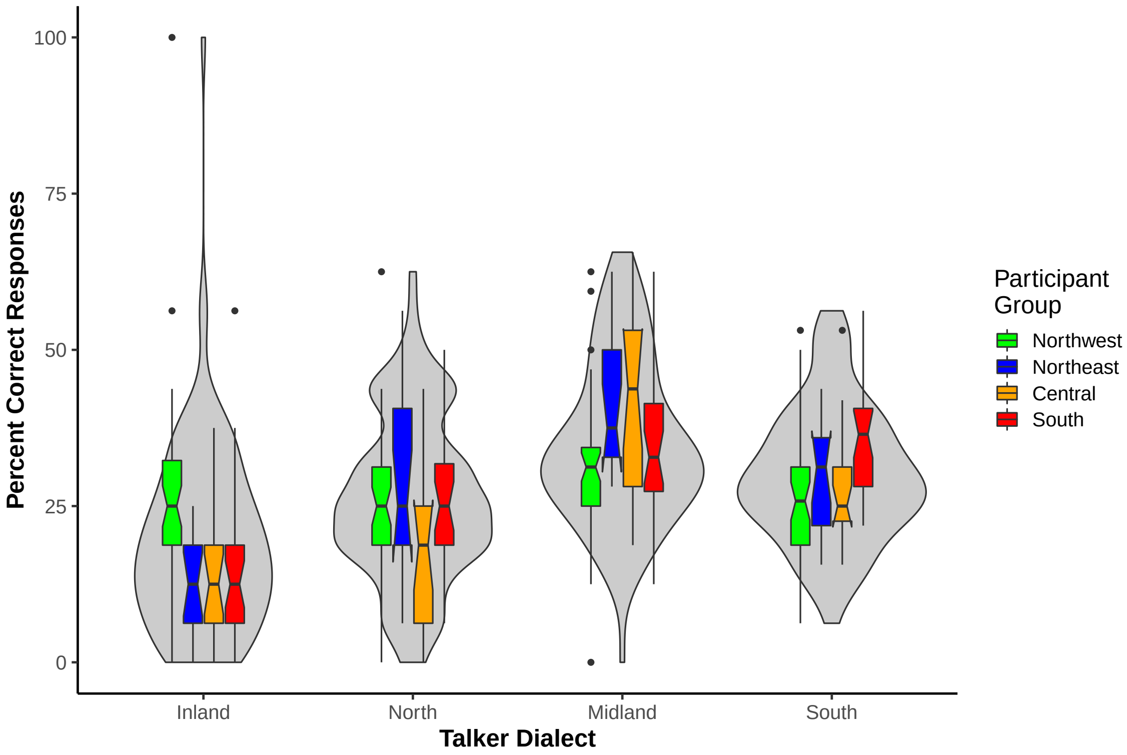

Figure 6.11: Percent correct categorization by each participant group for each of the four talker dialect groups. The violin plot represents the relative density of all participants along the correct response rate scale. The boxes summarize the response of the participant groups and span the interquartile range with the middle line showing the median. The whiskers extend no more than 1.5 times the interquartile range. Notches in the boxes span the 95% confidence interval for the median. The notches resemble spikes when they exceed the interquartile range.

the responses from black and white northwest listeners. In addition to the general patterns presented already, black northwesterners performed better on Midlands talkers and worse on Southern talkers compared to the white listeners.

6.3.2 Response Bias

Independent of the accuracy of categorizations, participants may differ in the rates at which they select categories. Table 6.10 shows the responses biases for each listener group on the responses. A bias rate of 0.25 indicates a relatively unbiased response since there are four possible response categories.

Table 6.9: Percent correct categorization performance for each listener group for each talker dialect

Listener Group	Inland	North	Midland	South
NW-Black	25	29	38	19
NW-White	26	26	29	28
Ft Wayne	12	29	42	30
Indianapolis	15	18	42	28
South	15	27	35	36
Mean	19	26	37	28

Overall, participants were most likely to respond the Midland category, with only the northwest participants responding North as frequently as Midland. The second most frequent response category for each participant group tended to be their own region. Both northwestern groups, however, favored responding North. In fact, North was responded more frequently than Inland by all participant groups. This is unexpected since the northwest corner of Indiana is a highly salient dialect region according to the results of the draw-a-map task presented in Chapter 4. On average, the North and South category responses were relatively unbiased; however, biases increased for participants closer to the response categories (e.g., groups responded South more if they lived closer to the South region and vice-versa for North).

Table 6.10: Response biases for each listener group on the response categories. Bias rates close to 0.25 indicate relatively unbiased responses.

Participant Group	Inland	North	Midland	South
NW-B	0.19	0.32	0.31	0.17
NW-W	0.23	0.28	0.28	0.21
Northeast	0.12	0.28	0.37	0.23
Central	0.13	0.24	0.39	0.25
South	0.12	0.24	0.33	0.31
Mean	0.16	0.26	0.33	0.24

6.3.3 Perceptual Response Patterns

This section will consider the patterns of perceptual categorization responses. It uses the methods described in Section 6.2.6. The previous two sections have considered the accuracy of responses and the response rate for each category. We now consider patterns of responses, especially focusing on the kinds of response errors.

The data presented here come from the confusion matrix of correct categories and response classifications. The raw confusion matrices can be found in an Appendix at the end of this chapter. As a reminder, *sensitivity* is the accuracy on each talker category (comparable to the hit rate), *specificity* is the correct rejection rate, *precision* is how often the response is correct when a category is chosen, *detection prevalence* is the response bias on each response category, and the *balanced accuracy* is the average of *sensitivity* and *specificity*.

The network connection graphs summarize the response proportions in the confusion matrix. The nodes represent the category of the talker while the edges, or lines, represent the response category. The width of the edges correspond to the proportion of each response type, which is also depicted by the number near the target response category end of the arrow. For example, if an edge points from the Inland talker category ("I") to the North node ("N") with a value of 0.33, when the participants heard a talker from the Inland region, they selected the North category 33% of the time. Similarly, if the edge circles back to point at the Inland category with a value of 0.19, the correct response was selected 19% of the time.

Table 6.11 and Figure 6.12 summarize the results for all participants. We have already seen the overall accuracy, sensitivity, and detection prevalence in the previous sections. Participants were most accurate on the Midland category and least on the Inland category. The specificity for Midland, however, is rather low at 0.68, likely as a result of it being the default response category. Comparing detection prevalence to precision gives a sense of relative accuracy. When participants chose the Midland category, they were correct 36% of the time, but they selected the category 33% of the time. A similar pattern is shown for the Inland category with only a slightly greater precision (19%) than detection prevalence (16%). By comparison, participants were somewhat less accurate when presented with a talker from the North region with precision ten percent lower than detection prevalence. They were relatively more accurate when hearing a talker from the South region with a precision of 41% compared to a 24% detection prevalence. Considering the balanced accuracy, performance was fairly poor overall; all values are near chance level.

The directed network graph in Figure 6.12 summarizes the confusion patterns. Working clockwise starting with the Inland category, we can see that when participants heard an Inland talker they responded North or Midland a large majority (67%) of the time. It resembles the pattern of a confusion bias with a default response simulation shown in Figure 6.9 except in this case there are two default responses that are randomly selected. It appears that the Inland category is not being used in a reliable manner, but it is selected randomly. Notice that responses for Inland range from 15% to 20%, and the correct response rate is squarely within that range.

When talkers from the North region were heard, responses appear mostly random with a slight bias toward a Midland response and bias against an Inland response.

The Midland talkers were most often correctly categorized. The remaining responses show a random selection between North and South and the now-familiar bias against an Inland response. The Midland category behaves like the default category since it is selected more frequently than any other category including correct responses.

The South talkers seem to be the most reliably categorized with a relatively high sensitivity rate and confusions that reduce in frequency from Midland to Inland.

Overall, the confusions follow some fairly neat patterns. Inland is reliably confused for either North or Midland, not South. North and South are both confused most often with Midland, and Midland is equally confusable with both North and South. Such a pattern reflects that fact that the Midland lies geographically between North and South.

For the remainder of the graphs depicting responses from the participant groups, I will only discuss patterns that differ in notable ways from the aggregation of all participants' responses.

As seen earlier, the northwest listeners have the lowest accuracy and category sensitivities overall. Table 6.12 shows that this participant group has a precision advantage for the South talkers compared to other groups. Even though they only responded South 20% of the time, they were correct 44% of the time on that category.

Figure 6.13 shows that the northwest listeners regularly confuse Inland talkers for Northern and Midland talkers and very rarely categorize them as Southern. There is also a Midland-North confusion with Midland and North getting categorized as each other as often as themselves, 27% and 29% of the time. There is an interesting confusion in which Southern talkers are categorized as Northern 28% of the time.

Tables 6.13 and 6.14 and Figures 6.14 and 6.15 show the classification patterns of northwestern participants divided by race. Black and white northwesterners have equal accuracies overall; however, they show different categorization performance. Black listeners show a strong bias to classify Inland talkers as Northern (45%) and almost never as Southern (5%). While they have a low sensitivity for Southern talkers (19%), they have a relatively high precision for the category (37%). There is a notable balanced confusion between North and South talkers (31% and 34%). While the sensitivity for Midland is not the highest among all participant groups, the precision for the category is the highest at 40%. Overall the black northwestern participants' responses can be summarized as having a confusion between

Table 6.11: Summary of response measures for all participants

Overall Accuracy	0.29			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.19	0.26	0.35	0.30
Specificity	0.84	0.74	0.68	0.78
Precision	0.19	0.16	0.36	0.41
Detection Prevalence	0.16	0.26	0.33	0.24
Balanced Accuracy	0.51	0.50	0.52	0.54

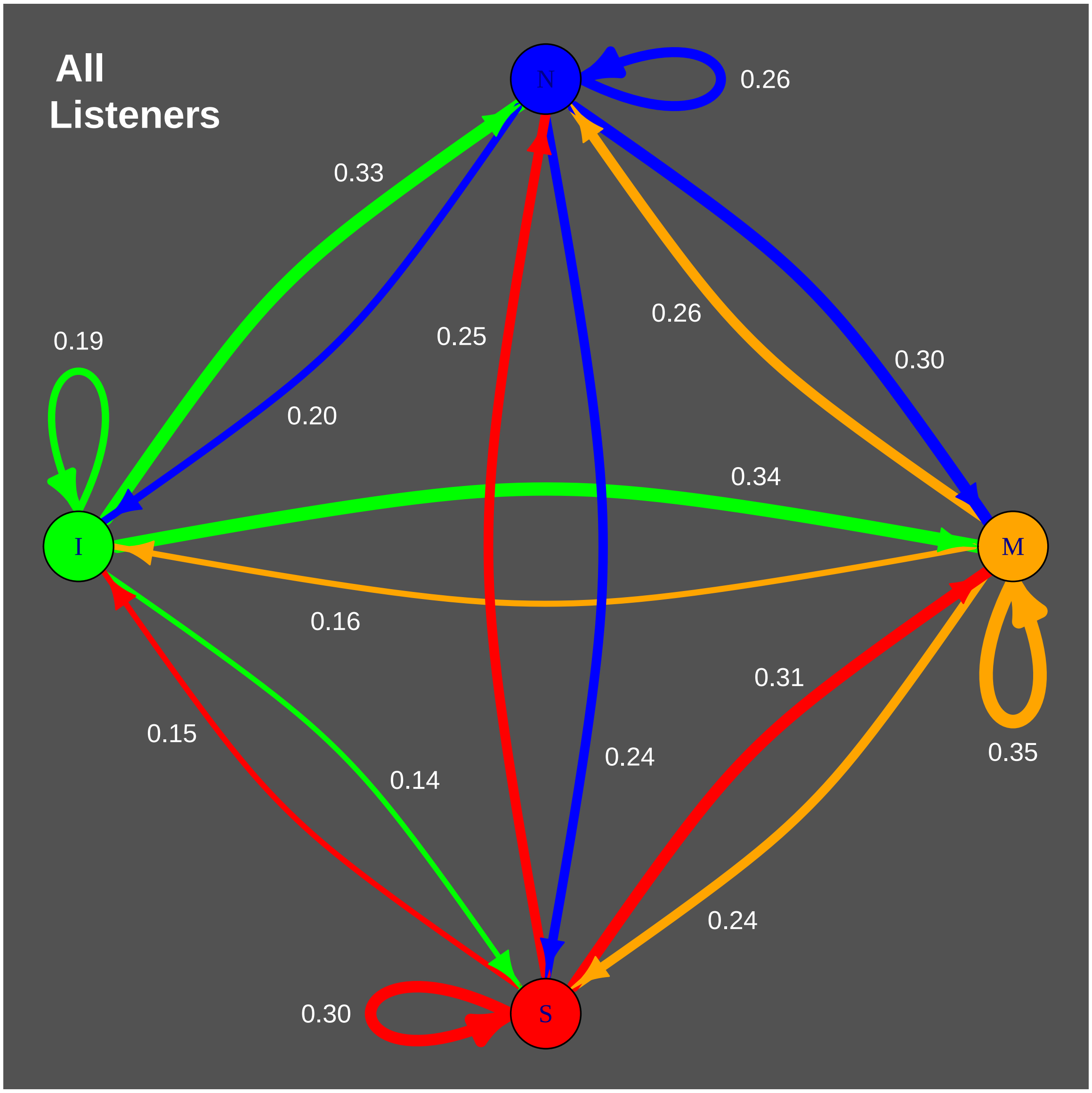

Figure 6.12: Response network graph for all participants

Table 6.12: Summary of response measures for all northwestern participants

Overall Accuracy	0.27			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.26	0.27	0.30	0.27
Specificity	0.78	0.71	0.72	0.83
Precision	0.19	0.16	0.35	0.44
Detection Prevalence	0.23	0.29	0.28	0.20
Balanced Accuracy	0.52	0.49	0.51	0.55

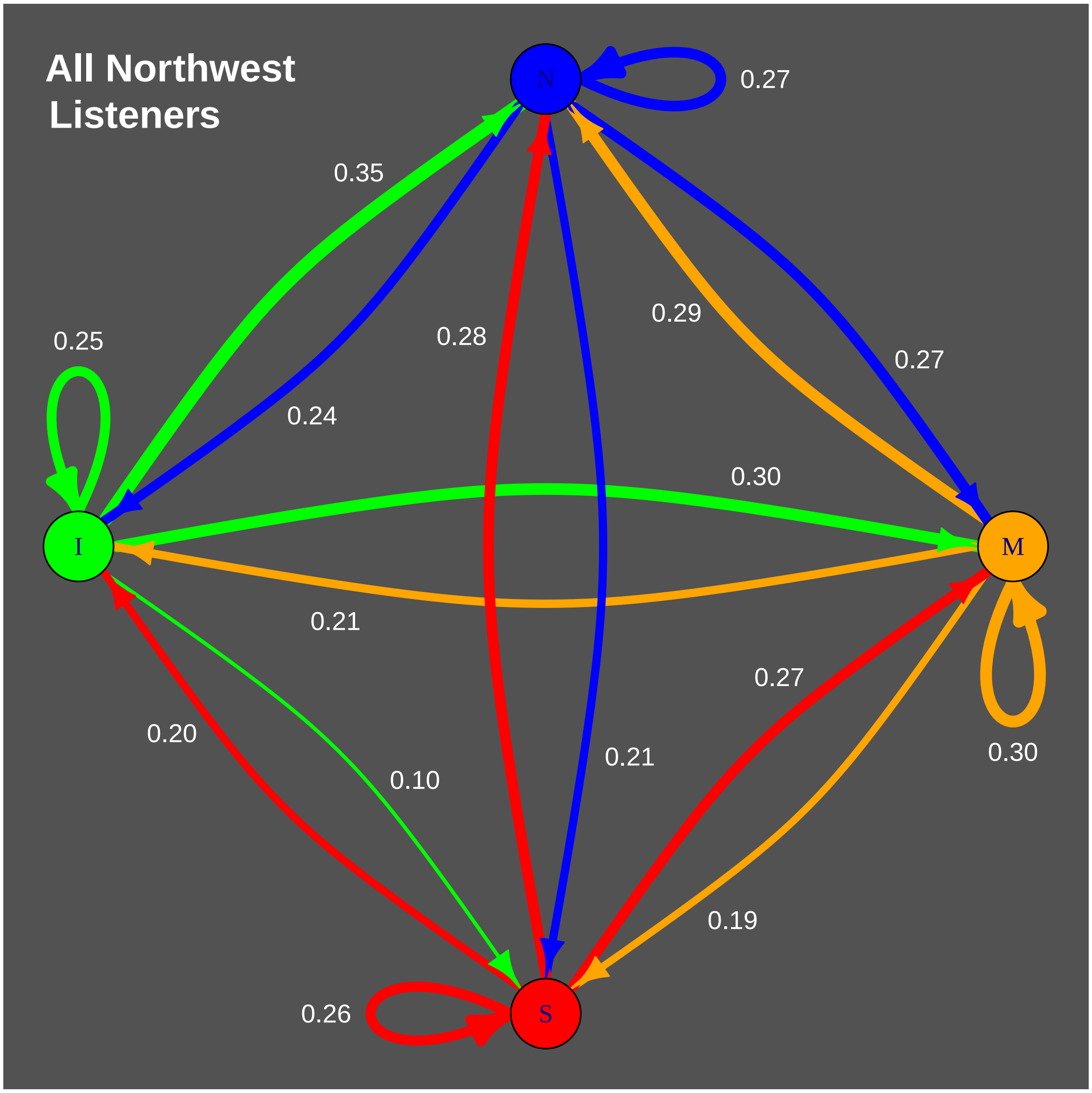

Figure 6.13: Response network graph for all northwestern participants

Inland and North with a default bias for North, a confusion between North and Midland with guessing, a general default response bias for Midland, and a confusion between North and South with guessing.

The white northwest participants generally have sensitivities and detection prevalences that most resemble random guessing, as seen in Table 6.14. Precision for Southern talkers is the highest among all participant groups. Figure 6.15 shows a three-way confusion between Inland, North, and Midland talkers with a response bias favoring North and Midland. The balanced North-South confusion found among the black northwesterners is not replicated among the white participants.

The northeast participants have the second highest overall accuracy among the participant groups. Their sensitivities, shown in Table 6.15, are each some of the highest of all participant groups, too. The low specificity for the Midland category (65%) suggests this is the default response category. Figure 6.16 shows that the Midland response bias is mostly due to a confusion with Inland and South. Inland talkers are almost always categorized as North or Midland. This pattern seems to be the result of a general default bias toward Midland, a two-way confusion between Inland and both North and Midland with responses defaulting almost equally toward the latter two, and selective sensitivities for South and North.

Table 6.16 shows that participants from central Indiana have a high overall accuracy (29%) but also the highest detection prevalence for their own dialect region (39% for Midland). At 35% the central participants have the lowest precision rate of any of the participant groups for the Midland category. Figure 6.17 confirms that there is a general response bias for Midland and that the correct response rate for Midland is approximately as high as the response bias from the other three categories. Midland talkers are classified as either North or South a quarter of the time. While southern talkers are most frequently

Table 6.13: Summary of response measures for black northwestern participants

Overall Accuracy	0.28			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.25	0.29	0.38	0.19
Specificity	0.82	0.67	0.72	0.84
Precision	0.22	0.15	0.40	0.37
Detection Prevalence	0.19	0.32	0.31	0.17
Balanced Accuracy	0.53	0.48	0.55	0.52

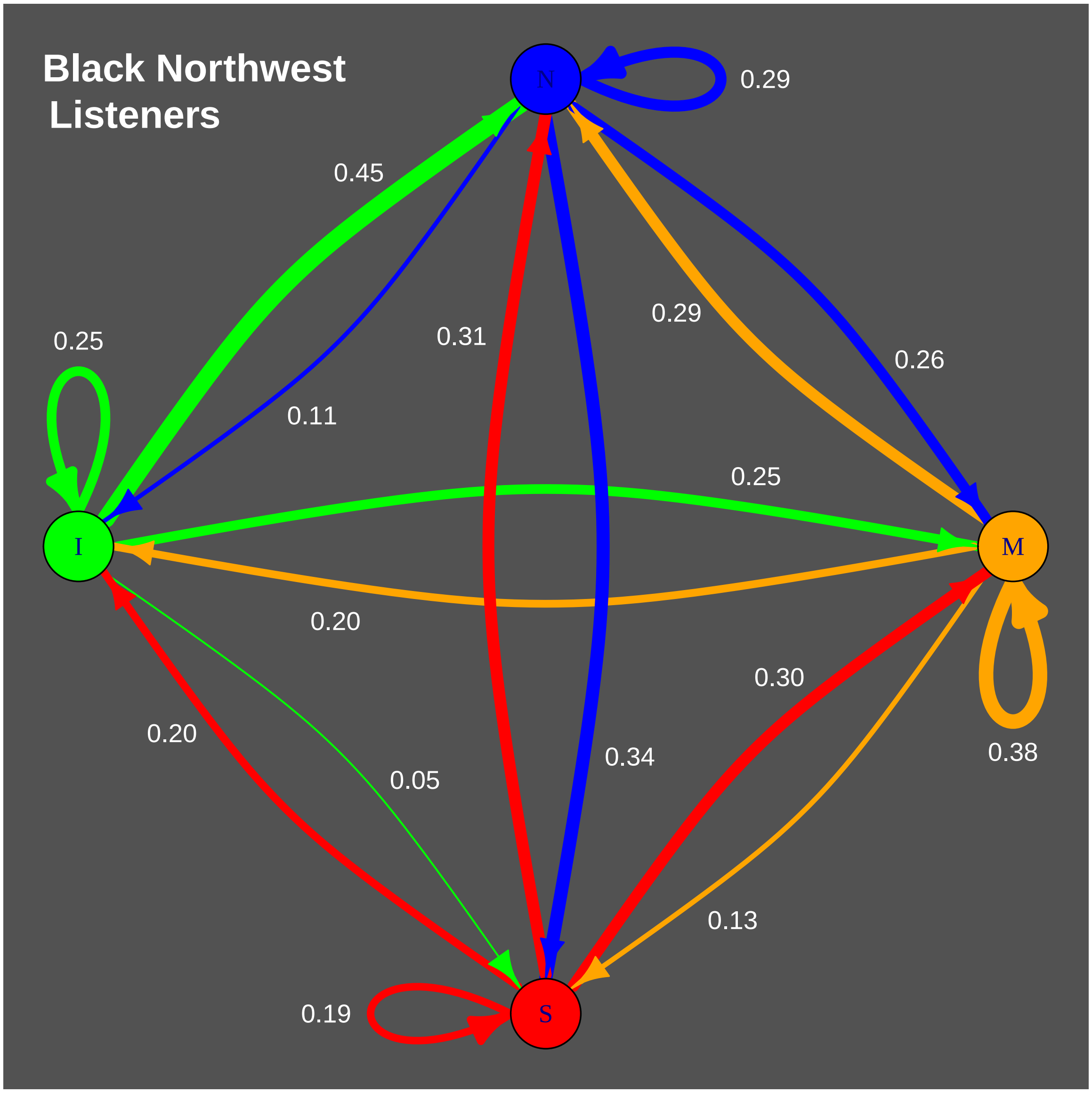

Figure 6.14: Response network graph for black northwestern participants

Table 6.14: Summary of response measures for white northwestern participants

Overall Accuracy	0.28			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.26	0.26	0.29	0.28
Specificity	0.77	0.72	0.73	0.83
Precision	0.19	0.16	0.35	0.45
Detection Prevalence	0.23	0.28	0.28	0.21
Balanced Accuracy	0.52	0.49	0.51	0.55

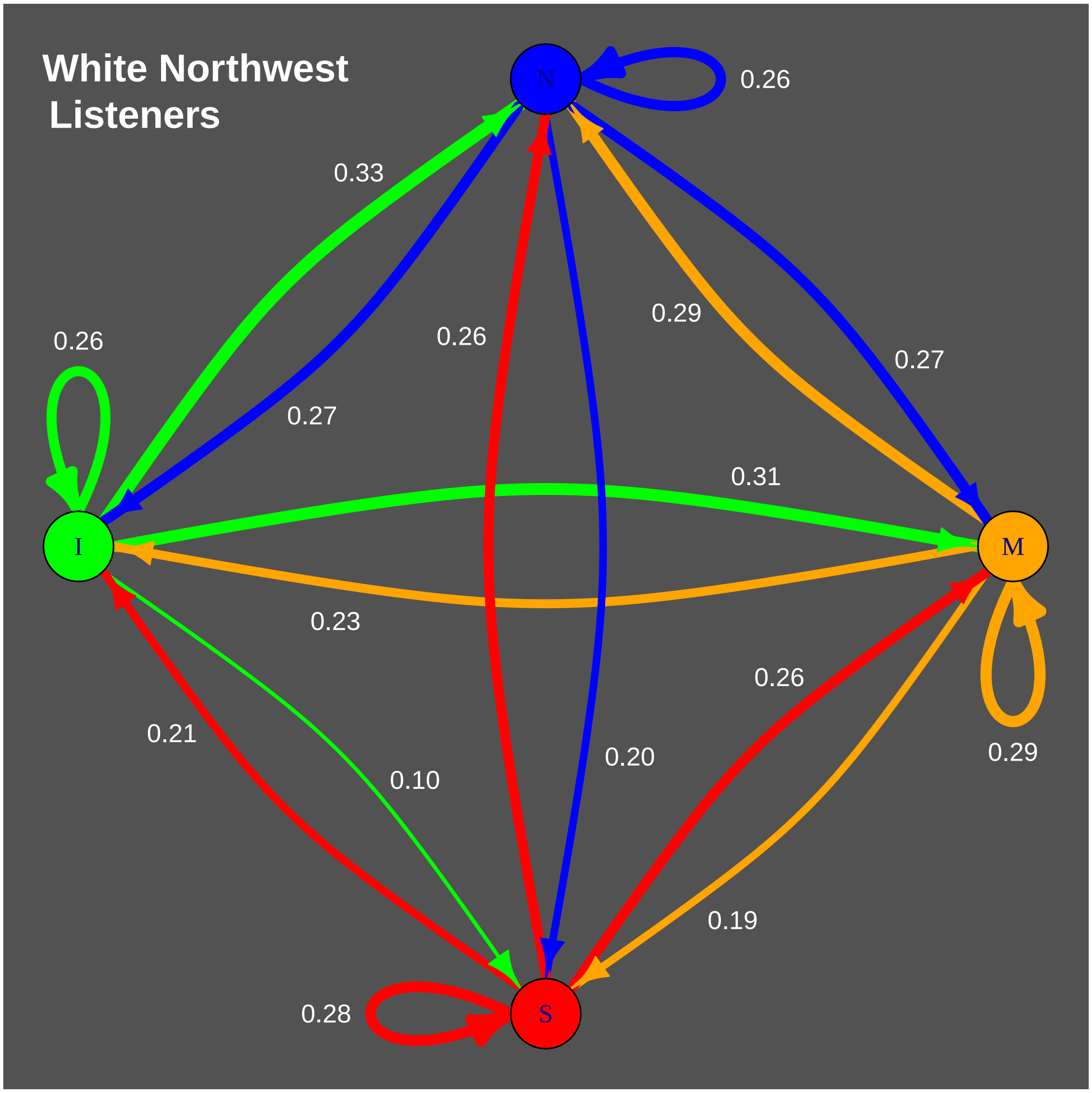

Figure 6.15: Response network graph for white northwestern participants

Table 6.15: Summary of response measures for northeastern participants

Overall Accuracy	0.30			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.12	0.29	0.42	0.30
Specificity	0.87	0.72	0.65	0.81
Precision	0.16	0.17	0.37	0.44
Detection Prevalence	0.12	0.28	0.37	0.23
Balanced Accuracy	0.49	0.51	0.53	0.55

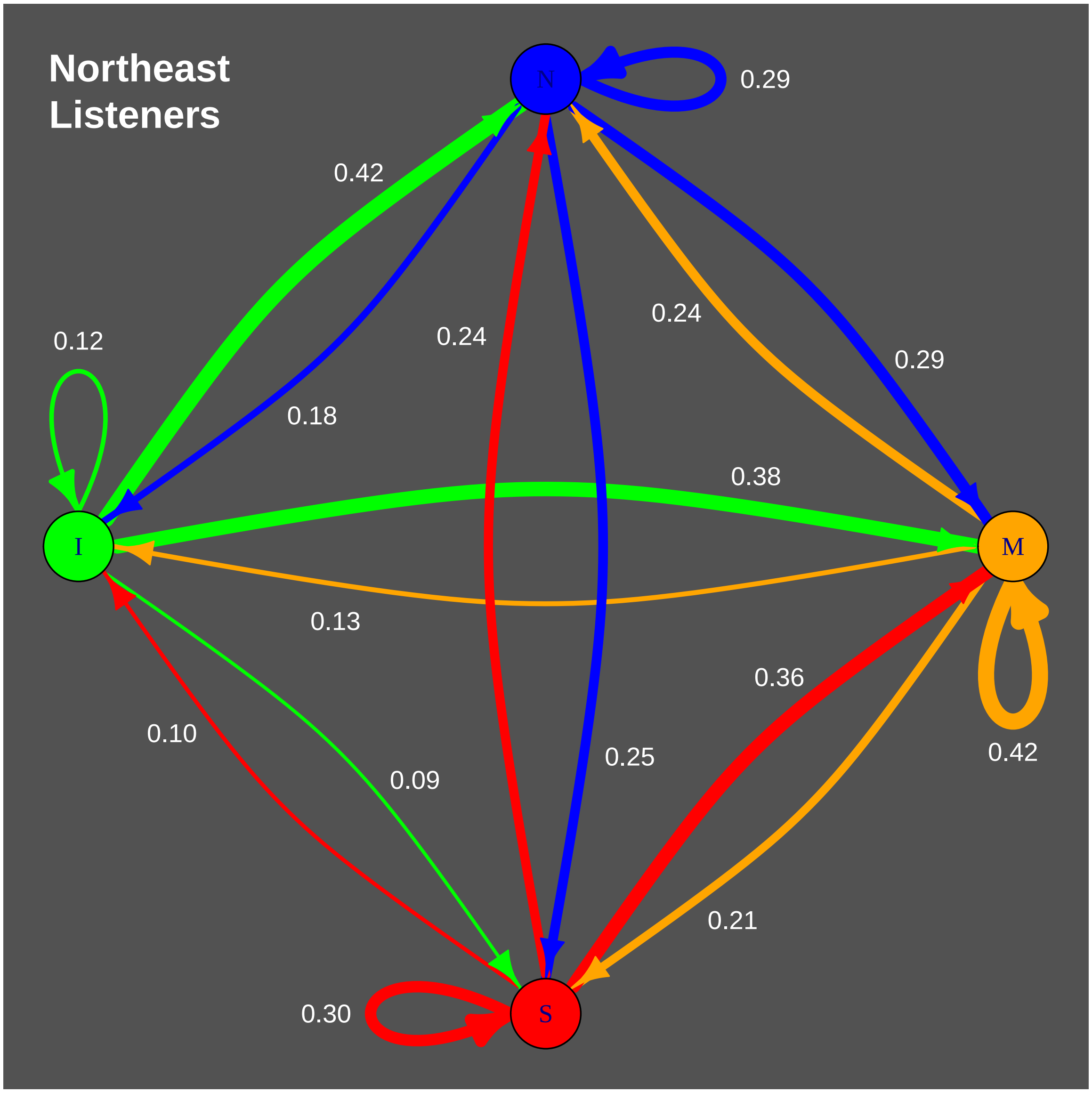

Figure 6.16: Response network graph for northeastern participants

labeled as Midland, North talkers are classified as South rather frequently (28%). These results mostly resemble responses with a default bias for a single category, Midland. In addition there is a confusion between Inland and North with a default bias toward the North category. There also appears to be three-way confusion between North, South, and Midland with guessing.

The southern listeners have the highest overall accuracy (31%) as seen in Table 6.17. Sensitivities for both Midland and South are well above chance and are also over the no information rate. Precision for both Inland and South categories are relatively high compared to their detection prevalence (21% and 39%, respectively), indicating that participants tend to be correct when they do select those categories. Figure 6.18 suggests that there is only a weak default response bias for Midland. Midland is evenly confused with South and unevely confused with North. Inland talkers are classified as each of three other categories and have the highest categorization as South (23%) of any of the participant groups. South and North are also fairly evenly confused. These results can be the result of two default-category response biases, Midland and South; a confusion between North and South with guessing; and a three-way confusion between Inland, North, and Midland with default response biases toward the latter two.

6.3.4 Effect of Phonological Content

This section considers the role of the phonological content of the sentences in the regional categorization. Of the 16 sentences participants heard, 8 did not include the phonemes /ae/ or /ay/ in content words. The remaining 8 contained either or both of these phonemes in content words. 3 contained /ae/, and 3 contained /ay/ while the remaining 2 contained both. The following will consider categorization patterns for each of the sentence types before considering all sentence types together.

Table 6.16: Summary of response measures for central participants

Overall Accuracy	0.29			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.15	0.18	0.42	0.28
Specificity	0.88	0.75	0.62	0.77
Precision	0.20	0.13	0.35	0.38
Detection Prevalence	0.13	0.24	0.39	0.25
Balanced Accuracy	0.51	0.47	0.52	0.53

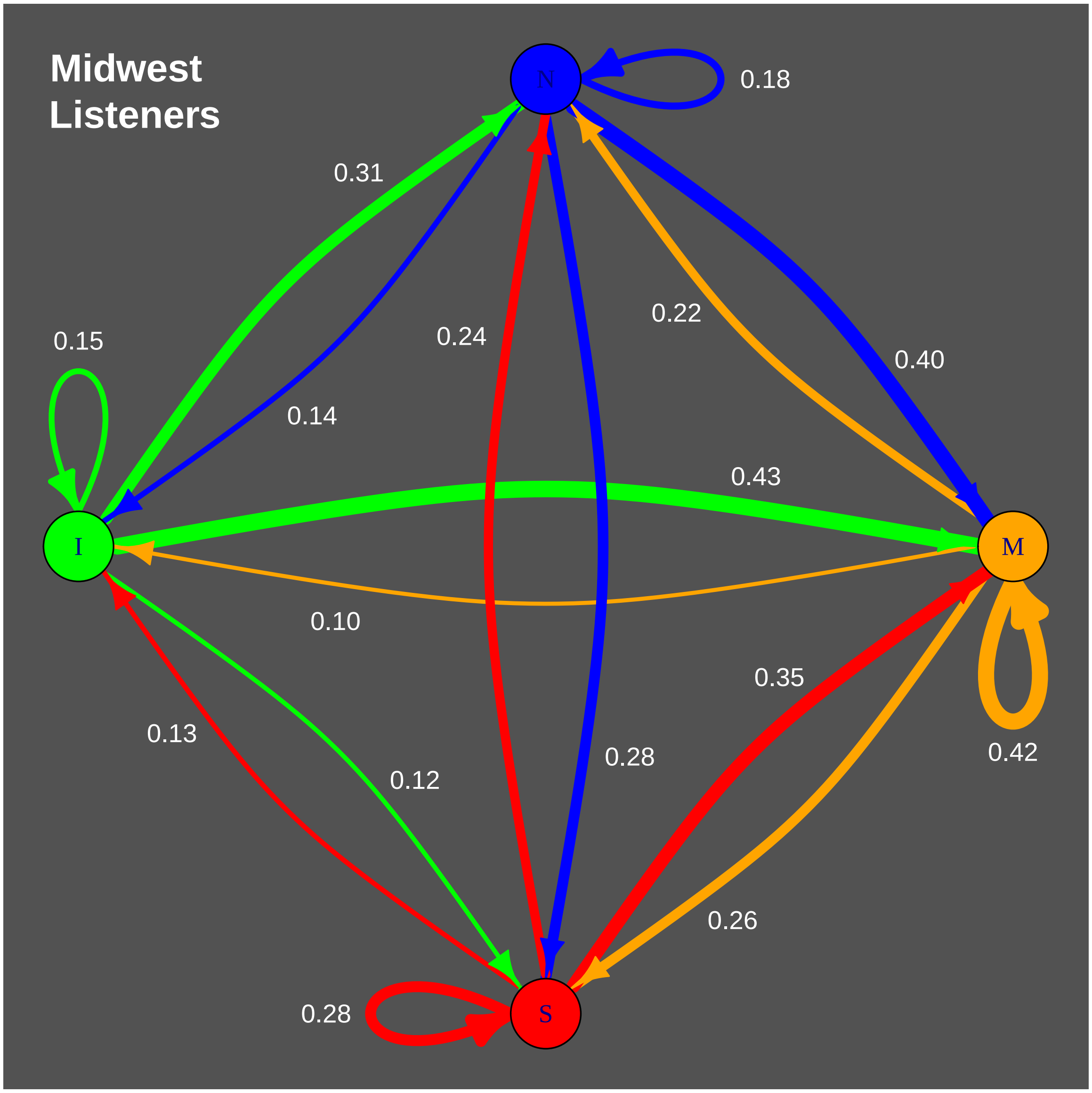

Figure 6.17: Response network graph for central participants

Table 6.17: Summary of response measures for southern participants

Overall Accuracy	0.31			
	Inland	**North**	**Midland**	**South**
Sensitivity	0.15	0.27	0.35	0.36
Specificity	0.88	0.76	0.68	0.72
Precision	0.21	0.18	0.36	0.39
Detection Prevalence	0.12	0.24	0.33	0.31
Balanced Accuracy	0.52	0.51	0.52	0.54

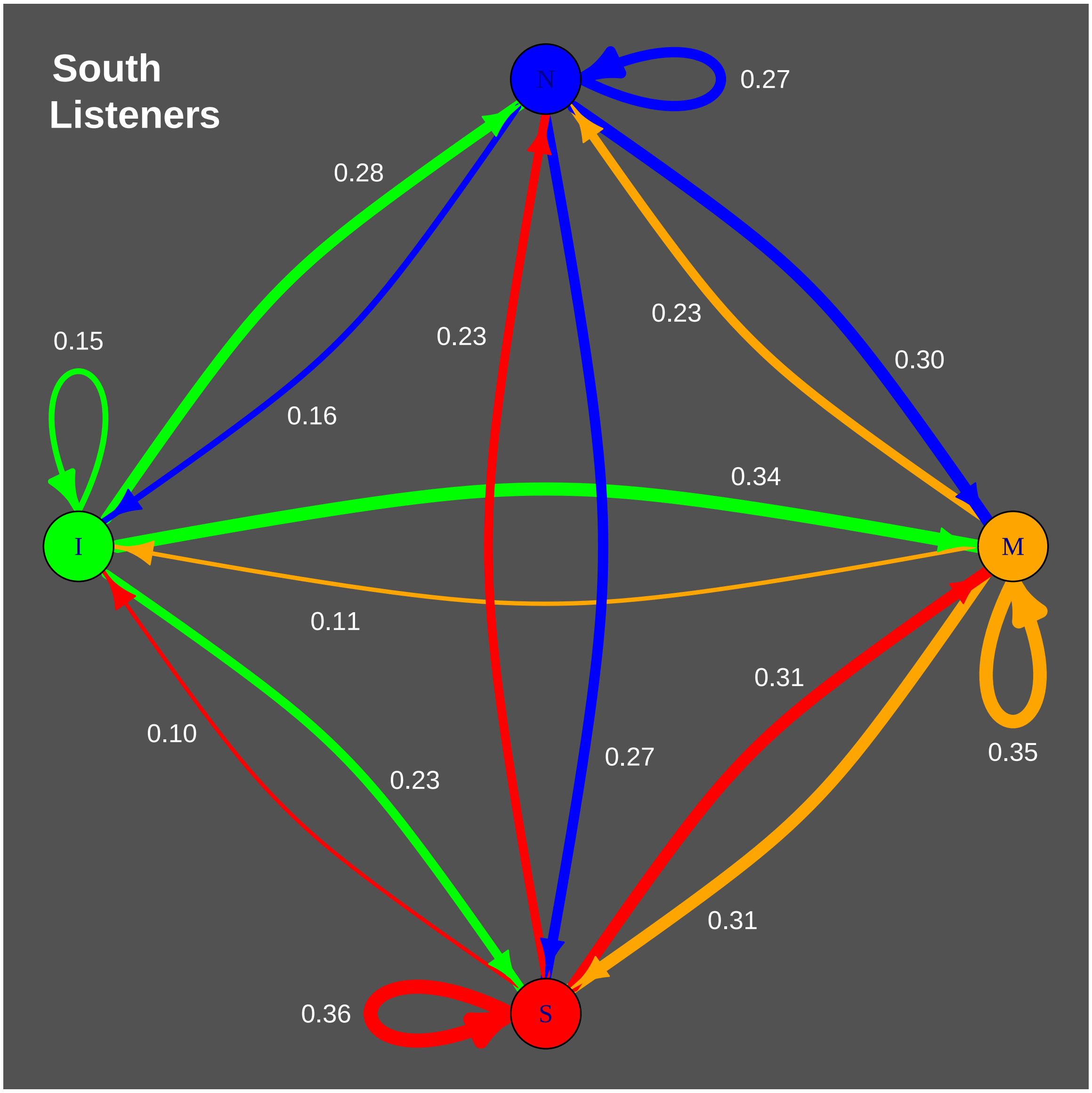

Figure 6.18: Response network graph for all southern participants

Table 6.19 presents the response frequencies for each response category and talker region by sentence type. This figure shows, for example, that when participants heard a sentence containing both /ae/ and /ay/ spoken by someone from Indianapolis, they responded with the Midland category, the orange bar, about 40% of the time. In this figure, each cluster of four bars, representing the response categories, add up to 1, or 100%.

Portraying these data in this way is useful for depicting patterns within the response categories; however, it is also problematic for comparing response frequencies between groups because of the response biases encoded in these data. For example, the responses to sentences that did not contain stereotyped phonemes show some really nice patterns in which response frequencies are highest for the "correct" responses and gradually decrease for talkers further away from the correct region. The red South bars are highest on the right where the talkers are from southern Indiana and lowest on the left where the talkers are from northern Indiana. Likewise, the green Inland North bars are tallest on the left where the talkers are from closer to the Inland North region and decrease in height toward the right as the talkers are further away. But the response biases obscure patterns in the relative response frequencies between response categories. That is, participants generally responded with the South category much more often than with the Inland North category, so even if participants responded South relatively less frequently for talkers from northern Indiana, their general tendency to choose South can overwhelm a category with a weaker response bias, such as the Inland North. If we consider the response behavior when participants heard talkers from the region of Indiana near Chicago speaking sentences containing both /ae/ and /ay/ (the far left set of bars in the lower-right of of Figure 6.19), it appears that the Inland category, the "correct" response, was selected the least. However, the response rate for South is actually rather low compared to all of the other red South bars in the figure. In

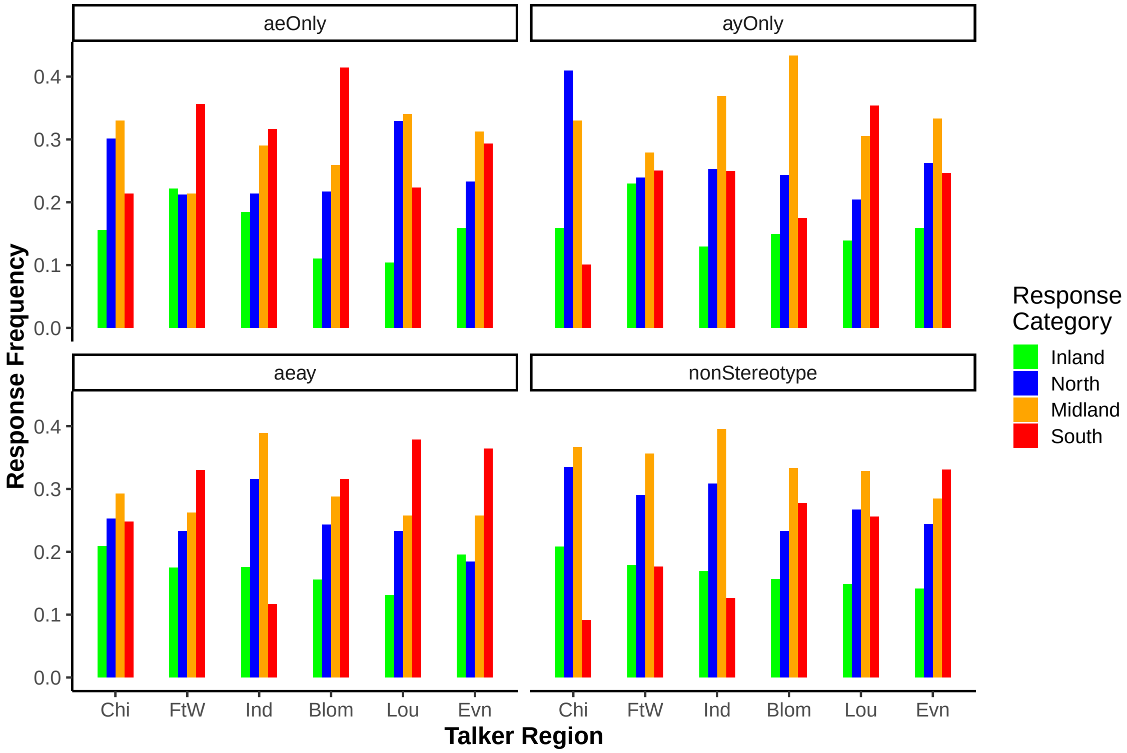

Figure 6.19: Categorization frequencies by sentence type for all talkers

fact, as will be shown later, responses for the Inland North for these talkers and sentences are relatively higher than most other sets of categories.

The between-group comparisons can be made more clear by standardizing the response rates. In the subsequent figures, the data within each response category have been standardized by subtracting the mean of the category and dividing by the standard deviation. This gives the data within each response category a mean of zero and a standard deviation of one. Figure 6.20 shows the effect of applying these operations on the data. We can now see that when people hear the talkers from near Chicago reading sentences that contain both /ae/ and /ay/, they respond with the Inland North category more than they usually respond with that category. In fact, participants tend to respond slightly less than average with the other categories.

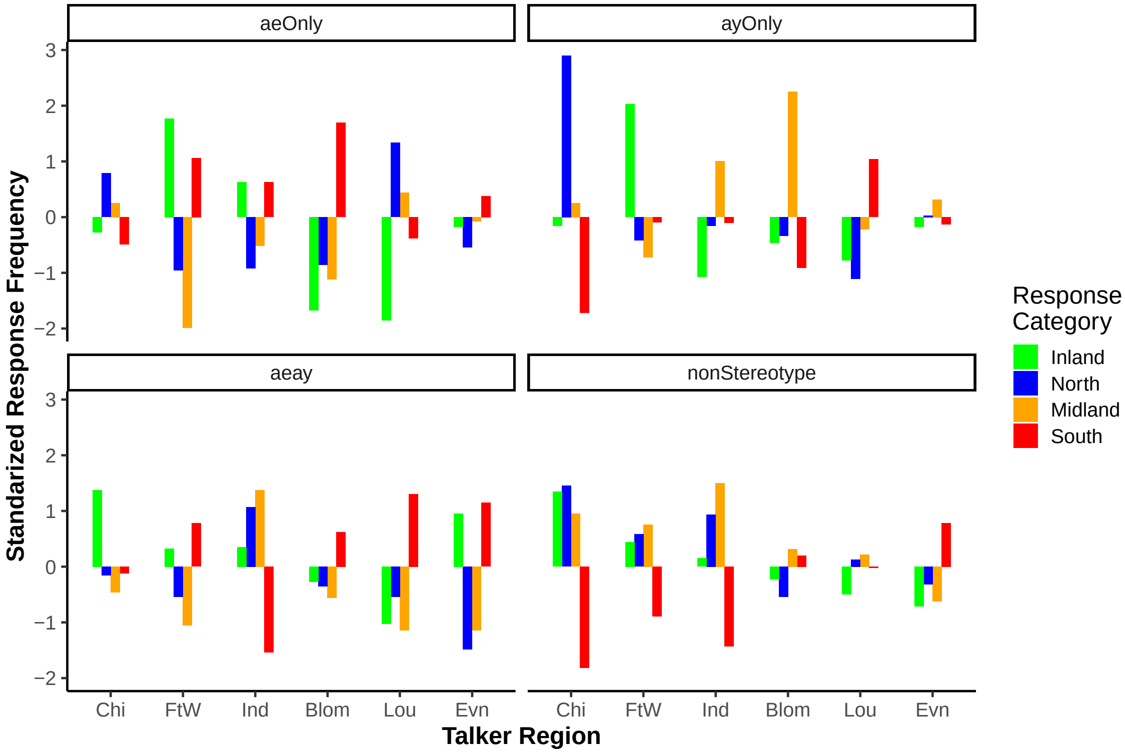

Figure 6.20: Standardized categorization frequencies by sentence type for all talkers

Figure 6.21 shows the standardized response rates for the group of sentences that contained /ae/ alone. We would expect that sentences that contain /ae/ would present listeners with a clear indication of the talkers' participation in an early stage of the Northern Cities Shift, /ae/ raising. If participants are sensitive to this feature, we would expect them to respond with the Inland North or, perhaps, the North category more frequently for talkers who are from the Inland North dialect region and also less frequently for talkers who are not from there.

Figure 6.21 also shows that talkers from Fort Wayne were categorized as Inland North most frequently, which is somewhat surprising since we would expect them to be categorized as North, but it is less surprising than that they were also categorized as South fairly frequently. The Inland North category was used least frequently for talkers from regions further south in Indiana, Bloomington, near Louisville, and Evansville, and this is expected.

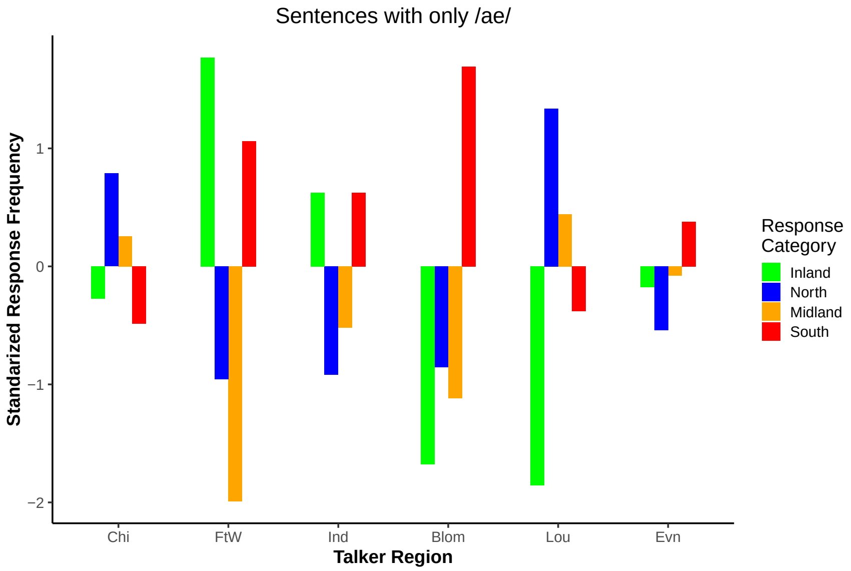

Figure 6.21: Standardized categorization frequencies for sentences containing /ae/ for all talkers

Talkers from near Chicago are classified as North. Taken with the results of the classification of people from Fort Wayne as Inland North, this could be the result of a confusion between the Inland North and North categories.

Figure 6.22 shows the standardized categorization frequencies for sentences of the type that only contain /ay/. Because monophthongization of /ay/ is a stereotypical speech pattern in the Southern dialect, we would expect that listeners would be more likely to categorize talkers from southern Indiana (Evansville, near Louisville, and maybe Bloomington) as South. They should also be less likely to categorize talkers from northern Indiana (near Chicago, Fort Wayne, and maybe Indianapolis) as South.

Here the expectations are not well supported. People from near Louisville are categorized as South most frequently, but people from Evansville are categorized at rates very close to the means of all categories. However, all of the talkers from Indianapolis and further south are categorized as Inland North or North less frequently than average, suggesting that listeners did not use the /ay/ feature as positive criteria for categorization but as negative criteria (i.e., "I don't know where you are from, but you aren't from the north.") In-

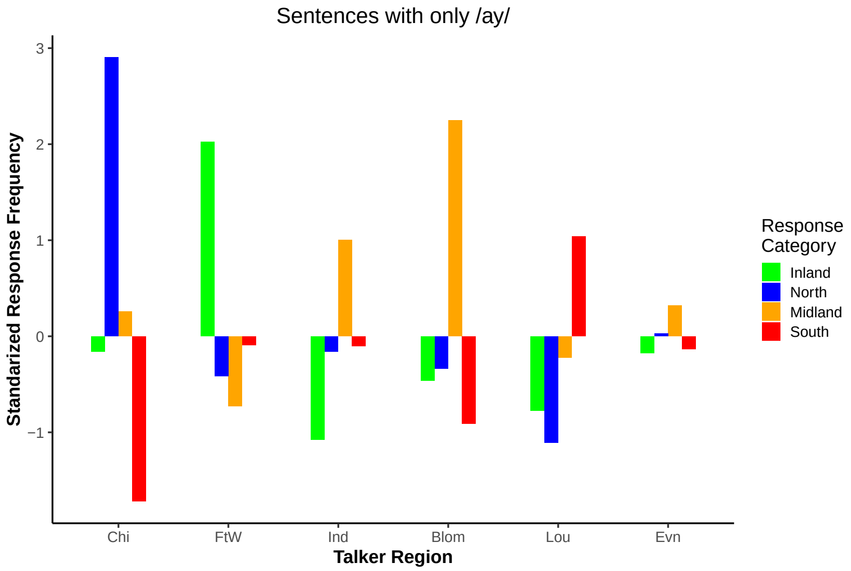

Figure 6.22: Standardized categorization frequencies for sentences containing /ay/ for all talkers

terestingly, this set of sentences was quite diagnostic for talkers from near Chicago and Fort Wayne, although the specific dialect categories are the reverse of what would be expected; the talkers from near Chicago are categorized as North while those from Fort Wayne are categorized as Inland North.

Responses to sentences that contained both /ae/ and /ay/, seen in Figure 6.23, show that these phonemes are most useful for identifying talkers as southern. The Bloomington, near Louisville, and Evansville talkers were identified as southern based on these sentences much more often than when they spoke other sentences. The sociophonetic content of these sentences was also used to identify these same talkers as *not* from the other three regions (with the curious exception of Evansville talkers also being identified as Inland). Listeners also used these sentences to identify talkers from Indianapolis as Midland or Northern but not as Southern. Finally, talkers from near Chicago were identified as Inland based on these sentences, much more so than when speaking sentences containing only /ae/.

The response patterns to sentences not containing sociolinguistic stereotypes are shown in Figure 6.24. Talkers from the southern half of the state were not reliably identified

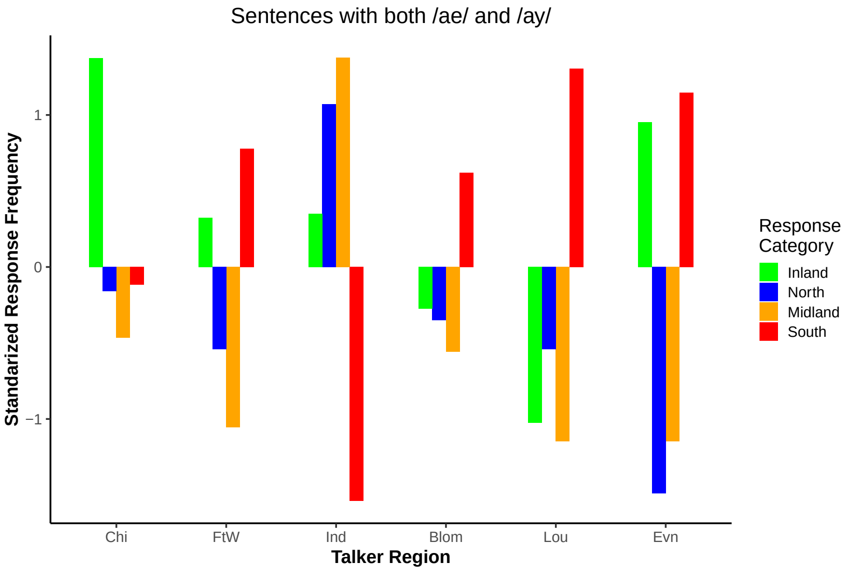

Figure 6.23: Standardized categorization frequencies for sentences containing both /ae/ and /ay/ for all talkers

based on these sentences. They were marginally less likely to be identified as Inland, and Evansville talkers were identified as Southern. The most action from these sentences is seen in the talkers from the northern half of the state. The content of these sentences allowed listeners to reject the talkers from near Chicago, Fort Wayne, and Indianapolis as Southern and to identify them as Northern. These sentences also helped listeners to identify the talkers near Chicago as Inland. It is also interesting to note the gradual step down in Inland responses from the near Chicago talkers to the Evansville talkers.

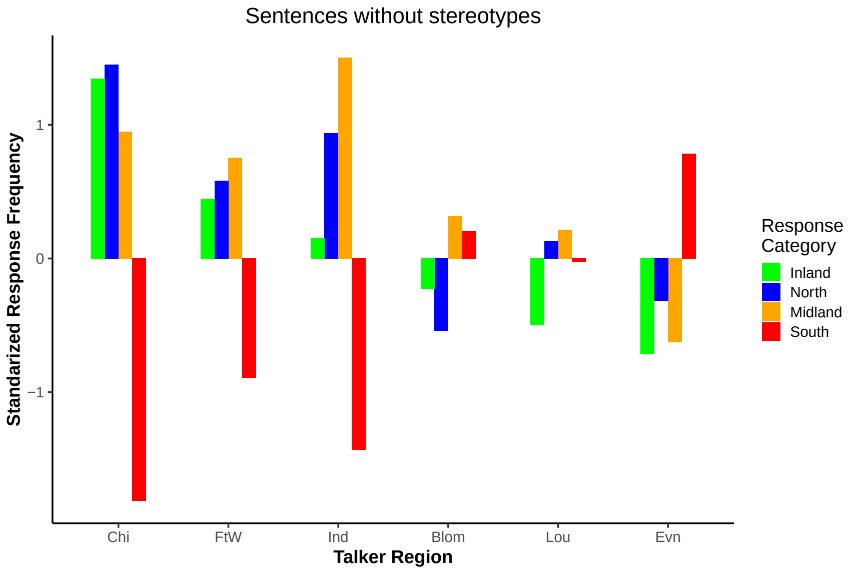

Figure 6.24: Standardized categorization frequencies for sentences containing neither of the stereotyped phonemes for all talkers

6.4 Discussion

The overall accuracy rates for all participants are rather low; however, they are above chance, and participants tend to be more accurate on their own dialect than on others. An important caveat on the accuracy rates concerns the *no information rate*, which is 33%. That is, if participants only selected the region with the greatest number of stimuli (i.e., their responses were not based on information from this stimuli), either the Midland or Southern region, they would have had an accuracy rate of 33%, a rate that no group achieved. Although this demonstrates an analytical difficulty, it also shows that participants' response biases did not overwhelm their signal detection abilities.

The response biases reveal an overall preference for the Midlands region followed by the participants' own region. Interestingly, participants were reluctant to select the Inland region near Chicago even though it is one of the most salient dialect regions in Indiana. Perhaps this is because the dialect of the Inland region is stereotyped and participants required a higher threshold to detect the dialect. A more mundane explanation is that the region was represented by a smaller surface area on the response screen than the other regions, and this fact could have primed the participants to also think responses for this area were less likely.

The perceptual categorization response patterns show how people categorize talkers. The four most notable patterns that appear to varying degrees among all participant groups are 1) a default response bias for Midland, 2) a confusion with default responses between Inland, North, and Midland with both North and Midland tending to share the response bias equally, 3) relatively high precision rates for the Inland category, and 4) relatively high precision rates for the South category.

The response patterns overall suggest a perceptual categorization continuum: Inland > North > Midland > South. Inland is very rarely confused for South. Inland is most confused with—and mostly classified as—North and Midland; however, the consistently high precision rates for Inland suggest that, while people are generally reluctant to categorize talkers as Inland, they tend to be correct when they do. Perhaps this reflects a stereotyped representation of the dialect that requires a high concentration of Inland features to achieve the classification threshold.

North talkers are generally not categorized as Inland, further supporting the hypothesis that categorization for Inland has a high threshold. North is frequently categorized as Midland. Some of the apparent confusion comes from Midland being a default response category, however, the rates at which Midland is categorized as North suggests that the two categories are genuinely confused in some cases. North is also categorized as South but less frequently than as Midland. Interestingly, however, black participants from northwest Indiana categorize North talkers most frequently as South, even more than as North. This North-South confusion appears to be equal as South talkers are categorized as North just as frequently for this participant group.

Midland talkers are most frequently categorized as Midland. This is partially due to a general response bias for Midland. Across listener groups, though, Midland talkers are frequently categorized as North and South. Both groups of participants in the north are more likely to claim a Midland talker is Northern than Southern. In a similar manner, southern listeners are almost as likely to consider a Midlands talker Southern as a Southern talker. As would be expected in a continuum, the listeners from central Indiana, geographically between the North and South dialect regions, are almost as likely to classify a Midlands talker as Northern or Southern.

Finally, being at an extreme end of the continuum, Southern talkers are categorized as either Southern or Midland, sometimes as Northern, and almost never as Inland. Southern talkers are categorized as Northern about as often as Northern Talkers are categorized as Southern suggesting an equal degree of confusability between the two. The rates of confusion between North and South are approximately around chance, 25%, so perhaps there is nothing more to make of the pattern than that. However, I do not suspect the higher rates of confusion between North and South talkers among the black northwestern participants is a coincidence given the structural similarities between Southern speech and AAVE.

The phonological content of the sentences played a role in the response patterns. Talkers from the Inland dialect region were identified correctly more often when they read sentences containing both /ae/ and /ay/ phonemes. These same sentences were also useful to listeners for identifying Southern talkers. Sentences that only contained a single sociolinguistic stereotype phoneme were generally not useful to listeners in identifying dialects. The sentences that were selected to include neither of the most salient stereotype phonemes of the Inland North and Southern dialects showed that listeners attended to some aspect of the talkers' speech patterns to determine if they were non-Southern, but only for people from the northern half of the state. This suggests that listeners rely on other dialect features to identify a dialect beyond the most stereotypical features.

The continuum of peoples' mental representations of these dialects is fairly smooth and consistent. The responses of northwestern listeners and southern listeners are basically inverses of each other, and the northeast and central responses are graded transitions between the two extremes. It is interesting to note an asymmetry in the classifications of the dialects on the extremes of the continuum, Inland and South. The Southern talkers are about as likely to be classified as Inland as Northern talkers are to be classified as Inland, suggesting that participants' mental representation of the Inland dialect place it as far away from the North dialect as North from South.

The general takeaway from this experiment is that Hoosiers as a whole distinguish two dialects in their state, corresponding to a North-South continuum. However, the continuum is punctuated depending on people's individual location along the continuum. The white northwestern participants behaved as if they only perceive reliable differences between two categories: South and Non-South. Black northwestern listeners show a notable confusion between North and South talkers suggesting a more complicated interaction with the continuum that deserves further consideration. Northeast listeners had the second highest overall accuracy and behaved as if they use three categories: North, South, and Other. Participants from central Indiana have confusion patterns that suggest only two reliable categories: North and South. The Southern listeners have the highest overall accuracy rate at 31%. While their confusion patterns are relatively complex, they suggest a three or even four category distinction considering their high precision for the Inland category.

Considering the response patterns of all the listener groups, it appears that some of the differences in categorization accuracy can be explained by the number of dialect categories the groups have access to. The Northwest and Central listeners had the lowest accuracies and also behaved as if they could only reliably access two categories. The Northeastern and Southern listeners, on the other hand, had higher accuracies and behaved as if they were sensitive to three or four dialect distinctions. An analysis of the acoustic characteristics of the talker tokens would shed further light on the relative abilities to discriminate dialects within the listener groups.

6.5 Appendix III: Raw Response Rate Confusion Matrices

Table 6.18: All Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	311	322	513	478
North	547	426	847	824
Midland	565	499	1155	1008
South	223	400	777	980

Table 6.19: All Northwest Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	174	168	295	272
North	239	188	403	379
Midland	207	186	417	365
South	66	146	258	359

Table 6.20: Northwest Black Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	20	9	32	32
North	36	23	47	49
Midland	20	21	60	48
South	4	27	21	31

Table 6.21: Northwest White Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	125	130	221	197
North	159	127	274	246
Midland	147	129	278	251
South	48	94	186	266

Table 6.22: Northeast Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	28	42	61	49
North	101	69	117	114
Midland	90	69	200	175
South	21	60	102	142

Table 6.23: Central Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	41	39	57	69
North	83	50	121	129
Midland	116	108	226	190
South	32	75	139	153

Table 6.24: South Participants

Actual	**Selected Region**			
Provenance	**Inland**	**North**	**Midland**	**South**
Inland	68	73	100	88
North	124	119	206	202
Midland	152	136	312	278
South	104	119	278	326

6.6 Appendix IV: Simulus Materials for 4AFC Task

Talker	SPIN Number	Sentence Type	Sentence Text
B07	101	nonStereotype	A round hole won't take a square peg.
B07	164	aeay	The flashlight casts a bright beam.
B07	170	nonStereotype	Follow this road around the bend.
B07	185	aeOnly	Paul hit the water with a splash.
B10	106	nonStereotype	Get the bread and cut me a slice.
B10	124	ayOnly	Her entry should win first prize.
B10	137	aeay	They tracked the lion to his den.
B10	179	nonStereotype	Banks keep their money in a vault.
B11	108	nonStereotype	Greet the heroes with loud cheers.
B11	129	aeOnly	Instead of a fence, plant a hedge.
B11	132	ayOnly	Her hair was tied with a blue bow.
B11	189	nonStereotype	Break the dry bread into crumbs.
B17	113	nonStereotype	He rode off in a cloud of dust.
B17	115	ayOnly	The super highway has six lanes.
B17	120	aeOnly	Old metal cans were made with tin.
B17	173	nonStereotype	Cut the meat into small chunks.
C01	101	nonStereotype	A round hole won't take a square peg.
C01	164	aeay	The flashlight casts a bright beam.
C01	170	nonStereotype	Follow this road around the bend.
C01	185	aeOnly	Paul hit the water with a splash.
C06	106	nonStereotype	Get the bread and cut me a slice.
C06	124	ayOnly	Her entry should win first prize.
C06	137	aeay	They tracked the lion to his den.
C06	179	nonStereotype	Banks keep their money in a vault.
C18	108	nonStereotype	Greet the heroes with loud cheers.
C18	129	aeOnly	Instead of a fence, plant a hedge.
C18	132	ayOnly	Her hair was tied with a blue bow.
C18	189	nonStereotype	Break the dry bread into crumbs.
C12	113	nonStereotype	He rode off in a cloud of dust.

C12	115	ayOnly	The super highway has six lanes.
C12	120	aeOnly	Old metal cans were made with tin.
C12	173	nonStereotype	Cut the meat into small chunks.
E01	101	nonStereotype	A round hole won't take a square peg.
E01	164	aeay	The flashlight casts a bright beam.
E01	170	nonStereotype	Follow this road around the bend.
E01	185	aeOnly	Paul hit the water with a splash.
E02	106	nonStereotype	Get the bread and cut me a slice.
E02	124	ayOnly	Her entry should win first prize.
E02	137	aeay	They tracked the lion to his den.
E02	179	nonStereotype	Banks keep their money in a vault.
E10	108	nonStereotype	Greet the heroes with loud cheers.
E10	129	aeOnly	Instead of a fence, plant a hedge.
E10	132	ayOnly	Her hair was tied with a blue bow.
E10	189	nonStereotype	Break the dry bread into crumbs.
E11	113	nonStereotype	He rode off in a cloud of dust.
E11	115	ayOnly	The super highway has six lanes.
E11	120	aeOnly	Old metal cans were made with tin.
E11	173	nonStereotype	Cut the meat into small chunks.
F04	101	nonStereotype	A round hole won't take a square peg.
F04	164	aeay	The flashlight casts a bright beam.
F04	170	nonStereotype	Follow this road around the bend.
F04	185	aeOnly	Paul hit the water with a splash.
F05	106	nonStereotype	Get the bread and cut me a slice.
F05	124	ayOnly	Her entry should win first prize.
F05	137	aeay	They tracked the lion to his den.
F05	179	nonStereotype	Banks keep their money in a vault.
F06	108	nonStereotype	Greet the heroes with loud cheers.
F06	129	aeOnly	Instead of a fence, plant a hedge.
F06	132	ayOnly	Her hair was tied with a blue bow.
F06	189	nonStereotype	Break the dry bread into crumbs.

F08	113	nonStereotype	He rode off in a cloud of dust.
F08	115	ayOnly	The super highway has six lanes.
F08	120	aeOnly	Old metal cans were made with tin.
F08	173	nonStereotype	Cut the meat into small chunks.
I01	101	nonStereotype	A round hole won't take a square peg.
I01	164	aeay	The flashlight casts a bright beam.
I01	170	nonStereotype	Follow this road around the bend.
I01	185	aeOnly	Paul hit the water with a splash.
I02	106	nonStereotype	Get the bread and cut me a slice.
I02	124	ayOnly	Her entry should win first prize.
I02	137	aeay	They tracked the lion to his den.
I02	179	nonStereotype	Banks keep their money in a vault.
I08	108	nonStereotype	Greet the heroes with loud cheers.
I08	129	aeOnly	Instead of a fence, plant a hedge.
I08	132	ayOnly	Her hair was tied with a blue bow.
I08	189	nonStereotype	Break the dry bread into crumbs.
I10	113	nonStereotype	He rode off in a cloud of dust.
I10	115	ayOnly	The super highway has six lanes.
I10	120	aeOnly	Old metal cans were made with tin.
I10	173	nonStereotype	Cut the meat into small chunks.
L01	101	nonStereotype	A round hole won't take a square peg.
L01	164	aeay	The flashlight casts a bright beam.
L01	170	nonStereotype	Follow this road around the bend.
L01	185	aeOnly	Paul hit the water with a splash.
L19	106	nonStereotype	Get the bread and cut me a slice.
L19	124	ayOnly	Her entry should win first prize.
L19	137	aeay	They tracked the lion to his den.
L19	179	nonStereotype	Banks keep their money in a vault.
L08	108	nonStereotype	Greet the heroes with loud cheers.
L08	129	aeOnly	Instead of a fence, plant a hedge.
L08	132	ayOnly	Her hair was tied with a blue bow.

L08	189	nonStereotype	Break the dry bread into crumbs.
L17	113	nonStereotype	He rode off in a cloud of dust.
L17	115	ayOnly	The super highway has six lanes.
L17	120	aeOnly	Old metal cans were made with tin.
L17	173	nonStereotype	Cut the meat into small chunks.