Effects of Incomplete Feedback on Response Bias in Auditory Detection: An Application of Bayesian Modeling to Real-world Listening Conditions

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts in the Graduate School of The Ohio State University

By

Shuang Liu, B.A., M.A.

Graduate Program in Speech and Hearing Science

The Ohio State University

2020

Thesis Committee

Lawrence L. Feth, Ph.D. Advisor

Robert A. Fox, Ph.D.

Abstract

The goal of many laboratory studies using the signal detection paradigm is to produce a bias-free estimate of listener sensitivity (Green and Swets, 1966). Forced-choice procedures with equal *a priori* probability of signal occurrence, a balanced payoff matrix, and trial-by-trial feedback are designed to assess sensitivity with response bias forced to zero. Attempts to approximate real-world listening situations in controlled laboratory settings do not lend themselves to traditional listening paradigms. Listeners in real-world listening tasks rarely encounter N-interval forced-choice opportunities. Further, many real-world listening situations do not lend themselves to trial-by-trial feedback.

Davis (2015) reported the effects of incomplete feedback on response bias for a simple tone-in-noise listening experiment with a single interval yes-no procedure. Feedback provided ranged from no feedback on any trial, to eight conditions with feedback for some signal–response combinations, to feedback on every trial. Davis reported a descriptive data analysis for each subject. The current study conducted an inferential data analysis with Bayesian statistics to estimate the effect of incomplete feedback on response bias in each experimental condition. The main finding is that, as expected, complete feedback drives response criteria toward the optimum, and incomplete feedback conditions result in various degrees of deviation from the optimal

criterion. The results can be interpreted in terms of how feedback conveys information about *a priori* probabilities of different states of the world and the values and costs of correct and incorrect responses.

Acknowledgments

I would first like to express my gratitude to my advisor Dr. Lawrence Feth for his support, guidance, and patience over the years. He opened the door to science for me and enlightened me on the breadth and depth of hearing science. He encouraged me to step out of my comfort zone, gave me the freedom to explore various research topics, and guided me back to where I started when I got lost. Without his guidance, I would not have the opportunity to appreciate the interdisciplinary nature of scientific research.

I am also very grateful to Dr. Robert Fox, who provided valuable suggestions on the thesis, to make it more understandable to a broader audience. I also appreciate his understanding and support of my situation whenever I needed them over the years.

I would also like to thank Dr. Jay Myung, who encouraged me to pursue more quantitative training, which is not only beneficial for this project but also my future career.

I would like to express special thanks to Ms. Jordan Vasko and Ms. Shuyuan Yu, my best friends in Columbus. Without your company and emotional support over the years, I would not be able to finish this journey with so much pleasure in my heart.

Vita

Jun. 2013……B.A., English,
Sun Yat-sen University

Aug. 2015…….....M.A., Linguistics,
University of Colorado Boulder

Aug. 2016 to May 2020……...Graduate Teaching Associate,
The Ohio State University

Aug. 2016 to May 2020…….....Graduate Research Associate,
The Ohio State University

Fields of Study

Major Field: Speech and Hearing Science

Table of Contents

List of Tables

List of Figures

Chapter 1. Statement of the Problem

The degree to which the results of a laboratory experiment reveal what happens in the real world depends on what factors or variables of the real-world problem are captured and modeled in the experiment. Auditory tasks in the real world include detecting warning sounds in noisy work environments, detecting a phone ringing in the presence of loud music, and detecting an ambulance siren in traffic. Many factors could have an impact on the detection performance, such as physical characteristics of the target and background sounds, the importance of the information carried by the target sound, how often the target sound occurs, whether the target sound occurs regularly or not. In addition, the biological characteristics of the observer and the number of other tasks that the observer must handle at the same time could influence performance as well. These factors together determine the actual detectability of auditory signals in the real world. The effects of some real-world factors are minimized in controlled laboratory experiments to test computational models for predicting human sensitivity to signal characteristics. However, many real-world listening situations do not lend themselves to traditional listening paradigms. To improve our ability to predict signal detectability in real-world listening situations, the gap between traditional listening paradigms and real-world listening situations should be mitigated by incorporating real-world factors into controlled laboratory design.

1.1 Examples of real-world listening situations

Bars and restaurants are common noisy environments where loud music and conversations are mixed in the background. If someone's phone ringtone is a piece of music, the ringtone may not be heard every time it occurs. Sometimes the ringtone could be easily masked by music in the background, as the background fluctuates. Both the intensity of the background noise and any similarity between the ringtone and the background will contribute to the difficulty of detecting the occurrence of a phone call in that environment. In the psychoacoustics literature, these effects are normally called energetic and informational masking (Brungart, 2001).

In addition, phone calls are of different importance. A phone call could come from "the boss" or it could come from an unknown caller. The cost of missing a phone call from someone's boss is probably higher than that of missing an unwanted call. The costs and values of missing versus answering phone calls may affect the level of alertness of the observer. If the phone number has been compromised and most phone calls are scam calls, the owner may not care much about missing a call at a party. If the owner is a busy executive who receives lots of working phone calls every day, they would probably check the phone very often during the party.

Warning sounds in noisy workplaces, such as the cockpit of an airplane, a nuclear power plant or a large manufacturing plant, are designed to alert workers to dangerous events. The alarm system is a critical part of the safety system of many workplaces. Desired detectability of the warning signals must be achieved through appropriate design and installation of the warning sounds. Design and installation of warning sounds should

consider characteristics of the noise background, acoustical properties of the work area, hearing status of workers, and use of hearing protection (Giguère et al. 2008). Dangerous events associated with alarms often require the operator's immediate attention. Failure to react to alarms can have significant negative outcomes. For example, in a military cockpit, dangerous events could be "Surface-to-air Missile", "Air Interceptor" and "Unknown Threat" (Smith et al. 2004). In a hospital, alarms could signify "Cardiac Crisis", "Staff Emergency" or "Patient Call" (Rayo et al., 2019). The cost of missing an event, versus the reward for responding to it, as well as the frequency with which each dangerous event occurs, might affect the operator's performance in the detection task.

Another real-world listening situation is detecting an ambulance siren in traffic. For the ambulance to move efficiently through traffic, other road users must be alerted by the ambulance siren and stop their vehicles immediately. However, the sound of the ambulance siren could be missed due to soundproofing or loud music inside the vehicle (Rane et al., 2019). Not reacting to ambulance siren in time could delay the arrival of the ambulance vehicle, which could result in serious negative outcomes.

Each detection situation has its unique features as well as some common features shared by most situations. To improve the sound design and better predict the detectability of auditory signals in the real world, one could conduct either field studies or laboratory studies to test the effects of various factors. Each type of study has its advantages and limitations.

1.2 Modelling real-world listening situations and testing them in a laboratory

Laboratory experiments and field experiments are complementary methods for the study of cause-and-effect relationships (Aziz, 2017; Coppock and Green, 2015). A field experiment is conducted in a "real-world" setting, the results of which are more likely to be generalized beyond the current study (VandenBos, 2007). The independent variables may or may not be deliberately manipulated by the researcher. Participants may or may not be aware that they are being studied and observed for their reactions. Subjects often are not randomly assigned to experimental conditions.

On the other hand, laboratory research is conducted in a setting completely designed by the experimenter. It is a tightly controlled investigation in which the researcher manipulates the factor under study to determine if such manipulation generates a change in the subjects' performance (Aziz, 2017). Compared to a field experiment that has less control over extraneous variables, a laboratory experiment is more likely to be free of flaws in its conclusions about cause-and-effect relationships among variables (VandenBos, 2007).

Both types of experiments are important for the understanding of human auditory detection. An example is the detection of electric vehicles by pedestrians. Electric Vehicles (EV) produces much less noise compared to normal vehicles, however, the absence of warning sounds entails a risk for pedestrians, especially for those who are visually impaired (González-Hernández et al., 2017). The investigation of the detection of EV alerts ranges from perceptual experiments in the laboratory using static listeners to explore the influence of various timbre parameters on sound detectability (Parizet et al.,

2013) to outdoor experiments with real background noise and subjects walking in a real pedestrian area to explore responses produced in a dynamic urban environment (González-Hernández et al., 2017).

Before any field experiment is carried out, laboratory testing usually comes first. One should carefully examine the difference between the laboratory setting and real-world situations, incorporate real-world factors into the laboratory experiments, and control them as much as possible. This will probably benefit research in at least two respects: it will improve the external validity of the results obtained from the laboratory experiment, and it will help eliminate the effects of extraneous variables that are not under control in a field experiment.

In psychophysics, laboratory experiments are usually designed to test computational models. Models are logical frameworks composed of a set of assumptions that describe the underlying mechanism of the problem under investigation. Models help predict the result of an experiment and experiments test whether the prediction is accurate or not. In psychophysics, a group of models based on the Theory of Signal Detection (TSD) (Green and Swets, 1966) are of great importance in separating the effects of sensory factors and non-sensory factors. Section 1.3 briefly reviews the theoretical importance of TSD in psychophysics, elements of TSD, and psychophysical procedures of TSD. Section 1.4 discusses the gap between typical psychophysical experiments based on TSD and signal detection in real-world situations.

1.3 The contribution of TSD to psychophysics, elements of TSD and psychophysical procedures

In the early psychophysics experiments, it was assumed that the probability of a yes response for a stimulus presentation was entirely determined by the stimulus and the biological state of the sensory system (Gescheider, 2013, pp. 93). Later, it was noted that at least two non-sensory factors also influence the probability of yes responses: stimulus probability and response consequences (Gescheider, 2013, pp. 98). TSD was the major theoretical advance made in the 1950s to separate the influence of non-sensory factors upon detection from sensitivity (Harvey, 2014).

TSD was built on statistical decision theory (Green and Swets, 1966, pp. 7). The decision that an observer must make when detecting a brief tone in white noise, such as the task in Davis (2015), is assumed to be a statistical one. A simple tone-in-noise task was used to minimize the effects of signal and noise characteristics on listeners' performance. The decision is based on evidence that is ambiguous in the sense that it does not completely support one of several hypotheses. According to TSD, the level of the *noise* against which a signal is detected may be either external or internal to the detecting device (or both) and is assumed to vary randomly from moment to moment (Gescheider, 2013, pp. 105). A stimulus is either noise alone, or a weak signal plus noise. The observer must make a sensory *observation* x (an internal representation of the stimulus) of the stimulus and then decide whether the observation is due to noise alone (NA) or signal plus noise (SN). The magnitude of the observation due to NA varies randomly. The randomness is described by a probability distribution. When a signal is added, the probability distribution is shifted to the right. That is, the average sensory

observation magnitude is assumed to be greater for signal plus noise (SN) than for noise alone (NA). The more overlap between the two distributions, the more difficult to make a correct decision (Gescheider, 2013, pp. 106).

In TSD, the basis for making a decision is the *likelihood* ratio corresponding to the magnitude of the observation, which is the ratio of the probability, or likelihood, of the observation given the stimulus is SN to the probability of the observation given the stimulus is NA.

$$l(x) = \frac{P(x \mid SN)}{P(x \mid NA)}$$

If $l(x)$ is greater than 1, it indicates that it is more likely that the stimulus presented was SN rather than NA. If $l(x)$ is smaller than 1, it indicates that it is more likely that the signal was not there. If $l(x)$ is equal to 1, it indicates that SN and NA are equally likely (Gescheider, 2013, pp. 107). TSD assumes that the observer operates by a *decision rule*. A particular value of $l(x)$ is set as the cutoff point above which the observer responds SN and below which the observer responds NA. It is called the *likelihood ratio criterion*, denoted by β. The *optimal likelihood ratio criterion*, which optimizes the performance in a long series of observations, is a function of the *a priori* probabilities of SN and NA trials and the costs and values of the various decision outcomes (Gescheider, 2013, pp. 112; Green and Swets, 1966, pp. 23).

$$\beta_{opt} = \frac{P(NA)}{P(SN)} \frac{[V(CR) + C(FA)]}{[V(H) + C(M)]}$$

where V(CR) is the value of a CORRECT REJECTION, C(FA) is the cost of a FALSE ALARM), V(H) is the value of a HIT, and C(M) is the cost of a MISS. (Values and costs

are all entered as positive numbers.) The concept of *optimal likelihood ratio criterion* will be critical for the discussion in chapter 4.

The unique contribution of TSD to psychophysics is that it offers a means to measure sensitivity and criterion independently (Gescheider, 2013, pp. 118). The index of stimulus detectability is d', which is the difference between the means of the SN and NA distributions divided by the standard deviation of the NA distribution. The index of *response bias* (the tendency of the observer to favor one response over another) is C, which is the number of standard deviation units that the response criterion is above or below the *zero bias point* where the NA and SN distributions cross. At the point where the NA and SN distributions cross, $l(x)$ is 1 and C is 0. Above that point, $l(x)$ is larger than 1 and C is positive. Below that point, $l(x)$ is smaller than 1 and C is negative.

Response proportions, from which the theoretical constructs of sensitivity and criterion are estimated, are obtained using psychophysical procedures such as the single interval yes-no procedure (SIYN) and the two-alternative forced-choice (2AFC) procedure (Gescheider, 2013, pp. 142-146). The SIYN procedure presents a sequence of trials to the observer. One trial only contains one observation interval where a signal may be present or absent. The observer must judge whether a trial contains a signal or not. 2AFC differs from SIYN in that two observation intervals are presented in a trial. The observer must report which interval contained a signal after a trial.

1.4 Limitations of Experiments based on TSD

This section discusses several discrepancies between typical signal detection experiments based on TSD and real-world detection situations. One issue is that typical signal detection experiments based on TSD often clearly define an observation interval by visual indicators (e.g., lights). Observers know when to pay attention and when to relax. However, usually there is no visual cues telling the observer when to start to pay attention to a signal in the real world. In some situations, observers must pay sustained attention in searching for a signal during a long period of time. This has stimulated a large amount of research on the vigilance problem since the 1950s (Mackworth, 1956; Watson and Nichols, 1976; Swets, 1977).

Another issue is that although 2AFC is often chosen as the psychophysical procedure in many signal detection experiments to estimate listener sensitivity which is uncontaminated by variations in response criterion (Green and Swets, 1966, pp. 43-44; Gescheider, 2013, pp.146), listeners rarely encounter tasks similar to 2AFC in the real world. To better resemble real-world detection tasks, a SIYN procedure or a vigilance task should be used.

In a typical signal detection experiment, equal *a priori probabilities* of SN and NA events and equal values for correct responses and costs of incorrect responses are assigned to direct the response criterion toward the unbiased position. However, these conditions rarely occur in the real world. For example, the probability of signal occurrence may be considered low when the task is to detect an ambulance siren. The cost of missing an ambulance siren may be considered as larger than the cost of a false

alarm (“heard” a siren when there was none). These real-world conditions could shift the response criterion from the unbiased position.

Usually, a running account of the observer’s performance is provided by trial-by-trial feedback. Trial-by-trial feedback gives the observer complete knowledge of whether the decision outcome is a hit, a miss, a false alarm, or a correct rejection. However, in the real world, an observer usually does not have access to complete knowledge of results. For example, when an ambulance vehicle passes other vehicles stopped at the roadside, drivers at the roadside know that they have correctly detected a siren. When a driver misheard a sound as an ambulance siren and prepared to stop, he might realize that it was a false alarm when he saw all other vehicles do not stop. When a driver missed a siren behind him and shortly turned right or left, he may never know that he has missed a siren. Incomplete knowledge of results is a non-sensory factor, which could influence an observer’s response criterion. The purpose of the research carried out by Davis (2015) was to study the effects of incomplete feedback on response bias. A detailed review of Davis (2015) will be given in Chapter 2.

In summary, detection tasks in real-world conditions, such as detecting a siren in traffic, could differ significantly from a typical signal detection experiment in the laboratory in many aspects. Davis (2015) focuses on the discrepancy in an observer’s knowledge of results. Chapter 2 reviews the experiment design, data analysis, and results of Davis (2015), and proposes a further analysis of the results of Davis (2015).

Chapter 2 Literature Review

Feedback provides information about the outcomes of past responses in a psychophysical experiment. Since the 1950s, knowledge of results was used to train observers to achieve asymptotic performances (Davis, 2015; Green and Swets, 1966, pp. 395). Since it is rare that observers will always know whether each response is correct or incorrect in real-world detection situations, the effects of two forms of incomplete knowledge of results have been studied. One type of study, known as providing "partial feedback", manipulated the proportion of trials for which the observer receives feedback (Lurie and Swaminathan, 2009; Szalma et al., 2000). Another type of study, known as "incomplete feedback", manipulated which response type (HIT, MISS, False Alarm or Correct Rejection), or combinations of response types received feedback (Szalma et al., 2006, Davis, 2015). The subtle difference between the two is that "partial feedback" is given for all four response types, but only on a percentage of the experimental trials. "Incomplete feedback" can occur on any trial for the designated response type(s). The experiment of Davis (2015), reviewed below, provided "incomplete feedback".

2.1 The approach used by Davis (2015)

2.1.1 Description of the experiment

Davis (2015) investigated the influence of incomplete feedback on response bias for a simple tone-in-noise detection task. The psychophysical procedure used was a

single-interval yes-no procedure. A signal-plus-noise (SN) trial contained a 500 ms broadband, white noise and a 20 ms 1 kHz tone presented in the temporal middle of the white noise (see Figure 1). A noise-alone (NA) trial contained only the 500 ms white noise. The reason for using a pure tone as the signal and a white noise as the background was to minimize the effects of signal and noise characteristics on listener performance. Subjects were asked to respond whether a trial contained the 1 kHz pure tone or not by clicking a button on the user interface shown in Figure 1. Then visual feedback was given to show whether the response "yes" or "no" was correct or incorrect. A correct response was marked with a check with the response button turned green. An incorrect response was marked with a cross with the response button turned red.

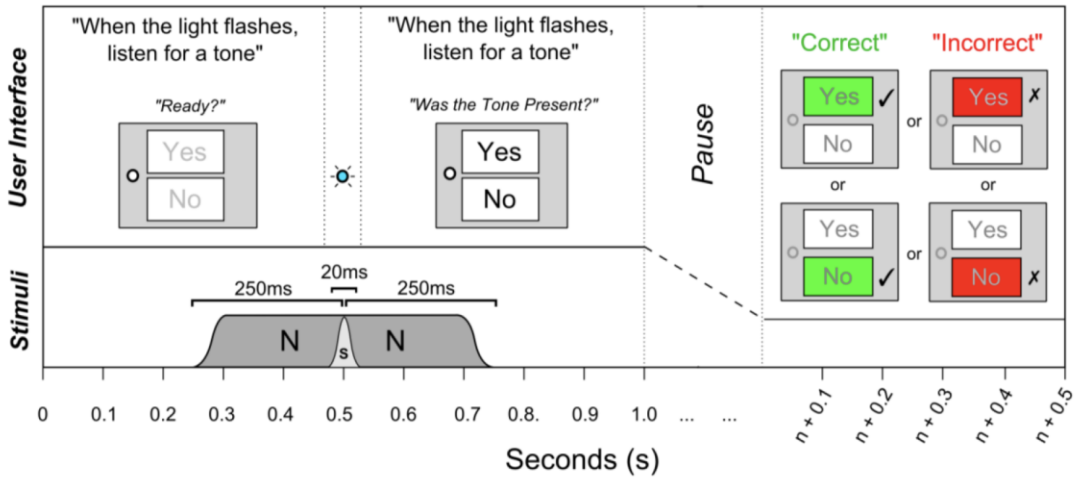

Figure 1. Timing and organization of a trial for the *familiarization, threshold estimation,* and *IKR tasks* used by Davis (2015). The trial is divided into three sections: (1) *stimulus presentation* lasing 1000 ms, (2) a *pause* giving the subject time to respond, and (3) display of *feedback* information lasting 500 ms. Reproduced from Davis (2015).

Each experimental condition provided a unique type of feedback to the listener, ranging from no feedback on any trial to eight conditions (listed below) with feedback for some combinations of HIT, MISS, FA and CR, to complete feedback on every trial. The organization of the experimental conditions is shown in Figure 2. The 10 feedback conditions are (1) no feedback, (2) feedback only for HIT, (3) only for MISS, (4) only for FA, (5) only for CR, (6) only for "yes" trials (HIT and FA), (7) only for "signal" trials (HIT and MISS), (8) only for "correct" trials (HIT and CR), (9) only for HIT, MISS and FA, and (10) for all SR combinations (all trials). The No Feedback condition (1) was presented first and the All Feedback condition (10) was presented last for all listeners. Conditions (2) to (8) were presented in a unique random order for each subject. Each condition contained 10 blocks of 50 trials. In each block, half of the trials were signal trials and half were noise trials presented in a random order.

Ten young adults with normal hearing participated in the experiment. Davis (2015) attempted to maintain sensitivity as constant as possible across the experimental conditions. Subjects were matched for sensitivity by a single-interval adaptive procedure (Kaernbach, 1990). Signal-to-noise ratio (SNR) for 75% detection threshold was established for each subject. Then that SNR was used for each subject throughout the experiment.

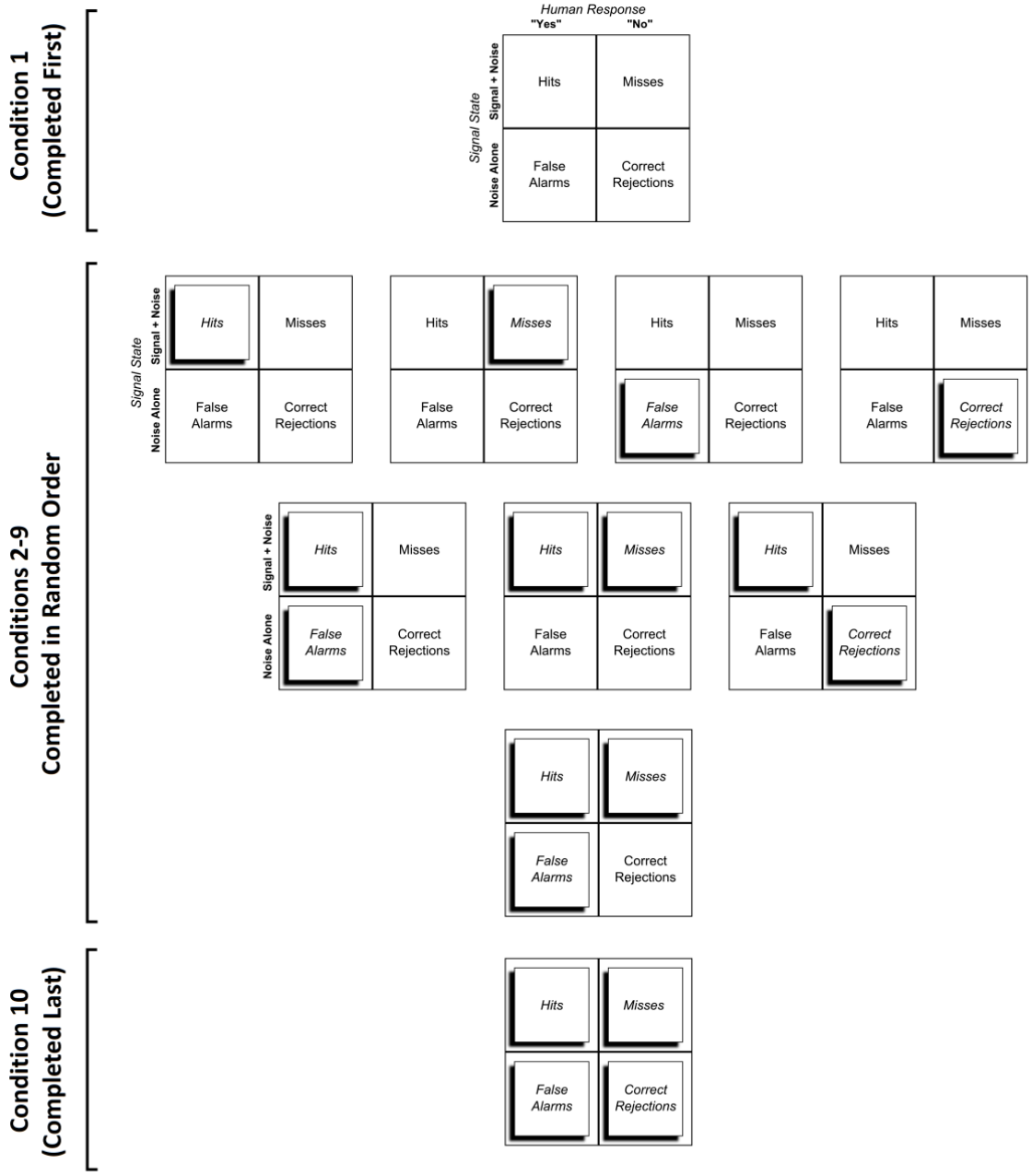

Figure 2. Experiment conditions in the Davis (2015) study. Condition (1) was presented first and condition (10) was presented last. All other conditions were completed between conditions 1 and 10 and were completed in random order. Each matrix represents a condition in which feedback is provided for the shaded box. Reproduced from Davis (2015).

2.1.2 Descriptive data analysis and its limitation

The optimal likelihood ratio criterion for the experiment in Davis (2015) is 1, since *a priori* probabilities of SN and NA trials were equal, and values of correct responses and costs of incorrect responses were not specified. The optimal response criterion is located at the unbiased position ($C_{opt} = 0$) for this experiment. When feedback is complete, it is expected that the observers' response criterion would be at the unbiased position. When feedback is incomplete, the observers' response criterion might deviate from the unbiased position.

Davis (2015) labeled his research questions with four keywords: *Symmetry*, *Organization*, *Implicitness*, and *Amount*.

Symmetry: Were the response biases for the four conditions [HIT], [MISS], [FA], and [CR] all equal?

Organization: Did bias differ when only "yes" trials got feedback, or only "signal" trials got feedback, or only "correct" trials got feedback? That is, were the response biases for the conditions [HIT, FA], [HIT, MISS], and [HIT, CR] all equal?

Implicitness: Were subjects able to utilize missing feedback to achieve the unbiased response criterion? That is, were the response criteria in conditions [HIT, MISS], [HIT, CR] and [HIT, MISS, FA] unbiased?

Amount: Did providing feedback for more types of decision outcomes reduce response bias?

Davis (2015) reported a descriptive data analysis for each subject. In each condition, every subject completed 10 blocks, which yielded 10 pairs of *HIT* and *FA rates*. Each pair of HIT and FA rates yields a pair of sensitivity and bias scores (Macmillan, & Creelman, 2004), as shown by the following equations.

$$d' = Z(hit\ rate) - Z(false\ alarm\ rate)$$

$$C = -0.5[Z(hit\ rate) + Z(false\ alarm\ rate)]$$

For each subject, a 95% confidence interval of response bias was constructed based on the 10 data points in each condition. For each subject, pairwise comparisons of biases across conditions was done by evaluating the amount of overlap of the confidence intervals. There are 6 possible pairwise comparisons among the 4 conditions in *symmetry* (HIT, MISS, FA, CR). For example, the evaluation of the difference between the means of the [HIT] and [MISS] conditions compared the differences between the two conditions where the target signal was present. For each subject, Davis (2015) counted the number of significant differences out of the 6 pairwise comparisons. For 9 out of 10 subjects, more than 3 out of 6 pairwise comparisons yielded significant results, indicating response biases were not all equal across [HIT], [MISS], [FA], and [CR].

Davis (2015) also analyzed the direction and degree of response bias for each condition in [HIT], [MISS], [FA] and [CR]. Most subjects showed response bias toward "No" (being conservative; showing a tendency to report signal absence; C is positive) in most conditions. Most subjects showed the least amount of response bias when only HITs received feedback and the most response bias when only CRs received feedback.

There are 3 possible pairwise comparisons among the 3 conditions in *organization* ([HIT, MISS], [HIT, FA] and [HIT, CR]). For 8 out of 10 subjects, more than 2 out of 3 pairwise comparisons yielded significant results. Again, this indicates response biases were not all equal across [HIT, MISS], [HIT, FA] and [HIT, CR]. Since most subjects showed response bias toward "Yes" (being liberal; showing a tendency to report signal presence; C is negative) in some conditions and toward "No" in other conditions, Davis (2015) concluded that response bias shows no apparent pattern toward "Yes" or "No" in these three conditions.

To explore whether subjects could utilize missing feedback to reduce response bias, Davis (2015) compared response criterion in [HIT, MISS], [HIT, CR] and [HIT, MISS, FA] to the unbiased position as well as the criterion in [ALL-KR] which was assumed to be least biased among all conditions. Most subjects yielded criterion close to the unbiased position or the criterion when HITs, MISSes, FAs, and CRs all received feedback.

To explore whether response bias is reduced when more feedback is provided, Davis (2015) analyzed the relationship between response bias and the number of types of feedback provided in each condition. Each condition provided feedback for 1, 2, 3, or all 4 types of decision outcomes. Results from 6 out of 10 subjects suggested that the more the feedback provided, the less the response bias.

The 10 subjects in Davis (2015) showed some common characteristics as well as idiosyncrasies in their response biases. The main limitation of the individual analysis is that it is purely descriptive, and it does not offer inference about the behavior of the

population. An estimate of the group-level response bias for each experimental condition may reveal new insights into the underlying mechanism of how incomplete feedback influences response bias in auditory signal detection. An approach to estimating model parameters of TSD (sensitivity and response bias) has been advocated by some Bayesian modelers (Lee, 2008; Lee and Wagenmaker, 2014; Rouder and Lu, 2005) for certain statistical advantages. The next section briefly describes this modeling and the inference procedure.

2.2 Inferential data analysis using Bayesian statistics

2.2.1 The probability model

The equal-variance Gaussian signal detection framework with an extension is used to model the data generating process in each condition. It is assumed that the probability distributions of the magnitude of observation due to NA and SN are both Gaussian. The mean of the distribution under NA is set as 0. The variances of both distributions are set as 1. Compared to the distribution under NA, the distribution under SN is shifted to the right by *d'* standard deviations. Thus, the distribution under NA is the standard normal distribution $N(0, 1)$, and that under SN is $N(d', 1)$.

It is assumed that each subject has an individual sensitivity and an individual response bias, denoted by d_i and c_i. The left panel of Figure 3 shows the model for subject i in a condition. It is assumed that individual response biases (c_is) and sensitivities (d_is) for this group of subjects is a sample from the population. The right

panel of Figure 3 shows the model on the left panel with an extension that models the individual biases or sensitivities as observations from a group-level Gaussian distribution.

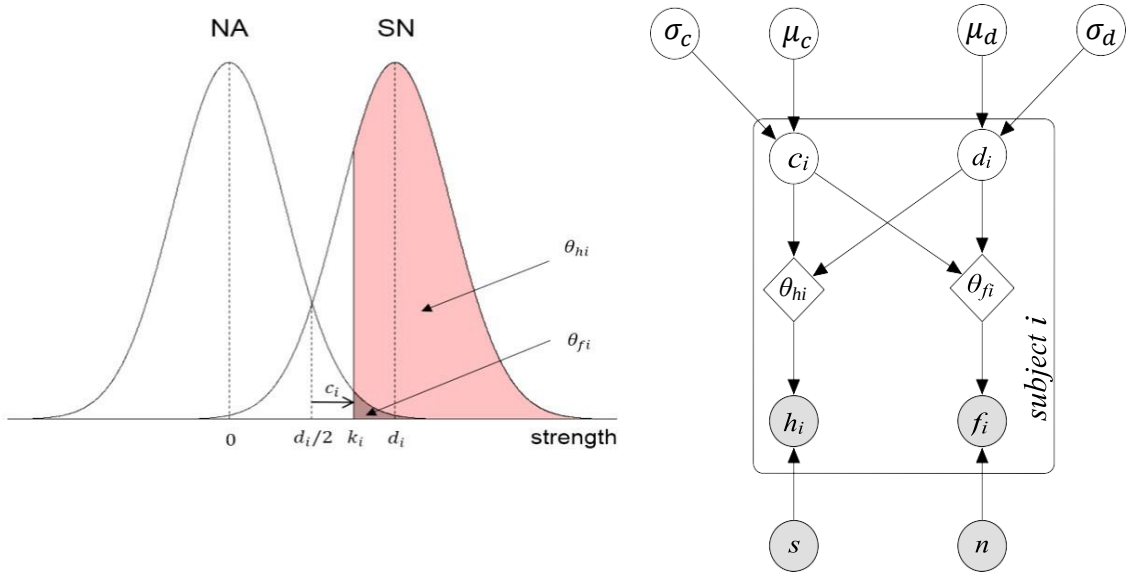

Figure 3. Left: Equal-variance Gaussian signal detection theory framework for subject i, where c_i is the bias parameter, d_i is the sensitivity parameter, k_i is the response criterion, θ_{hi} is the HIT rate, and θ_{fi} is the FA rate. Right: Equal-variance Gaussian signal detection theory framework with hierarchical extension. s is the number of signal trials and n is the number of noise trials in a condition. h_i is the observed number of HITs and f_i is the observed number of FAs for subject i in a condition. μ_c and σ_c are the mean and standard deviation of the parent distribution of individual bias parameters. μ_d and σ_d are the mean and standard deviation of the parent distribution of individual sensitivity parameters.

Each condition contains 500 trials (10 blocks × 50 trials / block). Half of them are signal trials and half of them are noise trials. In the right panel of Figure 3, s = n = 250. The observed data for each subject are the HIT and FA counts (h_i and f_i), which are assumed to follow binomial distributions. θ_{hi} and θ_{fi} are the HIT and FA probabilities.

$$h_i \sim Binomial(s, \theta_{hi})$$

$$f_i \sim Binomial(n, \theta_{fi})$$

HIT and FA probabilities, θ_{hi} and θ_{fi} , are determined by sensitivity d_i and response bias c_i.

$$\theta_{hi} = \Phi(\frac{1}{2} d_i - c_i)$$

$$\theta_{fi} = \Phi(-\frac{1}{2} d_i - c_i)$$

Individual sensitivities, d_is, are modeled as random draws from a Gaussian parent distribution with mean μ_d and standard deviation σ_d. Individual biases, c_is, are viewed as random draws from a Gaussian parent distribution with mean μ_c and standard deviation σ_c.

$$d_i \sim Gaussian(\mu_d, \sigma_d)$$

$$c_i \sim Gaussian(\mu_c, \sigma_c)$$

2.2.2 Bayesian inference procedure

In Bayesian statistics, the uncertainty about a parameter is quantified with a probability distribution. Before one sees any data, a *prior distribution* is given to each parameter to be estimated. Bayes theorem (Bayes, 1763) lies at the core of Bayesian statistics, which specifies how the *prior distribution* is updated with the information collected from the data to arrive at the *posterior distribution* (Lee & Wagenmakers, 2014, pp. 3). Diffuse priors were set for μ_c, μ_d, σ_c, and σ_d. Markov Chain Monte Carlo sampling of posterior distributions of μ_c, μ_d, σ_c and σ_d were done in WinBUGS (Spiegelhalter et al., 2003) and R(Ripley, 2001). Figure 4 shows the script for setting up the model in WinBUGS. For each parameter, an empirical density of the posterior distribution was obtained from the samples.

In addition to reporting the empirical posterior distribution, the location of the posterior is usually summarized with mean, median, or mode. The selection of the summary statistic depends on the shape of the posterior. Although mean is very frequently used, if the posterior has a heavy tail, then mean may not be a good choice, since it is easily influenced by extreme values and may be far away from where the most probability is located (Bolstad and Curran, 2016, pp. 150). The spread of the posterior is summarized with variance or standard deviation (Bolstad and Curran, 2016, pp. 151). To find a high probability interval for the parameter, a credible interval (highest density interval) is often reported (Bolstad and Curran, 2016, pp. 153). The lower bound of a 95% credible interval is the 2.5% quantile of the posterior and the upper bound is the 97.5% quantile of the posterior.

```
model{
  for (i in 1:k){

    # observed counts

    h[i] ~ dbin(thetah[i],s)
    f[i] ~ dbin(thetaf[i],n)

    # reparameterization using equal-variance Gaussian SDT

    thetah[i] <- phi(d[i]/2-c[i])
    thetaf[i] <- phi(-d[i]/2-c[i])

    # bias and sensitivity

    c[i] ~ dnorm(muc,lambdac)
    d[i] ~ dnorm(mud,lambdad)

  }

  # priors

  muc ~ dnorm(0,.001)
  mud ~ dnorm(0,.001)
  lambdac ~ dgamma(.001,.001)
  lambdad ~ dgamma(.001,.001)
  sigmac <- 1/sqrt(lambdac)
  sigmad <- 1/sqrt(lambdad)

}
```

$$h_i \sim Binomial(s, \theta_{hi})$$

$$f_i \sim Binomial(n, \theta_{fi})$$

$$\theta_{hi} = \Phi(\frac{1}{2}d_i - c_i)$$

$$\theta_{fi} = \Phi(-\frac{1}{2}d_i - c_i)$$

$$c_i \sim Gaussian(\mu_c, \sigma_c)$$

$$d_i \sim Gaussian(\mu_d, \sigma_d)$$

$$\mu_c \sim N\left(0, \frac{1}{.001}\right) \quad \mu_d \sim N(0, \frac{1}{.001})$$

$$\lambda_c \sim Gamma(.001, .001)$$

$$\lambda_d \sim Gamma(.001, .001)$$

$$\sigma_c = \frac{1}{\sqrt{\lambda_c}} \quad \sigma_c = \frac{1}{\sqrt{\lambda_c}}$$

Figure 4. Script for setting up the equal-variance Gaussian signal detection theory model with an extension in WinBUGS.

In summary, Davis (2015) manipulated incomplete feedback in a simple tone-in-noise detection experiment to study the effects of incomplete feedback on response bias. Davis (2015) analyzed the data carefully for each subject. Incomplete feedback indeed resulted in changes in response bias on an individual level. However, Davis (2015) did not make any inference for the population's response bias under incomplete feedback conditions. The results of an experiment should not be limited to the observers who participated in the study. Thus, an inferential data analysis is needed. The next chapter presents the results from inferential data analysis with Bayesian statistics.

Chapter 3 Reanalysis and Results

3.1 Data collected in Davis (2015)

Davis (2015) provided the HIT rate and FALSE ALARM (FA) rate averaged over 10 blocks for each subject in each condition (Table 1). HIT count and FA count were recovered from the HIT rate and FA rate. There were 500 trials in each condition, 250 of which were SN trials and 250 were NA trials. Thus, the HIT count in 250 trials is 250 multiplied by the HIT rate. The FA count in 250 trials is 250 multiplied by the FA rate. For each condition, HIT counts and FA counts for all subjects were then fed into the Markov Chain Monte Carlo sampling procedure to estimate the group-level response bias (μ_c) and sensitivity (μ_d).

Table 1. HIT rate and FA rate averaged over 10 blocks for each condition and each subject

Condition	Rate	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
NO-KR	HIT	0.69	0.45	0.61	0.45	0.72	0.67	0.49	0.44	0.59	0.39
	FA	0.17	0.08	0.06	0.09	0.28	0.08	0.22	0.07	0.21	0.06
[H]	HIT	0.63	0.65	0.71	0.67	0.78	0.71	0.64	0.66	0.74	0.50
	FA	0.23	0.41	0.13	0.15	0.27	0.33	0.42	0.30	0.52	0.11
[M]	HIT	0.67	0.64	0.51	0.59	0.66	0.53	0.67	0.56	0.67	0.56
	FA	0.20	0.45	0.11	0.30	0.20	0.18	0.28	0.45	0.31	0.07
[FA]	HIT	0.65	0.57	0.70	0.43	0.71	0.56	0.67	0.52	0.71	0.51
	FA	0.09	0.24	0.16	0.26	0.33	0.34	0.24	0.40	0.41	0.20

Continued

Table 1 Continued

Condition	Rate	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
[CR]	HIT	0.72	0.83	0.46	0.29	0.61	0.56	0.63	0.52	0.57	0.35
	FA	0.27	0.28	0.22	0.34	0.24	0.13	0.31	0.19	0.41	0.07
[H, M]	HIT	0.78	0.70	0.82	0.82	0.66	0.68	0.58	0.60	0.69	0.47
	FA	0.30	0.38	0.41	0.36	0.23	0.34	0.51	0.12	0.37	0.09
[H, FA]	HIT	0.58	0.60	0.53	0.74	0.74	0.78	0.60	0.50	0.70	0.63
	FA	0.12	0.27	0.18	0.24	0.35	0.29	0.41	0.42	0.42	0.12
[H, CR]	HIT	0.70	0.51	0.86	0.53	0.74	0.59	0.53	0.62	0.71	0.53
	FA	0.19	0.17	0.47	0.40	0.26	0.31	0.44	0.20	0.38	0.18
[H,M,FA]	HIT	0.72	0.73	0.82	0.81	0.76	0.54	0.70	0.57	0.65	0.43
	FA	0.19	0.31	0.37	0.26	0.19	0.34	0.27	0.43	0.44	0.14
All-KR	HIT	0.77	0.75	0.68	0.66	0.62	0.57	0.52	0.46	0.69	0.63
	FA	0.26	0.40	0.34	0.50	0.34	0.33	0.30	0.40	0.41	0.16

3.2 Results of inferential data analysis

Stimulus strength was kept constant throughout the experiment in terms of the individualized signal-to-noise ratio. Thus, group-level sensitivity (μ_d) is expected to be relatively stable across conditions. Response bias is assumed to be affected by non-sensory information conveyed by feedback. Thus, group-level response bias (μ_c) is expected to vary across conditions. Posterior means and 95% credible intervals for μ_c, μ_d, as well as σ_c (spread of response bias in the group) and σ_d (spread of sensitivity in the group) in all conditions are reported in Table 2. Figure 5 shows the posterior distributions of μ_c for all conditions. Figure 6 shows the posterior distributions of μ_d for all conditions. Posterior distributions of σ_c and σ_d are shown in Figure 7 and Figure 8, respectively.

Table 2 Posterior means and 95% credible intervals for μ_c, μ_d, σ_c and σ_d.

Feedback Condition	μ_c		μ_d		σ_c		σ_d	
	mean	95% CI	mean	95% CI	mean	95% CI	mean	95% CI
No Feedback	0.52	0.31, 0.74	1.31	1.07, 1.54	0.32	0.19, 0.53	0.33	0.18, 0.58
CR	0.29	0.07, 0.52	0.87	0.53, 1.21	0.34	0.21, 0.58	0.51	0.31, 0.87
FA	0.19	0.05, 0.34	0.93	0.62, 1.23	0.22	0.13, 0.36	0.46	0.27, 0.8
MISS	0.22	0.04, 0.4	0.98	0.7, 1.26	0.27	0.16, 0.46	0.41	0.24, 0.71
HIT	0.08	-0.11, 0.27	1.05	0.77, 1.32	0.28	0.17, 0.48	0.4	0.23, 0.69
HIT + CR (correct trials)	0.1	-0.11, 0.29	0.91	0.63, 1.18	0.3	0.18, 0.5	0.4	0.24, 0.69
HIT+ FA ("yes" trials)	0.12	-0.05, 0.29	0.98	0.68, 1.28	0.26	0.15, 0.44	0.44	0.26, 0.76
HIT + MISS (signal trials)	0.02	-0.21, 0.25	1.02	0.76, 1.29	0.35	0.21, 0.59	0.38	0.22, 0.65
HIT + MISS + FA	0.05	-0.12, 0.23	1.03	0.73, 1.34	0.26	0.16, 0.44	0.47	0.28, 0.79
All Feedback	0.03	-0.1, 0.16	0.77	0.5, 1.04	0.19	0.11, 0.34	0.39	0.23, 0.66

A comparison between the two extreme conditions (No Feedback and All Feedback) suggests that providing complete feedback to the subjects had effectively reduced response bias. Subjects showed a significant amount of response bias when no feedback was provided (No Feedback, $\hat{\mu}_c$ = 0.52). As can be seen from the first row of Figure 5, the posterior distribution of μ_c is far away from the unbiased position indicated by the vertical dashed line. Subjects were conservatively biased in this condition. When complete feedback was provided, response bias was reduced almost to none (All

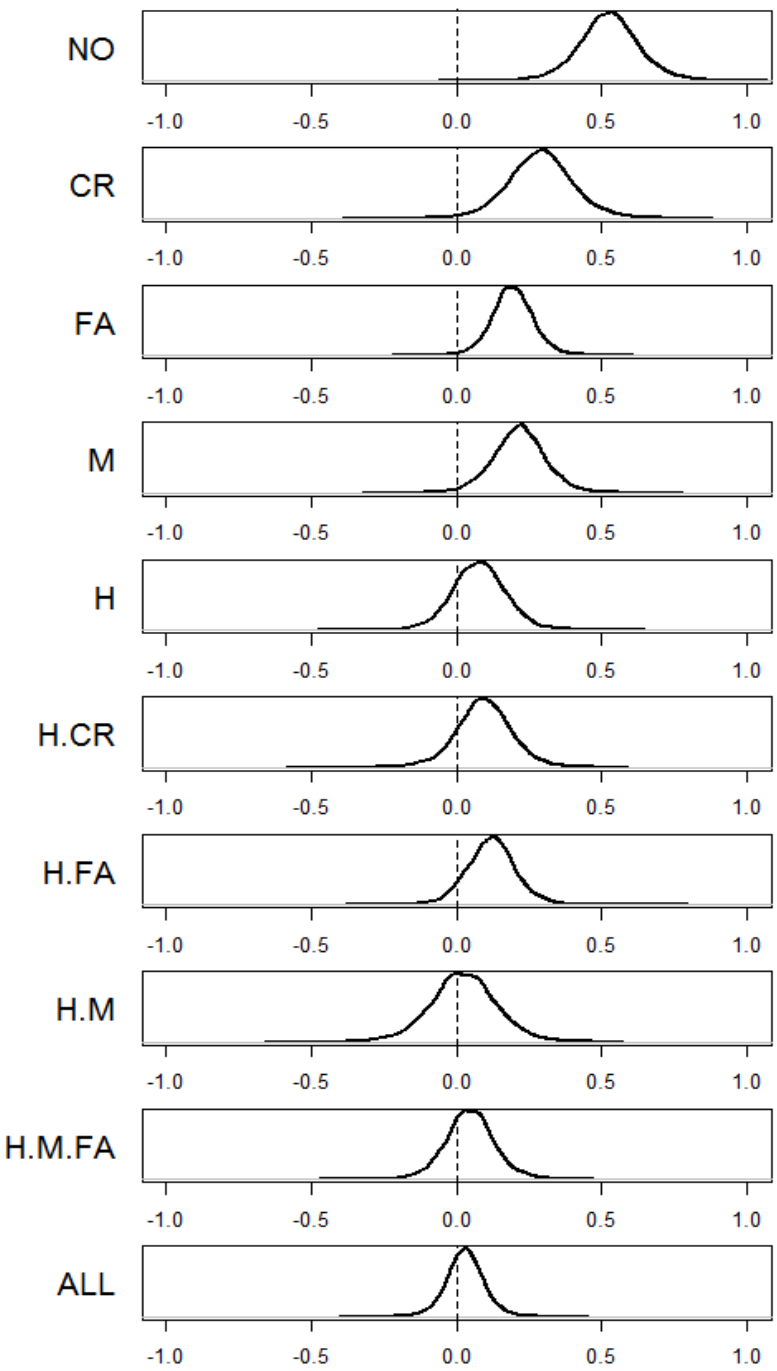

Figure 5. Posterior distributions of the mean of bias (μ_c) in all conditions. The vertical dashed line in each row shows the unbiased position. (NO-No Feedback; CR-CORRECT REJECTION; FA-FALSE ALARM; M-MISS; H-HIT; H.CR- correct trials; H.FA-"yes" trials; H.M-signal trials; All-All Feedback)

Feedback, $\hat{\mu}_c$ = 0.03). As can be seen from the last row of Figure 5, the posterior distribution of μ_c is very close to the unbiased position.

When feedback was incomplete, the response criterion in most conditions was not optimal. As can be seen from Figure 5, the posterior distributions of μ_c deviate from the unbiased position in these conditions. When feedback was provided for just one Signal-Response combination (HIT, MISS, FA, or CR), bias was reduced compared to the No Feedback condition (CR, $\hat{\mu}_c$ = 0.29; FA, $\hat{\mu}_c$ = 0.19; MISS, $\hat{\mu}_c$ = 0.22; HIT, $\hat{\mu}_c$ = 0.08; No Feedback, $\hat{\mu}_c$ = 0.52). Posterior distributions of μ_c in these conditions are much closer to the unbiased position than in the No Feedback condition. Among these four conditions, providing feedback for HIT has the greatest effect in reducing response bias. The 95% credible interval of μ_c in HIT condition contains 0.

Providing feedback for signal trials (HIT + MISS) further reduces bias almost to none ($\hat{\mu}_c$ = 0.02). Providing feedback for "yes" trials (HIT + FA) does not reduce bias further (HIT + FA, $\hat{\mu}_c$ = 0.12). Providing feedback for correct trials has a similar effect (HIT + CR, $\hat{\mu}_c$ = 0.1). Providing feedback for HIT, MISS and FA reduces bias almost to none, although the effect is not as great as providing feedback for signal trials only (HIT + MISS+ FA, $\hat{\mu}_c$ = 0.05; HIT + MISS $\hat{\mu}_c$ = 0.02).

As can be seen from Figure 6, most conditions show relatively similar sensitivities except the first and last conditions. For the eight conditions that were randomly presented, posterior means of μ_d are between 0.87 and 1.05. Sensitivity in the first condition, No Feedback, is the highest among all conditions ($\hat{\mu}_d$ = 1.31). Sensitivity in the last condition, All Feedback, is the lowest among all conditions ($\hat{\mu}_d$ = 0.77). Davis

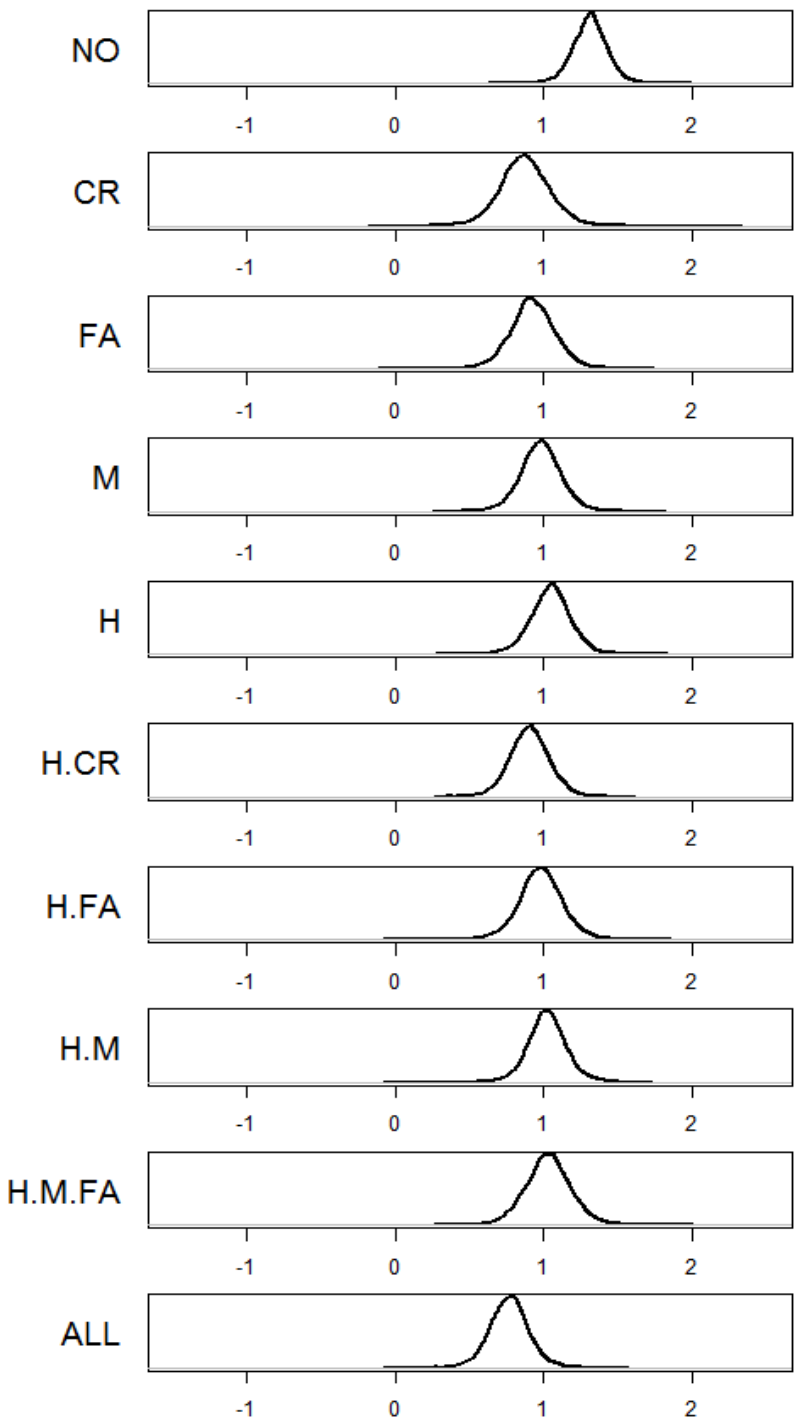

Figure 6. Posterior distributions of the mean of sensitivity (μ_d) in all conditions (NO-No Feedback; CR-CORRECT REJECTION; FA-FALSE ALARM; M-MISS; H-HIT; H.CR-correct trials; H.FA-"yes" trials; H.M-signal trials; All-All Feedback)

(2015) reported that most subjects (9 out of 10) showed sensitivity above 1 in the first condition, while most subjects (8 out of 10) showed sensitivity below 1 in the last condition (many were around 0.7 and one was as low as 0.16).

The estimated spread of the parent distribution of response bias ($\hat{\sigma}_c$) in each condition is reported in the fifth column of Table 2. It is between 0.19 and 0.35. The All-Feedback condition shows the smallest spread among all conditions ($\hat{\sigma}_c$ = 0.19), indicating that response bias varies the least across subjects when feedback was complete. The estimated spread of the parent distribution of sensitivity ($\hat{\sigma}_d$) in each condition is reported in the sixth column of Table 2. It is between 0.33 and 0.51. The variability of sensitivity among subjects is similar across conditions.

In summary, response bias varied across conditions while sensitivity was relatively constant. When there was no feedback, subjects showed a substantial amount of response bias toward "No". When feedback was complete or when feedback was provided for signal trials, response bias was reduced almost to none. Chapter 4 will discuss how feedback might convey non-sensory information (*a priori* probabilities and costs and values of responses) to the subjects and how response bias was influenced accordingly.

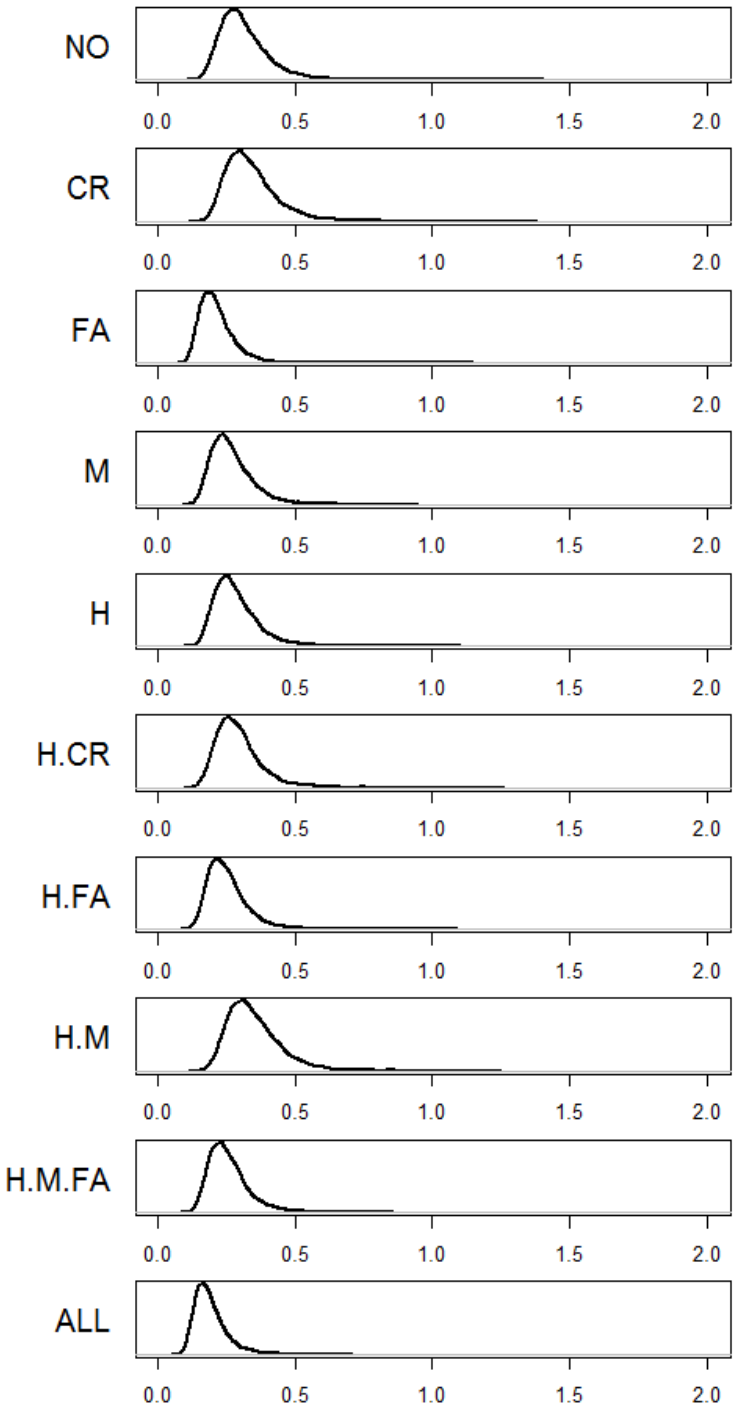

Figure 7. Posterior distributions of the standard deviation of bias (σ_c) in all conditions (NO-No Feedback; CR-CORRECT REJECTION; FA-FALSE ALARM; M-MISS; H-HIT; H.CR- correct trials; H.FA-"yes" trials; H.M-signal trials; All-All Feedback)

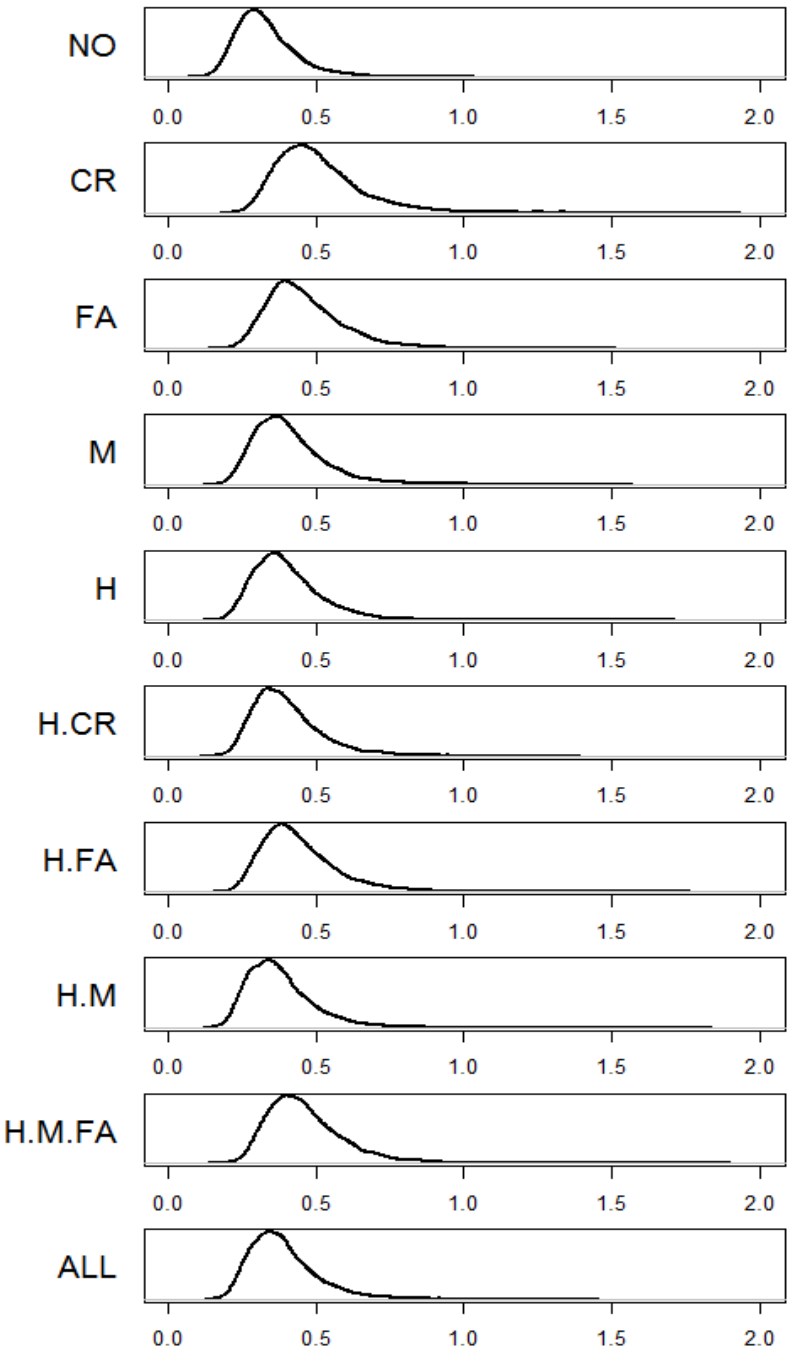

Figure 8. Posterior distributions of the standard deviation of sensitivity (σ_d) in all conditions (NO-No Feedback; CR-CORRECT REJECTION; FA-FALSE ALARM; M-MISS; H-HIT; H.CR- correct trials; H.FA-"yes" trials; H.M-signal trials; All-All Feedback)

Chapter 4 Discussion and Conclusions

4.1 Effects of Incomplete Feedback on Response Bias

To maximize performance in a long series of observations (specifically, to maximize expected values), recall that the optimal likelihood ratio criterion, β_{opt}, is given by the equation

$$\beta_{opt} = \frac{P(NA)}{P(SN)} \frac{[V(CR) + C(FA)]}{[V(H) + C(M)]}$$

where V(CR) is the value of a CORRECT REJECTION, C(FA) is the cost of a FALSE ALARM), V(H) is the value of a HIT, and C(M) is the cost of a MISS. Values and costs are all entered as positive numbers. If costs and values are irrelevant and the goal is to simply maximize the proportion of correct responses, the optimal likelihood ratio criterion is simplified to $\beta_{opt} = \frac{P(NA)}{P(SN)}$ (Green and Swets, 1966, pp. 23). For the auditory detection task in Davis (2015), values of correct responses and costs of incorrect responses were not specified numerically or as monetary gains or losses. Subjects were simply asked to "get as many correct answers as possible". Therefore, the optimal criterion for the experiment is simply $\beta_{opt} = \frac{P(NA)}{P(SN)}$. The ratio of the *a priori probabilities* is 1, so β_{opt} = 1, which is also the unbiased response criterion.

Since subjects were not told before the experiment that SN and NA events were equally likely to occur, they could have assumed that P(SN) and P(NA) were equal, or

that SN was more likely to occur than NA, or the other way around. Thus, the subjective *a priori* probabilities ratio could be equal to, less than, or greater than 1. If there is no feedback, it is up to each subject to decide whether to behave conservatively, liberally, or neutrally. When complete feedback is available, subjects could learn that the *a priori* probabilities P(SN) and P(NA) were equal from the feedback. That is, an indication of "HIT" or "MISS" indicates that a SN trial has occurred; a "FALSE ALARM" or "CORRECT REJECTION" indicates that a NA trial has occurred. When feedback is incomplete, subjects may, or may not, gain an accurate estimate of the *a priori* probabilities.

A visual display was used to indicate whether a response was correct or not. A check and a green response button were used to convey a HIT or a CORRECT REJECTION. A cross and a red response button were used to convey a MISS or a FALSE ALARM. With complete visual feedback, the values of correct responses and costs of incorrect responses were reinforced equally. With incomplete feedback, only some values or costs were reinforced. This might have result in an imbalanced internal payoff matrix. That is, the value of a HIT, the value of a CORRECT REJECTION, the cost of a MISS and the cost of a FALSE ALARM could be considered unequal by the observers. From the perspective of an observer, the decision goal might be not just to maximize the number of correct responses, but also to maximize the expected "value". The observers might have used $\frac{P(NA)}{P(SN)}\frac{[V(CR)+C(FA)]}{[V(H)+C(M)]}$ as the optimal response criterion.

The magnitude of $\frac{[V(CR)+C(FA)]}{[V(H)+C(M)]}$ might increase or decrease when some decision outcomes are reinforced by feedback. When HITs were reinforced, subjects knew the

occurrence of a HIT, but did not know when a MISS, a FALSE ALARM or a CORRECT REJECTION had occurred. In this case, the value of a HIT might have been reinforced, and the magnitude of the ratio $\frac{V(CR)+C(FA)}{V(H)+C(M)}$ might have been decreased, because part of the denominator was increased (indicated by the blue text color).

Subjects were very conservative in the first condition where no feedback was available. The estimated response bias was positive. This indicates that the corresponding likelihood ratio criterion is greater than 1. That is, the product of the unknown subjective ratio of costs and values and the unknown subjective ratio of *a priori* probabilities is greater than 1 ($\frac{P(NA)}{P(SN)}\frac{V(CR)+C(FA)}{V(H)+C(M)} > 1$). This indicates that subjects might have expected NA trials occur more often than SN trials, and/or the cost of a FALSE ALARM and the value of CORRECT REJECTION is greater than the cost of a MISS and the value of a HIT. It is more probable that the subjective ratio of *a priori* probabilities was the dominant factor since no visual display of green check or red cross was shown to the subjects at all in this condition. The conservative behavior might just be a characteristic of this group of subjects. This group of subjects was a convenient sample. They had a lot of experience in other auditory detection experiments. If a different sample of subjects were recruited from a more general population, liberal or neutral behavior might have been observed.

When complete feedback was provided in the last condition, the subjective ratio of *a priori* probabilities $\frac{P(NA)}{P(SN)}$ might be close to 1 for each subject. Subjects could tell whether an SN event occurred or not because feedback occurred on every trial. Subjects

might have cumulated this information over time and learned the equal *a priori* probabilities of SN and NA events. Also, when feedback was complete, the subjective magnitude of $\frac{V(CR)+C(FA)}{V(H)+C(M)}$ might be close to 1, since the values and costs of all decision outcomes were equally emphasized. Both ratios were close to 1, driving the response criterion toward the unbiased position.

It is quite interesting that the response criterion was almost at the unbiased position when feedback was provided for "signal" trials (HIT and MISS). In this condition, signals and feedback occurred together, subjects could learn the *a priori* probabilities by comparing the number of trials with feedback and the number of trials with no feedback, or at least they could tell that SN trials were not as rare as they might have expected at the beginning of the experiment. Thus, the subjective ratio of $\frac{P(NA)}{P(SN)}$ might be still greater than 1, but not as large as that when no feedback was available. The subjective ratio of $\frac{V(CR)+C(FA)}{V(H)+C(M)}$ might be less than 1, because HIT trials and MISS trials were reinforced by the visual feedback, and subjects might have thought that the value of a HIT and the cost of a MISS were greater than the value of a CORRECT REJECTION and the cost of a FALSE ALARM. Taken together, the product of the two ratios might be close to 1.

When feedback was provided for "correct" trials (HIT and CORRECT REJECTION), subjects could have learned that the *a priori* probabilities were equal, but it was not as easy as when feedback was provided for "signal" trials. To learn how likely an SN trial occurs in this condition, subjects must add the number of HITs ("yes"

responses with feedback) and the number of MISSes ("no" responses without feedback). Similarly, to learn how likely a NA event occurs, subjects must add the number of CORRECT REJECTIONs ("no" responses with feedback) and the number of FALSE ALARMs ("yes" responses without feedback). Keeping track of these sums would not be not easy, especially when the task is to pay attention to a weak signal embedded in noise. This may explain why the response criterion in this condition was not as unbiased as when feedback was provided for "signal" trials. No speculation can be made for the impact of proving feedback for correct trials on the ratio $\frac{V(CR)+C(FA)}{V(H)+C(M)}$ in this condition, since part of the numerator is increased and part of the denominator is increased.

When feedback was provided for "yes" trials (HIT and FALSE ALARM), information about the *a priori* probabilities was not complete. It was not possible to count the number of SN trials or NA trials in a block. One must know either the number of HITs and MISSes, or the number of FALSE ALARMs and CORRECT REJECTIONs, to learn the *a priori* probabilities. This may explain why the response criterion in this condition was not as unbiased as when "signal" trials received feedback or when all trials received feedback. Again, no speculation can be made for the impact of providing feedback for "yes" trials on the ratio $\frac{V(CR)+C(FA)}{V(H)+C(M)}$, since one element in the numerator and one element in the denominator were reinforced.

It may be interesting to compare response biases when only HIT received feedback and when only CORRECT REJECTION received feedback. When HIT received feedback, an observer could count the number of a large portion of SN trials. This would correct the observer's initial belief that SN trials may be rare. Similarly, when

CORRECT REJECTION received feedback, an observer could count the number of a large portion of NA trials. This would correct the observer's initial belief that NA trials are very common. Thus, providing feedback for HIT and providing feedback for CORRECT REJECTION may be considered as providing similar information about the *a priori* probabilities. The subjective ratio $\frac{P(NA)}{P(SN)}$ should be similar in these cases. However, providing feedback for CORRECT REJECTION leads to a more conservative bias than providing feedback for HIT. Knowing when a CORRECT REJECTION occurs might have emphasized the value of a CORRECT REJECTION, which increased the ratio $\frac{V(CR)+C(FA)}{V(H)+C(M)}$, resulting in a greater (more conservative) likelihood ratio criterion. Knowing when a HIT occurs might have emphasized the value of a HIT, which decreased the ratio $\frac{V(CR)+C(FA)}{V(H)+C(M)}$, resulting in a smaller (less conservative) response criterion.

Above all, the effect of incomplete feedback on response bias in the experiment of Davis (2015) could be interpreted in terms of how feedback conveys information about *a priori* probabilities of SN and NA trials and the relative importance of decision outcomes to the subjects. When information was complete, observers were able to achieve the optimal response criterion associated with the setting of the experiment (equal *a priori* probabilities and a balanced payoff matrix), which was also the unbiased criterion. When information was incomplete, the criterion used by the observers deviated from the unbiased criterion. It seems the observers maintained an optimal criterion which was determined by the available information.

4.2 Sensitivity

In general, sensitivity was relatively stable across 8 out of the 10 conditions, indicating that the internal representation of the presence or absence of the signal was controlled well by the individualized signal-to-noise ratios. Feedback did not have a substantial effect on sensitivity. However, it was not completely clear why this group of subjects showed higher sensitivity in the first condition and lower sensitivity in the last condition. This may have been a decrease of sensitivity over time due to fatigue, but one may also question why it was not the case that sensitivity increases as subjects become more experienced with the stimuli. In a future study, one may want to present all experimental conditions in random order instead of fixing the condition with no feedback at the beginning and the condition with complete feedback at the end.

4.3 Summary and Conclusions

Subjects showed little or no response bias when complete feedback was provided, while the amount of bias was significant when there was no feedback, as expected. Once some feedback was provided, although incomplete, response bias was reduced significantly. For this group of subjects, providing feedback for HIT reduced bias more than for providing feedback for MISS, FALSE ALARM and CORRECT REJECTION. When feedback was incomplete, response bias was not reduced as much as when feedback was complete, except when feedback was provided for signal trials (HIT and MISS). Providing feedback for signal trials was as effective as providing feedback for all decision outcomes in reducing response bias for this group of subjects. Perhaps the result

of the current study is uniquely defined by the conservative character of this group of subjects. A larger sample might yield subgroups with liberal or neutral pre-testing bias allowing a study to determine the effects of incomplete feedback on those subgroups.

The issue in real-world signal detection tasks highlighted by the study of incomplete feedback, is that information about the environment available to the observer may be limited. which could result in a ceiling for the observer's behavior. In the experiment of Davis (2015), feedback carries information about the *a priori* probabilities as well as costs and values of correct and incorrect responses. When part of the feedback was taken away, subjects could not respond without bias to optimize their long-term performance.

Together, *a priori* probabilities, values of correct responses and costs of incorrect responses (the payoff matrix), and the amount and type of feedback could all vary in real-world listening situations or other signal detection scenarios. *A priori* probabilities of different states of the world could be unequal. Values and costs for different responses may be unbalanced. Feedback for certain responses may or may not be accessible. These could all affect an observers' response criterion which drives his/her performance in a signal detection task in the real world. When designing a laboratory signal detection experiment, non-sensory factors mentioned above should be incorporated into the design to improve the external validity of the experimental results.

4.4 A brief proposal for a follow-up study

The results of the current study and theory of signal detection suggest that response bias exists (response criterion is not optimal) when feedback is incomplete. One may be curious about whether response bias could be trained out of a listener who must operate in situations for which incomplete feedback is the norm. An experiment could be designed to test this hypothesis. Below is a brief proposal for the experiment.

The simple tone-in-noise detection task from Davis (2015) will be used. SN and NA trials will be set equally likely to occur. For this experiment, let us specify the values of HIT and CORRECT REJECTION as monetary gain of 5 dollars and costs of MISS and FALSE ALARM as monetary loss of 5 dollars. Listeners will be recruited and randomly assigned to the control group and the treatment group. Individualized signal-to-noise ratio will be determined for each listener in the same way as described in Davis (2015). Listeners will go through a familiarization task as that in Davis (2015) to get familiar with the stimuli and the psychophysical procedure. Listeners in the treatment group will receive an additional training session. The purpose of the training session is to reinforce the non-sensory information that will direct response criterion to the unbiased position according to theory of signal detection. During the training session, listeners in the treatment group will be told explicitly that SN and NA trials are always equally likely to occur. They will also be told that each correct response results in a monetary gain of 5 dollars and each incorrect response results in a money loss of 5 dollars. To emphasize this information, listeners will be required to practice with several blocks of trials. A visual display will show them whether the signal was present or absent. After a response,

a counter will show the monetary gain/loss. After the training session, a test session will be given to both groups of listeners. The experimental conditions will be the same as those in Davis (2015). The only difference is that the visual display will show the total value at the end of a block. Finally, performance will be compared between the two groups.

If listeners in the treatment group show less response bias than listeners in the control group, this suggests that non-sensory information could be reinforced and fixed in advance by training. Later, subjects can keep the unbiased response criterion even if they have to operate in situations where incomplete knowledge of results is the norm. If listeners in both groups show similar results, this indicates that it is difficult to fix non-sensory information by training, and response bias will continue to be influenced significantly by incomplete knowledge of results.

Bibliography

Aziz, H. (2017). Comparison between field research and controlled laboratory research. *Archives of Clinical and Biomedical Research*, 1(2), 101-104

Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical transactions of the Royal Society of London*, (53), 370-418.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101-1109.

Coppock, A., & Green, D. P. (2015). Assessing the correspondence between experimental results obtained in the lab and field: A review of recent social science research. *Political Science Research and Methods*, *3*(1), 113.

Davis, M. J. (2015). *The effects of incomplete knowledge of results on response bias in an auditory detection task* (Doctoral dissertation, The Ohio State University).

Gescheider, G. A. (2013). *Psychophysics: the fundamentals*. Psychology Press.

Giguère, C., Laroche, C., Osman, A., & Zheng, Y. (2008, July). Optimal installation of audible warning systems in the noisy workplace. In *Proceedings of the 9th International Congress on Noise as a Public Health Problem (ICBEN)* (pp. 197-204).

González-Hernández, J. M., Peral-Orts, R., Campillo-Davo, N., Poveda-Martínez, P., Campello-Vicente, H., & Ramis-Soriano, J. (2017). Assessment of warning sound detectability for electric vehicles by outdoor tests.

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. 1988 reprint edition.

Harvey Jr, L. O. (2014). Detection theory: sensory and decision processes. *University of Colorado, Boulder*.

Kaernbach, C. (1990). A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. *The Journal of the Acoustical Society of America*, *88*(6), 2645-2655.

Lee, M. D. (2008). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods*, *40*(2), 450-456.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lurie, N. H., & Swaminathan, J. M. (2009). Is timely information always better? The effect of feedback frequency on decision making. *Organizational Behavior and Human decisión processes*, *108*(2), 315-329.

Mackworth, N. H. (1956). Vigilance. *Nature*, *178*(4547), 1375-1377.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Parizet, E., Robart, R., Chamard, J. C., Schlittenlacher, J., Pondrom, P., Ellermeier, W., ... & Hatton, G. (2013, June). Detectability and annoyance of warning sounds

for electric vehicles. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, No. 1, p. 040033). Acoustical Society of America.

Rane, D., Shirodkar, P., Panigrahi, T., & Mini, S. (2019, March). Detection of Ambulance Siren in Traffic. In *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)* (pp. 401-405). IEEE.

Rayo, M. F., Patterson, E. S., Abdel-Rasoul, M., & Moffatt-Bruce, S. D. (2019). Using timbre to improve performance of larger auditory alarm sets. *Ergonomics*, *62*(12), 1617-1629.

Ripley, B. D. (2001). The R project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, *1*(1), 23-25.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic bulletin & review*, *12*(4), 573-604.

Smith, S. E., Stephan, K. L., & Parker, S. P. (2004). *Auditory warnings in the military cockpit: A preliminary evaluation of potential sound types* (No. DSTO-TR-1615). DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION EDINBURGH (AUSTRALIA) AIR OPERATIONS DIV.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS user manual.

Swets, J. A. (1977). Signal detection theory applied to vigilance. In *Vigilance* (pp. 705-718). Springer, Boston, MA.

Szalma, J. L., Hancock, P. A., Warm, J. S., Dember, W. N., & Parsons, K. S. (2006). Training for vigilance: Using predictive power to evaluate feedback effectiveness. *Human factors*, *48*(4), 682-692.

Szalma, J. L., Parsons, K. S., Warm, J. S., & Dember, W. N. (2000, July). Continuous vs. Partial Knowledge of Results in Training for Vigilance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, No. 21, pp. 3-386). Sage CA: Los Angeles, CA: SAGE Publications.

VandenBos, G. R. (2007). *APA dictionary of psychology*. American Psychological Association.

Watson, C. S., & Nichols, T. L. (1976). Detectability of auditory signals presented without defined observation intervals. *The Journal of the Acoustical Society of America*, *59*(3), 655-668