Hierarchical Bayesian Modeling of Criterion Variance in Probabilistic Categorisation as an Analogue to Signal Detection.

A thesis presented in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

at Massey University, Manawatū,

New Zealand.

Robert Taylor

2015

Abstract

Variance in the decision criterion across trials induces response inconsistencies which in turn result in suboptimal performance. Criterion variability is largely thought to be driven by internal mechanisms; however, factors external to the observer may also affect response consistency. Specifically, how trial-by-trial feedback is delivered can influence the stability of the criterion across trials. This thesis examined how two types of feedback (stochastic and deterministic) influenced performance in probabilistic categorization tasks, which served as analogues to the orthodox detection task. Critically, feedback that is related to the statistical properties of the stimulus distributions (i.e., feedback for which event had occurred) results in lowered performance when compared to feedback that is provided deterministically (i.e., relative to the optimal cut-off). This result held more consistently in conditions where there was greater (probabilistic) confusability among the stimuli. The effects upon the criterion were also examined by comparing dynamic signal detection models that allowed for trial-by-trial criterion shifts. Hierarchical Bayesian modeling was implemented to fit the dynamic criterion models, allowing for model comparisons to proceed using Bayes Factors. It was found that simple error-correcting models predicted the data less well than models that included shifts after correct decisions. However, criterion shifts after correct decisions can be better described by a weighted moving average criterion which shifts toward the current stimulus, rather than away. This finding arose through the explicit modeling of the stimulus magnitudes on each trial. Finally, a model was contrived that both allowed stimulus magnitudes to influence criterion shifts and make the effects of feedback more overt. The model suggests that the way feedback information is stored over trials drives shifts in the criterion, and that feedback will influence how storage is facilitated. However, the model could not completely describe the effects of feedback nor fit the empirical data as well as already established dynamic criterion models.

Acknowledgements

I would first like to acknowledge my primary supervisor John Podd for his tireless support, encouragement, and guidance. John has been supervising my research for the better part of six years now and has been a rock in my research career. It saddens me that I will be one of his last PhD students and I feel Massey University is losing one of its quantitative stalwarts as he moves into semi-retirement. I am incredibly lucky to have had John looking over my shoulder for as long as he has, and as my research career continues onward I hope that he still will. John provided me with an unprecedented level of freedom that allowed me to follow my nose, often aimlessly at times, and indulge my curiosity in mathematics and computational modeling. This freedom, however, did not come without a cost. The eternal frustration with trying to get my head around learning R, writing objective functions, implementing and simulating models, and all the associated mathematics, often left me disheartened, to the point of giving up. However, his encouragement and enthusiasm with every new skill that I mastered allowed me to immerse myself in these fields and provided me with perhaps the best possible environment in which I could teach myself. Moreover, when I, rather abruptly, decided that I should jump the ditch to try to establish myself amongst the modeling community in Australia, John went with it. I simply would not be in the position I am today, nor be the researcher that I am, without John, and I cannot thank him enough for all the effort and time he has put into me, and my research.

I would also like to thank Chris Donkin at the University of New South Wales for providing me not only with a job over the latter stages of my thesis, but also for his guidance, mentoring, and willingness to take on such a “green” modeler. The exposure to the host of new methods, techniques, and people that Chris allowed me completely transfigured this thesis. I eagerly sought to apply everything that I had learnt – which involved a last minute rewrite so that all modeling was Bayesian – and so I hope that this thesis attests to my abilities as they presently stand; though as a modeler I feel I still have much to learn and a long way to go. The modeling undertaken in this work just would not have been possible had Chris not taken me under his wing, and I am eternally grateful for the opportunity Chris offered me.

Thanks also need to be given to my co-supervisor Stephen Hill. Stephen has overseen my research activities, also for the better part of six years, and has always provided a distant, but discerning, guidance – usually in the form of pointing out things that I have missed or simply not thought about. Stephen is also one half of the duo that are responsible for my love of experimental psychology, after taking John and Stephen's Experimental Psychology paper in third year. I am thankful to have had Stephen in the wings, and for his insights.

Finally, I would like to acknowledge my family, my friends and colleagues, and everyone that has supported me at some point during this ten year endeavour. For being more than understanding about my periodic absenteeism; letting me work, often obsessively and unabated, though always providing me with food, coffee, and beer when needed. For entertaining my frustrations, rants, and irrational perceptions about my work, and offering me nothing but encouragement and approbation in return. For providing me a creative outlet where I could play music and take my mind off my thesis, even if only for a few hours. For riding bikes, climbing mountains, camping out, and all the antics that made sure I realised that life was not completely about work. For being there to celebrate, lament, and support me through all the challenging life events that occurred during the past ten years. For picking me up and dusting me off when I wanted to quit, and for simply knowing when not to ask "how's the thesis going?" But most of all, thank you all for just believing in me; every last ounce was needed. I would not have completed this work without it, and I hope I have done you all proud.

This Thesis has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher named is responsible for the ethical conduct of this research.

Table of Contents

List of Figures

List of Tables

Chapter 1

An Introduction to Signal Detection Theory

For over 50 years signal detection theory (SDT) has provided a model with which one can analyse and interpret human decision-making under conditions where the evidence available to do so is uncertain, complex, or incomplete. Historically, much of the early work on SDT was conducted during the Second World War where advances were being made with the use of radio waves in the detection and ranging of objects, or RADAR. During the 1950's further advances in engineering and signal processing extended RADAR's practical applications, and in 1954 Peterson, Birdsall, and Fox (see also Van Meter & Middleton, 1954, as cited in Swets, 1961) published a unified theory of signal detectability, or signal detection theory as it is known today. During the same period detection principles were being applied in both auditory (Tanner, Swets, & Green, 1955, as cited in Swets, 1961; Munson & Karlin, 1956) and visual (Swets, Tanner, & Birdsall, 1955, as cited in Swets, 1961; Tanner & Swets, 1954) psychophysical experiments, and in 1966 SDT was formally incorporated as a psychological theory with Green and Swets' publication, *Signal Detection Theory and Psychophysics*.

1.1. A Primer on Signal Detection Theory

The psychophysical theories of antiquity (e.g., Fechner, 1860) were predominantly concerned with empirically determining a sensory threshold – that is, the magnitude a stimulus must reach in order for it to be detected 50% of the time. However, dissatisfaction arose with the classical methodologies (e.g., method of limits, method of constants) in their inability to account for biases in responding. This limitation unfortunately meant that estimates of sensitivity were potentially confounded with response bias. The absence of catch trials (i.e., no signal present) meant that no elucidation between genuine detections or guesses could be made. Simply put, it is not enough to just report whether a stimulus has been detected, but the nature of the report, which includes any inherent biases, must also be considered. While the sensory aspects of the detection task remain intrinsic, SDT views the reporting of an event as a critical determinant of performance, and explicitly considers both the sensory and decisional contributors to task performance.

Knowing that uncertainty is axiomatic in any detection environment, SDT combined detection principles with statistical decision theory (Wald, 1950). SDT views the detection task as comprising of two internal processes: the first concerns perceptual processes that determine the ability of the observer to distinguish between stimulus events – the *sensory stage*; the second process concerns the decision-making abilities of the observer– the *decision stage*. In its most basic application SDT requires an observer to discriminate between two possible classes of event, though it may be applied to situations where discrimination between multiple events is required. Conventionally, observers are presented with a series of trials where either noise alone or a tone in noise is present. The sensory stage requires the transduction of a sensory input, where the perceptual effect must be registered by the appropriate sensory system. However, random perturbations in the perceptual effect across trials induce confusability in the evidential value of the effect. Accordingly, the perceptual effect is assumed to be distributed along an internal psychological continuum – referred to as the *decision axis* – for each stimulus class (Figure 1.1).

A simplifying assumption made by SDT is that the perceptual effect, or evidence[1], is distributed as a random Gaussian variable , $X_i \sim N(\mu, \sigma)$, though other distributional forms can justifiably be used (e.g., Uniform, Poisson, Exponential, Weibull, and Cauchly; see DeCarlo, 1998; Green & Swets, 1966; Knoblauch & Maloney, 2012). The probability density functions (PDF) for each event class are expressed as follows:

$$f(x|i) = (2\pi\sigma_i^2)^{-\frac{1}{2}} e^{\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]}, \qquad \text{Eq. 1.1}$$

where μ_i and σ_i are the mean and the standard deviation of the i^{th} evidence distribution, and $i = \{0, 1\}$ refer to the noise and signal distribution, respectively[2]. For the case considered here the

[1] Perceptual effect and perceptual evidence may be used interchangeably. For consistency the distributions will be referred to as evidence distributions.

[2] This thesis only concerns the binary detection case, and so the convention will be adopted where each event is a dummy coded variable. Accordingly, all stimulus events and associated responses, R_j, will be coded as such.

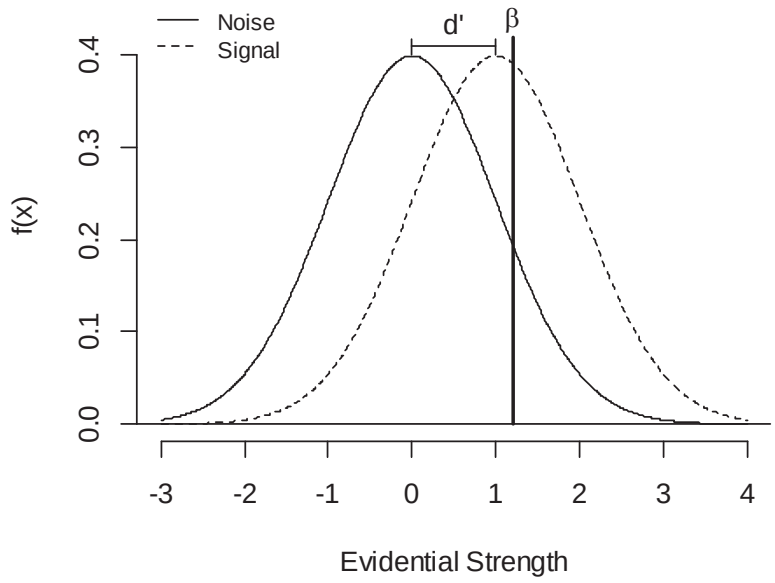

Figure 1.1: The equal-variance SDT model. Two overlapping Gaussian distributions are assumed to represent variability in the perceptual information associated with each event class. The distance between the means of the distributions determines how detectable the stimulus events are, with the signal event assumed to have a greater sensory mean. A decision criterion (β) is located along the decision axis that partitions the sensory space into response regions.

distributions are also assumed to have equal variances ($\sigma_0^2 = \sigma_1^2 = 1$). A signal event is expected to increase the mean sensory effect by some fixed amount. The distributional mean for signal events is assumed to be greater than for noise alone, where the mean of the noise alone distribution conventionally remains fixed at $\mu = 0$. The degree of separation between the means - called *d*-prime (d') - establishes the signal-to-noise ratio (SNR) and determines how detectable the signal is. Thus, the parameters of the signal and noise distributions completely define the actual limit of performance (i.e., asymptotic performance), quantified by calculating the standardised distance between the distributional means (scaled in σ_0 units):

$$d'_{actual} = \frac{\mu_1 - \mu_0}{\sigma_0}. \qquad \text{Eq. 1.2}$$

Complete distributional overlap ($d'_{actual} = 0$) indicates no discrimination ability and results in only chance performance. As the signal mean shifts further rightward (as $d'_{actual} \longrightarrow \infty$) signal detection increasingly becomes easier.

The decision stage, then, requires some response to be made based upon noisy perceptual evidence. On each trial the observer is required to specify whether the signal was present or not, responding with either a "yes" or a "no" - the *yes/no task*[3]. SDT assumes the observer locates a fixed criterion along the decision axis that partitions the sensory space into response regions. The observer is then assumed to compute a single statistic that quantifies the evidential weight for a specific response alternative – the *decision variable*- and finally, compare the decision variable with the criterion value. A response is completely determined by which side of the criterion the decision variable falls; a simple deterministic choice rule of the form

$$respond \begin{cases} "No", & X_i < k \\ "Yes", & X_i \geq k \end{cases}, \qquad \text{Eq. 1.3}$$

where k denotes the criterion. The characteristics of both the decision variable and criterion reflect an assumption made about the nature of the underlying decision axis. For example, the decision axis may reflect sensory strength, and so assuming some amount of perceptual evidence on a given trial (x_n), the observer may weigh the relative strength of x_n against the criterion (the sensory criterion, c). However, Green and Swets (1966) suggest that the most appropriate decision variable is the likelihood ratio, $\lambda(x_n) = f(x_n|s)/f(x_n|n)$, or some monotonic transform thereof (e.g., $\log \lambda(x_n)$). The observer may consider the likelihood that x_n originated from either the signal or noise distribution and weigh this against the criterion (the likelihood criterion, β). The likelihood is simply the ratio between the ordinates of each probability density at the value of x_n. The likelihood criterion specifies a critical likelihood ratio at a fixed value of x, and so $\beta = \lambda(x)$.

[3] Alternative formats may require observers to rate how certain they are that a signal was presented or not - the *rating task*. However, the foregoing is only concerned with the binary response task and so no further discussion of response formats will be pursued.

The choice of decision rule will also depend upon the values (V_x) associated with each decision. Generally, correct decisions have positive values associated with them (V_{Hit}, V_{CR}), whereas incorrect decisions have negative values attached to them (V_{FA}, V_{Miss}). SDT can specify the optimal decision criterion (β_o) in terms of both the values associated with each decision and the prior probability of a signal event (π), where

$$\beta_o = \frac{1-\pi}{\pi}\left[\frac{V_{CR} + V_{FA}}{V_{Hit} + V_{Miss}}\right]. \qquad \text{Eq. 1.4}$$

For general purposes the typical decision rule is the one that maximises the percentage of correct decisions. For the case considered in the present example it is assumed that there are no differential values associated with each decision. Further, the priors for each stimulus event are assumed to be equal ($\pi = .5$), and so the optimal decision criterion is simply the reciprocal of the prior odds, $\beta_o = (1-\pi)/\pi = 1$. It is also the point at which the ordinates of the signal and noise distributions cross (see Figure 1.1). The optimal criterion also specifies the neutral point; that is, the point where responding is unbiased. Deviations in the location of the criterion from the optimal point suggest biases in the way an observer responds (the observer is considered unbiased if $\beta = 1$). If $\beta > \beta_o$ the observer is considered to be *conservative,* having a strict criterion (biased toward "no" responses). The converse case, where $\beta < \beta_o$, suggests the observer is *liberal,* having a lax criterion (biased toward "yes" responses). Equation 1.3 leads to four possible trial outcomes (Table 1.1).

Table 1.1:
Response Contingencies for the Yes/No Response Task Conditional Upon Signal (S) and Noise (N) events.

	Response	
Event	Yes	No
Signal	Hit	Miss
	$HR = P(Yes\|S)$	$1 - HR$
Noise	False Alarm	Correct Rejection
	$FAR = P(Yes\|N)$	$1 - FAR$

The rates associated with each decision outcome can be determined by calculating the relative proportion of responses made to each stimulus event; more formally these are conditional posterior probabilities. Conventionally, the rates of interest are the hit rate (HR) and the false alarm rate (FAR); the proportion of "yes" responses made on signal and noise trials, respectively. The HR and FAR will vary together for different criterion values; for example, a lax criterion will increase the HR but will also increase the FAR. Conversely, a strict criterion will reduce the FAR but the HR will also decrease. The concomitant changes in the HR and FAR reflect the change in area to the right of the criterion beneath the signal and noise distributions, respectively (Figure 1.2a); these areas are proportional to

$$\theta_i = 1 - \left[\mathbf{\Phi}\left(\frac{k - \mu_i}{\sigma_i}\right)\right] = 1 - \left[\int_{-\infty}^{k} f(x \mid i)\, dx\right], \qquad \text{Eq. 1.5}$$

where θ_i is the rate associated with the number of "yes" responses conditional upon the i^{th} distribution, $\mathbf{\Phi}$ is the CDF of the standard normal curve, and k denotes the criterion location.

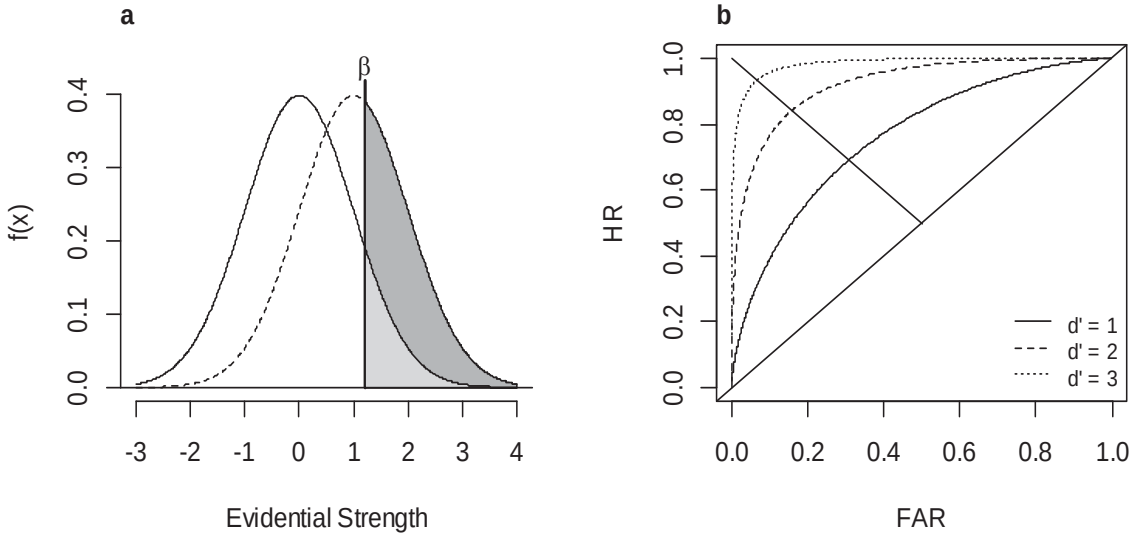

Figure 1.2: a) The area to the right of the criterion beneath the signal distribution (dark grey) corresponds to the HR, whereas the area to the right of the criterion beneath the noise distribution (light grey) corresponds to the FAR; **b)** the ROC curve which plots the trade-off between the HR and FAR for varying criterion locations. The area beneath the ROC curve is relative to the distance between the means of the signal and noise distributions. The minor diagonal reflects the neutral point, meaning that bias can be determined by assessing where the binary ROC point falls in relation to the minor diagonal.

The trade-off between the HR and FAR can be visually inspected by plotting the HR as a function of the FAR for every possible criterion value given a fixed level of d'_{actual} - the *Receiver Operating Characteristic* (ROC) curve (Figure 1.2b). In the binary response task only a single HR and FAR pair are obtained and so generating a complete ROC curve would require multiple sessions while varying the bias of the response criterion. The ROC curve illustrates that the ability to detect a signal event is independent of any proclivity to favour a particular response. Additionally, the area beneath the ROC curve (AUC) also reflects how detectable the signal event is, with the bow toward the upper left hand corner becoming increasingly bowed as d'_{actual} increases. The ROC curve can also be plotted as a linear function by implementing the transformation $\mathbf{\Phi}(z) = \mathbf{\Phi}^{-1}(\theta_i)$ on the obtained HR and FAR, where $\mathbf{\Phi}^{-1}$ is the inverse of the Gaussian CDF (the *quantile* function or *z*-transform). The resulting *normal deviate ROC* can be expressed in the usual linear form $y = mx + b$, where the slope is proportional to the ratio of the signal and noise standard deviations, $m = \sigma_1/\sigma_0$, and the intercept is equal to d'_{actual} , $b = \mu_1/\sigma_1 = d'_{actual}$.For the equal variance case considered here the slope is unity and so the equation can be simplified to $\mathbf{\Phi}^{-1}(HR) = \mathbf{\Phi}^{-1}(FAR) + d'_{actual}$.

The SDT model outlined during the foregoing is a two-parameter model that requires estimation of the observer's ability to detect the signal stimulus (*sensitivity*; measured by $d'_{obtained}$), and the observer's criterion value (β), which allows one to assess *response bias*. With the HR and FAR in hand one has all that is needed to estimate the model parameters, and each parameter can be calculated by

$$d'_{obtained} = \mathbf{\Phi}^{-1}(HR) - \mathbf{\Phi}^{-1}(FAR). \qquad \text{Eq. 1.6}$$

and

$$\beta = e^{\left[\frac{\mathbf{\Phi}^{-1}(FAR)^2 - \mathbf{\Phi}^{-1}(HR)^2}{2}\right]}. \qquad \text{Eq. 1.7}$$

Sensitivity may also be inferred by calculating the AUC (Verde, MacMillan, & Rotello. 2006), where

$$A_z = \mathbf{\Phi}\left(\frac{d'_{obtained}}{\sqrt{2}}\right). \quad \text{Eq. 1.8}$$

By convention, a likelihood ratio decision axis is assumed, so β is the most appropriate measure of response bias. However, problems can arise in the interpretation of bias when changes in d'_{actual} occur (see Pastore, Crawley, Berens, & Skelly, 2003). This occurs because β is on a ratio level scale. Instead, an alternate criterion measure based on an interval scale ($\mathbf{\Phi}^{-1}$) is recommended (MacMillan & Creelman, 2005):

$$c = -\frac{\mathbf{\Phi}^{-1}(HR) + \mathbf{\Phi}^{-1}(FAR)}{2} \quad \text{Eq. 1.9}$$

and because it is scaled in standardised units it is unaffected by changes in d'_{actual}. Critically, β and c are not direct transformations of each other (Pastore et al., 2003), and so changes in d'_{actual} that affect one measure (e.g., c) will not affect alternative measures in the same way. Claims of changes in bias as a function of changes in d'_{actual} may simply reflect an incorrect assumption about the nature of the decision axis and the associated measure (see Pastore et al., 2003). In more elementary treatments of SDT, however, the criterion is simply specified in terms of sensory magnitude, and so the sensory criterion is often estimated as $k = -\mathbf{\Phi}^{-1}(FAR)$[4].

1.2. The Theory of Ideal Observers

SDT states that the distributional properties of the evidence completely define the upper limit of performance. In turn, the distributions themselves are influenced by the immediate psychophysical environment. This can include anything from the parameters of the signal waveform, the power of the noise, or other experimental manipulations (e.g., word frequency, study time) that are assumed to affect the discriminability of target stimuli.

[4] The remainder of this work will only consider the sensory criterion and so for ease of exposition will simply adopt the convention of denoting the criterion by *c*.

Assuming these limits can be expressed mathematically, SDT offers a method in which human performance can be compared to an *ideal observer* - an observer who possesses all necessary knowledge[5] (e.g., distributional shape, parameters, power of noise, etc.) in order to perform a set detection task maximally.

During the development of SDT, ideal observers were determined by specifying the frequency, phase, duration, onset, and amplitude of a signal sinusoid. In doing so the maximum level of performance attainable could be specified as $d'_{actual} = \sqrt{2E/N_o}$, where E is the energy of the signal and N_o is the power of the noise (Green & Swets, 1966; Swets, 1961). Such an observer is called a *signal known exactly* (SKE) observer. Human performance can then be compared with the ideal observer to investigate the efficiency of human detection. Tanner and Birdsall (1958) derived a measure of observer efficiency where the expected level of human performance is

$$d'_{obtained} = \eta\sqrt{2E/N_o}, \qquad \text{Eq. 1.10}$$

and efficiency is measured more generally as the squared ratio of the actual and obtained sensitivity parameters,

$$\eta = \left(\frac{d'_{obtained}}{d'_{actual}}\right)^2. \qquad \text{Eq. 1.11}$$

The measure of efficiency is largely taken to determine the use of available information in identifying which stimulus event occurred. Human performance was shown to be consistently sub optimal across the vast majority of detection tasks; i.e., $d'_{obtained} \neq d'_{actual}$ (Tanner & Birdsall, 1958; Green & Swets, 1966; Swets, 1961). Critically, human observers are not SKE observers.

[5] Note the ideal observer cannot perform the detection task perfectly. That is, the observer is assumed to locate their decision criterion at the optimal location and so maximise the percentage of correct decisions. The ideal observer will produce the minimum number of errors possible and so attain performance equal to d'_{actual}.

Insights into these limitations can be gleaned by degrading the information available to the SKE observer. Namely, uncertainty may be introduced by leaving certain parameters unspecified which reduces the SKE observer to a *signal known statistically* (SKS) observer. It became apparent, however, that in order to account for the empirical data additional parameters were required (e.g., learning or attention parameters). The extant SDT model thus required additional complexity in accounting for potentially unaccounted for processes that introduce variability into response data. For example, Swets (1961; see also Green & Swets, 1966) suggests that discrepancies between human and optimal detection may reflect variability in an observer's internal processes, such as inherent biological or neural noise, or faults in cognitive processes; or what is generally referred to as *internal noise*. One of the more patent ramifications of internal noise is the inconsistency in observer responding, or what is often called *decision noise*. Indeed, Green and Luce (1974) note that "perhaps the single most pervasive characteristic of psychophysical data is the inconsistency of subjects when answering most questions we ask them about simple stimuli" (p. 373). Collectively, inconsistencies and inefficiencies in human performance challenge the capability of an inherently flawed system to maintain a single, long-term criterion representation.

1.3. Theoretical Challenges: Criterion Variance

A very real issue still remains for SDT regarding how best to deal with variations in observer response probabilities. Entailed within the theory's assumption of a fixed decision criterion is the idea that each trial is an independent event, or what is referred to as the Bernoulli assumption. While this certainly holds for stimulus presentation, the assumption is also extended to the observer's response on each trial. This implies a static decision process whereby responses are made in the absence of any decisional noise, and based solely upon the evidence available during any given trial. Perhaps the earliest evidence that this was not so were findings that demonstrated non-independence in observer responding, or *sequential dependencies* (Howarth & Bulmer, 1956; Verplanck, Collier, Cotton, 1952; Collier & Verplanck, 1958). In general, sequential dependencies - or *response autocorrelations* (Benjamin, Diaz, &

Wee, 2009) - are contextually mediated effects that describe correlations between the current response and preceding events, which can include the previous response, stimulus, and feedback (or *knowledge of results*; KR). More generally they are described in terms of decisional recency, where a positive correlation, or *assimilation*, occurs when the observer is inclined to respond with the previously reinforced response on trial $n - 1$. Conversely, a negative correlation, or *contrast*, is the reverse effect (Jones & Sieck, 2003; Jones, Love, & Maddox, 2006). Sequential dependencies have since been found in virtually all types of detection, including absolute identification and magnitude estimation (Brown, Marley, Donkin, & Heathcote, 2008; DeCarlo & Cross, 1990; Holland & Lockhead, 1968; Luce, Nosofsky, Green, & Smith, 1982; Petrov & Anderson, 2005; Stewart, Brown, & Chater, 2005; Ward & Lockhead, 1971), probabilistic and perceptual categorisation (Jones & Sieck, 2003; Jones et al., 2006; Stewart & Brown, 2004; Stewart, Brown, & Chater, 2002), and identification and detection tasks (Atkinson, Catterette, & Kinchla, 1964; Howarth & Bulmer, 1956; Tanner, Haller, & Atkinson, 1967; Tanner, Rauk, & Atkinson, 1970; Treisman & Williams, 1984).

From an empirical standpoint, the tendency for observer response patterns not to be independent point toward a dynamic decision process. The theoretical corollary, then, means that in order to account for the empirical data the criterion must be treated as a random variable in its own right. However, criterion variability precludes consistent performance and thus threatens the inferential validity of the signal detection model. More worryingly, SDT estimates may be grossly unrepresentative of true observer performance. Though antecedents to modern detection theory (Thurstone, 1927; see also Lee, 1969) explicitly considered variability in the decision threshold, this feature was not inherited within the SDT framework, despite its palpable commonalities with Thurstonian comparative judgement. This said, the notion of criterion variance was not without consideration during the development of SDT (Green & Swets, 1966).

During the theory's nascent years, Tanner and Swets (1954) noted that a variable criterion will introduce additional variance into the response data (see also Tanner, 1961).

Subsequent studies began to accumulate further evidence that suggested variability in the decision criterion was indeed present (e.g., Bonnel & Miller, 1994; Hammerton, 1970; Lee & Zentall, 1966; McNicol, 1975; Nosofsky, 1983). For example, criterion variance can be demonstrated by having observers shift their criterion between two locations during the course of an experiment. The shifting between locations has been shown to depress estimates of $d'_{obtained}$ with the corresponding ROC point falling below the optimal curve (Eijkman & Vendrik, 1964; Thomas & Myers, 1972; Wagenaar, 1973). Additionally, task demands, such as the number of criteria that need to be maintained, also have an effect upon criterion variability (e.g., Treisman, 1985). That is, the more criteria that need to be remembered the harder it is to maintain them. For example, Ell, Ing, and Maddox (2009) argued that this demand affects the ability of the working memory store to maintain multiple elements, and demonstrated an effect of cognitive load upon criterion noise. However, despite such demonstrations, most Thurstonian decision-bound theories, which includes SDT and multidimensional generalisations thereof (e.g., General Recognition Theory; GRT: Ashby & Gott, 1988), persist with the fixed criterion assumption.

For the most part, any additional variance associated with the decision stage was not considered the province of the stimulus-orientated psychophysicists. To further complicate matters, Green, McKey, and Licklider (1959) noted that "apparently, the parameters of the internal noise are not independent of the external noise "(p. 518). Accordingly, this places limits upon the ability to separate the contributions of sensory-based and decision-based noise, making it difficult to deal with criterion variability in any constructive sense. Where known model limitations would usually provide a basis for model extension and development, the effects of decision noise were largely considered a nuisance and unceremoniously lumped with the evidence distribution variance (Mueller & Weidemann, 2008; McNicol, 1972). However, such a practice offers no real solution to the problem and in fact can create further issues when it comes to interpreting model parameters. Notwithstanding these difficulties, a number of models do exist that try to account for decision noise through further extension of the signal detection model (e.g., Kac, 1962;

Dorfman & Biderman, 1971; Mueller & Weidemann, 2008; Thomas, 1975). In summary, that signal detection, *a priori*, is not able to account for variations in observer responding is of theoretical concern. This is not to say that the theory is so fundamentally flawed that it cannot be remediated in some way; rather, it highlights the need to acknowledge these limitations and find ways in which SDT may be modified

1.3.1. The Effects of Criterion Variance

A criterion that remains static across trials provides an appealing, though perhaps ambitious, characteristic of human decision-making. Conceptually, the criterion is viewed as a single vertical line upon the decision axis (cf. Figure 1.1). Mathematically, a criterion with zero variance is a special case of the normal PDF where the deviance parameter is set to zero. This is called a Dirac delta function (Rosner & Kochanski, 2009),

$$f(x) = \lim_{\sigma \to 0} \delta(x - \mu). \qquad \text{Eq. 1.12}$$

Strictly speaking the delta function is a generalised function and takes on values of zero for all $x \neq \mu$, with all the density located at μ. The integral of the Dirac delta function is the Heaviside function, where

$$H(x) = \int_{-\infty}^{x} \delta(x - \mu)\, dx \qquad \text{Eq. 1.13}$$

$$= \begin{cases} 0, & x < \mu \\ 1, & x \geq \mu \end{cases},$$

and is a step function. This satisfies the fixed criterion decision rule where μ denotes the criterion location, and implies that all evidence values that exceed the criterion will receive a "yes" response with a probability of 1, whereas all others will receive a "yes" response with a probability of 0. In light of the discussion so far, the evidence that observers respond so consistently is lacking. So, rather than viewing the criterion as a single long term value, a more appropriate interpretation views the estimated criterion location as the mean of a stochastic process that affects the trial-by-trial location of the criterion (e.g., DeCarlo, 2010; Thomas, 1975; Wickelgren, 1968). Specifically, trial-by-trial shifts in the criterion location

create a distribution of points, meaning the criterion can be treated like any other random variable. This then ascribes distributive properties to the criterion which results in a CDF that departs from the step function specified under the fixed criterion assumption.

A problem presents itself when we assume the estimated detection parameters, based upon the response data, accurately reflect the true state of nature. The presence of any additional variance beyond that specified by the model will confound measurement by degrading the quality of the response data. This occurs when additional noise is mixed with the perceptual variability, further degrading the decision variable which results in response distributions that have larger variances than the evidence distributions. McNicol (1972) further demonstrated that when the distributions are rescaled to have unit variance the means of both distributions become truncated. Consequently, observer performance is underestimated. In fact, $d'_{obtained}$ is underestimated by a factor of $\sqrt{1 + \sigma_c^2}$ (McNicol, 1972; Schoeffler, 1965; Wagenaar, 1973), where $d'_{obtained} = d'_{actual}/\sqrt{1 + \sigma_c^2}$, and σ_c^2 denotes criterion variance. This underestimation has serious implications when it comes to assessing performance in situations where an ideal observer cannot be specified – critically, the level of asymptotic performance is unknown. Unfortunately, in most cases the asymptote *is* undefined and we rely upon the estimates as veridical indicators of performance. Given that SDT makes no explicit provision for any variance other than sensory contributors, such a practice must result in spurious estimates of performance. Quite simply, if an observer is not performing at asymptote then a fixed criterion observer cannot be substantiated; however, the degree of departure from asymptote is unknown so we accept our estimates as true despite the very likely presence of criterion variability.

1.3.2. Externalising Decision Behaviour

Internally distributed perceptual variability places limits upon estimating the degree to which decision, or criterion, noise contributes toward the total noise variance. To gain some idea of the way the decision criterion may shift, researchers require a means through which sensory noise can be controlled. One successful approach has been to externally

distribute the stimuli (e.g., Healy & Kubovy, 1981; Kubovy & Healy, 1977a; Kubovy, Rapoport, & Tversky, 1971; Lee, 1963; Lee & Janke, 1964, 1965; Lee & Zentall, 1966; Podd, 1975; Ward, 1973; Ward, Livingston, & Li, 1988; Zak, Katkov, & Sagi, 2012); this approach has also been referred to as the *external noise method*, though here the method will be referred to as the probabilistic categorisation task (PCT; cf. Kubovy & Healy, 1977a).

Like the fundamental detection task, the PCT assumes that there are two possible events that can occur at any one time; the difference is, however, that the PCT assigns stimuli to one of two categories which are then distributed along a uni-dimensional continuum. So, the decision axis in the PCT is built from discrete stimulus values, and may be built along any dimension upon which a stimulus can vary (e.g., line length, tone frequency, greyness, dot position, numbers; see Kubovy & Healy, 1977a; Lee & Janke, 1964; Lee & Zentall, 1966; Podd, 1975); stimuli must be able to take on absolute values, though. The probability with which a stimulus is presented is usually determined by calculating probability mass functions, which are massed to best approximate a normal distribution for each category class. As Lee and Zentall (1966) describe

> "Such a distribution was called an externally distributed stimulus, since it is analogous to a stimulus such as a tone or attitude statement which is presumed in detection theory or Thurstone theory to give rise internally on each presentation to a sample point from a probability distribution. In the present case, however, such a sampling is done from a distribution which has variance external to S[subject]. This allows more information to be obtained on how S handles such a problem" (p. 120).

Confusability is thus introduced probabilistically by having the stimulus categories overlap by some amount, meaning category membership for any particular stimulus is not certain. Feedback is also probabilistic where the probability of reinforcement for each stimulus varies according to the stimulus location along the continuum (Lee, 1963).

In effect, the PCT creates an analogue to the signal detection task where what would normally be an internally distributed perceptual representation now becomes externally

distributed stimulus values. Furthermore, the nature and characteristics of the evidence (i.e., stimulus values and distributional parameters) are now observable and measurable. This allows researchers to calculate the optimal level of performance, in an ideal observer sense, which then provides a standard against which one can assess human performance (see Section 1.2). The utility of the PCT is such that decision behaviour may be studied independently from the contributions of sensory processes (Ward et al., 1988). As noted by Lee and Janke (1965), "The method of externalization of the distribution is useful to study the choice principles involved in such situations, since E [the experimenter] can control and observe the sampling" (p. 79). In order to validate such an assumption it was typically the case that all stimuli were separable (e.g., Lee & Janke, 1964; Zak et al., 2012) where in most instances perceptual noise is considered to be negligible (Glass, Maddox, & Markman, 2011). That is, in a sense, the stimuli are as close to veridical as possible (Maddox, 2002). That the stimuli reside at separable points along the stimulus continuum, and that these points can be set so each discrete value is discriminable from the next, significantly limits the role that variability in the perceptual effect can play. However, this is not to say that the perception of a stimulus will not be variable. Perceptual variability is of course axiomatic (Ashby & Lee, 1993); though it does mean that whatever noise *does* exist is unlikely to be due to perceptual confusability. Rather, it is likely to be due to decision noise. Letting Maddox (2002) summarise these points,

> "…in many cases, perceptual noise will be small, and the *perceptual representation will be close to veridical* [emphasis added]. In addition, experience with a task and certain types of decision criteria decrease the magnitude of criterial noise. Despite this fact, it is important to acknowledge these inherent sources of noise and to account for them within theories of categorization" (p. 575).

Having the stimuli externally distributed also means that some metric of criterion variance may be gleaned. Recently, Zak et al. (2012) derived estimates of criterion variability in a visual dual discrimination task using externally distributed stimuli composed of Gaussian

luminance blobs. By calculating the proportion of responses for each stimulus value along the decision axis, the slope of the response function may be used to estimate the amount of criterion variability (see also Lee & Zentall, 1966). As Zak et al. (2012) note,

> "The use of unidimensional external noise with well-defined statistical properties allows for the assessment of decision criteria in stimulus space (stimulus criterion) as the stimulus level at which observers switch from one response alternative to the other, thus largely avoiding the above-mentioned limitations inherent in the standard SDT analysis" (p.1043).

The PCT has allowed researchers to study the observer decision criterion under a number of conditions. An increased amount of interest has centred on criterion learning under conditions where the costs and benefits associated with specific categorisation environments are manipulated (Bohil & Maddox, 2003a; Maddox, 2002; Maddox & Bohil, 1998; Maddox & Bohil, 2004; Maddox & Dodd, 2001), and simiarly the study of gains and losses via the regulatory fit framework (Glass et al., 2011; Maddox, Baldwin, & Markman, 2006; Maddox, Markman, & Baldwin, 2007; Markman, Baldwin, & Maddox, 2005; Markman, Maddox, Worthy, & Baldwin, 2007). The associated effects of feedback in decision criterion learning have also received extensive attention (Bohil & Maddox, 2003b; Maddox & Bohil, 2001, 2005). However, the role that feedback plays is far from being completely understood.

1.4. The Role of Knowledge of Results

One potential, though often overlooked, source of decision noise is whether or not knowledge of results is provided. Knowledge of results (KR) refers to information that is provided with the express purpose of facilitating performance in some way, usually through providing specific distributional knowledge which enables the criterion to be optimally located. KR is viewed as an integral component in skill acquisition across a number of applied fields, which includes: visual inspection and quality control (e.g., Embrey, 1975; Micalizzi & Goldberg, 1989); motor task learning (e.g., Blackwell & Newell, 1996; Magill, 1994; Salmoni,

Schimdt, & Walter, 1984; Schmidt, Young, Swinnen, & Shapiro, 1989; Swinnen, Schmidt, Nicholson, & Shapiro, 1990), perceptual motor and visual tasks (e.g., Chun & Wolfe, 1996); vigilance (e.g., Weidenfeller, Baker, & Ware, 1962; Szalma, Hancock, Dember, & Warm, 2006); and probability learning (e.g., Atkinson et al., 1964; Estes & Johns, 1958; Tanner et al., 1967, 1970). The content of the KR, and how it is delivered, may also vary (e.g., Chinn & Alluisi, 1964; Szalma et al., 2006). For example, KR may inform observers that their decision was incorrect; or it may provide them with a percentage correct score at the end of a block of trials. Other forms may quantify by how much an observer was off in an estimate; for example, whether a drawn line was a number of centimetres too short or too long (see Salmoni et al., 1984, for a review).

It is almost universally accepted that KR improves task performance. However, despite an abundance of research that ostensibly supports this position (see Kluger & DeNise, 1996), various researchers have expressed doubt about the benefits KR supposedly bestows; for example, Latham and Locke (1991) state that

> "Few concepts in psychology have been written about more uncritically and incorrectly than that of feedback.... Actually, feedback is only information, that is, data, and as such has no necessary consequences at all" (p. 224);

Additionally, Salmoni et al. (1984) suggest that "To neglect KR as a variable in the study of learning would lessen our understanding of skill acquisition considerably" (p. 355). Balcazar, Hopkins, and Suarez (1985), perhaps more tersely, state that "feedback does not uniformly improve performance" (p. 65). While such positions stand in stark contrast to the prevailing notions, they are not without support.

A review by Kluger and DeNisi (1996) suggests that KR may have been viewed in an overly favourable light. This interpretation appears to have been largely compounded in a review by Ammons (1956) whereby empirical disparities in the reported benefits were overlooked, instead concluding that KR universally increases both learning and motivation. In wishing to elucidate a truer effect of KR within the extant literature, Kluger and DeNisi went

on to collate studies that could be identified with the keywords KR and feedback, which were drawn from a number of fields within the psychological domain. They then conducted a meta-analysis by calculating Cohen's *d* for all studies where an effect for KR was reported. This yielded some 607 effects sizes, based upon 22,663 observations (average sample size per effect was 32). The results suggested that there was a modest positive effect for KR (Cohen's d = .41; SE = .09); however, over 38% of the reported KR effects were negative; KR is far from universally beneficial. While a positive effect can be present, there is evidence to suggest that KR effects can be negative.

From a psychophysical standpoint, KR typically informs the observer as to which event occurred during the observation interval. For example, whether or not it contained the tone, or whether the tone was higher or lower than a comparison. KR usually occurs on a trial-by-trial basis which allows the observer to gauge the frequency with which each event occurs; KR provided in this way is referred to as trial-by-trial knowledge of results for stimulus events, or $TTKR_e$, One of the common uses for TTKR is to alter the observer's degree of bias, aiming to drive the criterion toward a desired location (e.g., Green & Swets, 1966; McNicol, 1975, Mueller & Weidemann, 2008). Indeed, this is the approach taken in generating empirical ROC curves using only the binary response format. It is often further assumed that feedback enables an observer to sufficiently learn and *stabilize* their criterion.

1.4.1. Schoeffler's Theory for Psychophysical Learning

In 1965, Schoeffler introduced a theory of psychophysical learning whereby response probabilities are sequentially updated across trials as a function of: a) the reinforcing event at the cessation of the trial (i.e., whether TTKR is provided or not); b) the difference in stimulus magnitude between the observation sampled on trial n and trial $n + 1$; and c) the learning rate. The model is probabilistically defined and does not necessitate the existence of a decision criterion and so is more appropriately based upon statistical learning theories (e.g., Bush, Luce, & Rose, 1964). Schoeffler's model assumes that there exists an ordered probabilistic relationship whereby the probability of making a signal response is a

monotonically increasing function of the magnitude of the observational sample, x. The nature of x is incidental; it may reflect stimulus magnitude in terms of overtly observable characteristics (e.g., pitch) or may reflect the internal evidential value of the decision variable (e.g., likelihood). Whatever the case, it suffices to establish that varying magnitudes of any evidential metric will influence the probability of specific responses. To demonstrate, assume on trial n that an observation is sampled from one of the evidential distributions, x_n. Further assume that, after a response is made, the observer is provided with TTKR and is told that x_n was sampled from the signal distribution, and so $x_n = S_{1,n}$. With this information in hand, the probability of making a future signal response for all perceived values greater or equal to x_n , i.e., $x \geq x_n$, is increased, and can be expressed as follows

$$P_{n+1}(R_1|S_{1,n}) = P_n(R_1|x_n) + [1 - P_n(R_1|x_n)]\beta \cdot \int_{-\infty}^{x} f(x_n|i = 1) f(x - x_n) dx_n \quad \text{Eq. 1.15}$$

for all possible values of x_n (Schoeffler, 1965). In breaking down the terms of the equation, the first term on the right simply reflects the probability of a signal response given x_n. The second term establishes the increase in probability, where β is the learning rate parameter. Note that the increase is proportional to the maximum increase possible; i.e., $[1 - P_n(R_1|x_n)]$. The final term concerns the probability that x is greater than the perceived magnitude of x_n and allows for a variable effect in the updating of the response probability. Note, we can also define

$$P_{n+1}(R_0|S_{0,n}) = P_n(R_0|x_n) + [1 - P_n(R_0|x_n)]\beta \cdot \int_{-\infty}^{x} f(x_n|i = 0) f(x_n - x) dx_n \quad \text{Eq. 1.16}$$

for all $x \leq x_n$. More generally, the probability of a signal response for any given value of x on trial n can be expressed via Luce's choice axiom, where

$$P_n(R_1|x) = \frac{P_n(R_1|S_{1,n})}{P_n(R_0|S_{0,n}) + P_n(R_1|S_{1,n})}. \quad \text{Eq. 1.17}$$

Further, note that in $\lim_{n\to\infty} \beta = 0$, and so asymptotic performance is equivalent to

$$P_n(R_1|x) = \lim_{n\to\infty} P_n(R_1|x) = \frac{\boldsymbol{\Phi}_1(x)}{\boldsymbol{\Phi}_0(x) + \boldsymbol{\Phi}_1(x)} \qquad \text{Eq. 1.18}$$

where $\boldsymbol{\Phi}_i$ is the CDF of the i^{th} evidence distribution. The function mapping the probability of making a signal response is thus a continuously increasing function for greater values of x, and is referred to as the *psychometric function*.

The asymptotic psychometric function implies that the level of decision noise will asymptote to some non-zero value. Schoeffler's model is thus not in agreement with the axioms of SDT, according to which the psychometric function assumes a step function. Indeed, Schoeffler discusses the implication the model has for the signal detection framework. Though probabilistic models are "criterion free", Schoeffler reconciled his theoretical predictions by suggesting that a continuously increasing response function could be generated via a criterion that has variance. That is, the slope of the function corresponds to a range of values at which the criterion was located at some point in time. So, the psychometric function can be interpreted as a distribution of criterion values (cf. Section 1.3.2).

While the reinforcing effects of TTKR are assumed to facilitate learning, Schoeffler went on to demonstrate that in the absence of TTKR the slope of the psychometric was *steeper* than the TTKR slope, implying improved performance without TTKR. Schoeffler explains that observers have an increased reliance upon TTKR owing to its higher informational value. However, circumstances can arise whereby the TTKR may not be consistent with the observer's expectation and so will be non-informative in this sense (Schoeffler used the term *misinformation*). For example, on some trials, upon perceiving a low magnitude sample, the observer will most likely respond with a noise response - though there is some probability that the observation was sampled from the signal distribution. When the observer receives TTKR that informs them that the sample was a signal event, the observer is expected to adjust their response probabilities so as to avoid making similar errors again in the future. In the absence of TTKR such adjustments cannot occur. Rather, the

observer has only their previous response and the magnitude of sensory effect to rely upon. Consequently, in the absence of any information that could potentially be viewed as misleading, the amount of noise than can enter the decision process is notably reduced. In sum, Schoeffler argues that learning is influenced by the availability of the information present within the environment, and when information is removed that can be misleading, performance is expected to improve.

Perhaps one of the more paradoxical realisations underpinning these predictions is that the TTKR is in no way untruthful – i.e., it is completely veridical. That it may actually *negatively* influence performance, then, may seem to some an anathema. However, during the same period further reports surfaced suggesting that TTKR may induce such negative effects. For example, Carterette, Freidman, and Wyman (1966; see also Carterette & Wyman, 1962; Gundy, 1961) established that providing TTKR may alter the sensitivity index by inducing the observer to shift criterion after errors have been made. Additionally, McNicol (1975) found that providing TTKR during an absolute identification task reduced estimates of d' compared to where no TTKR had been provided. McNicol concluded that shifts in the decision criterion must be contributing additional noise toward the perceptual processes and that feedback has a critical role in this.

The Probabilistic Categorisation Task (PCT) has also been used in demonstrating the effects of TTKR, with the obvious benefit that the effects are independent from sensory confound. Furthermore, the use of the PCT provides a nice medium through which the effects of TTKR can be explained. That is, the role of misinformation is played by the overlap between stimulus categories, where the probabilistic TTKR attaches evidential labels upon the stimuli that may not be in accordance with the observer's expectations. Accordingly, qualitatively identical stimuli may at varying times have different labels placed upon them by virtue of which distribution was sampled. One might expect such objective violations to have their greatest effect upon performance where there is greater stimulus confusability (i.e., greater distributional, or category, overlap). Conversely, as the task (theoretically) becomes

easier by limiting the degree of category overlap, fewer stimuli are confusable and the negative effects are expected to abate.

Lee and Zentall (1966) published a study that investigated factorial effects that influenced observer responding in probabilistic categorisation. Of the factors that were investigated the two most critical were task difficulty (d'_{actual}= 1.25, 2.25, 3.5) and whether TTKR was provided or not. The crucial finding was that in the absence of TTKR improvements were only observed in the difficult (d'_{actual}= 1.25) condition. This was evidenced by an increased slope in the psychometric function and an improvement in mean estimates of sensitivity. Lee and Zentall note that such improvements under difficult conditions are not usually observed in psychophysical studies (TTKR is usually seen as helpful). Upon interpretation, Lee and Zentall suggest that in removing TTKR "the hopeless attempt to eliminate error is forsaken, in favour of responding according to the most probable answer" (p. 124). In Chapter 2 I will discuss my own research that has investigated the effects of TTKR upon observer performance.

1.5. Modeling Criterion Shifts

The foregoing strongly suggests that feedback plays a fundamental role in influencing observer performance, and that this influence may not always be for the better. Moreover, the type of TTKR that is provided may differentially influence how labile the criterion is across trials (cf. Schoeffler, 1965; see also Bohil & Maddox, 2003b). However, investigations into the effects of TTKR have largely been confined to the learning of the appropriate criterion location (e.g., Healy & Kubovy, 1981; Maddox & Bohil, 2001, 2005) or have simply inferred the variability via indirect means (e.g., Lee & Zentall, 1966; Hammerton, 1970; Zak et al., 2012). Furthermore, probabilistic conceptualisations of the decision process (e.g., Schoeffler, 1965) completely preclude inferences about the way the criterion may shift across trials.

A few years prior to Schoeffler's learning theory, a procedure was proposed that described trial-by-trial criterion shifts. In his seminal paper, Kac (1962, see also Kac, 1969) introduced an error-correction model whereby the criterion was assumed to shift after errors

were committed. This type of model not only assumed that sequential dependencies do exist between observer responses (cf. Howarth & Bulmer, 1956; Collier & Verplanck, 1958), but introduced a theoretically viable mechanism to explain criterion shifts. In effect, observers engage in error-monitoring and attempt to reduce the probability of making similar errors again in the future by incrementally adding or subtracting a constant value to the current criterion location. Specifically, the criterion was assumed to increase by a fixed amount following errors on noise trials, and decrease by a fixed amount following errors on signal trials. For this reason such models have been referred to as *additive learner models* (ALM; Dusoir, 1980). Shifts are recursively expressed, where assuming a given S_iR_j outcome the criterion will be updated according to the following rule:

$$C_{n+1} = \begin{cases} C_n + \Delta_{ij}, & if\ i = 0 \\ C_n - \Delta_{ij}, & if\ i = 1 \end{cases} \qquad \text{Eq. 1.19}$$

where Δ_{ij} is the amount of shift, or what may be referred to as the learning parameter (Dorfman & Biderman, 1971), determined by the ij^{th} outcome. Additionally, Δ_{ij} is assumed to be constant for all shifts, where $\Delta_{ij} = 0$ for all $i = j$ reflects the error-correction model (cf., Kac, 1962; see also Larkin, 1971). A more general expression may be used, where

$$C_n = C_0 + \sum_{n=1}^{N} \Delta_{ij} K_{ijn}, \qquad \text{Eq. 1.20}$$

K_{ij} is the number of ij^{th} outcomes up to trial n, and C_0 is the initial criterion location. In this form the model can be fit as a generalised linear model (GLM) and provides a mathematically tractable approach in fitting the model.

The ALM assumes that the criterion shifts deterministically following errors, though this need not be the case. The model can be parameterised probabilistically by substituting $p_{ij}\Delta$ into Equation 1.20 above (e.g., Thomas, 1973, 1975). The error-correction model has been further extended by allowing shifts, whether they are probabilistic or deterministic, to occur following all trial outcomes (Dorfman, 1973; Dorfman & Biderman, 1971; Dorfman,

Saslow, & Simpson, 1975; Thomas, 1975); that is, $\Delta_{ij} \neq 0$ for all $i = j$. Dorfman and colleagues (1971, 1975) established that, in general, the size of the shift following correct decisions was smaller in comparison to the shifts after errors (i.e., $\Delta_{11} = \Delta_{00} < \Delta_{01} = \Delta_{10}$). The generalisation also caters to the assumption that observers assimilate toward to the previous correct response (cf. Treisman & Williams, 1984). The general model assumes that the criterion will decrease by some amount following correct signal responses, and increase by some amount following correct noise responses; so the criterion will shift in the same direction following each i^{th} event regardless of the observer's response; i.e., $-\Delta$ for all $i = 1$ and $+\Delta$ for all $i = 0$. While these generalisations introduced flexibility, on the whole they offered very little improvement in fit over the simpler error-correction model (see Dorfman et al., 1975; Dusoir, 1980; see also Thomas, 1973, where further problems with the generalised model are discussed). Dorfman and colleagues (1975) suggested that some element of random variability in the criterion is required following all decisional outcomes in order for the model to better fit observer data. Treisman and Williams (1984) also investigated similar processes through which they derived their Criterion Setting Theory (CST). In essence the CST model is no different to the ALMs described by Kac (1962) and Dorfman and colleagues (1971, 1975), with the only difference being the introduction of a process that allows the criterion to decay back toward a reference point during the inter trial interval.

The ALMs at heart provide systematic mechanisms through which the criterion shifts across trials; however, criteria may not necessarily shift in such ways. As already alluded to above, non-systematic (i.e., random) shifts may also account for the data. More recent modeling efforts have considered this point (Benjamin et al., 2009; Erev, 1998; Mueller & Weidemann, 2008; Turner, Van Zandt, & Brown, 2011). The basic tenet of such models is that the criterion is a sampled value on each trial, rather than a systematically altered value; though in some instances a combination of both systematic and random processes may be used (e.g., Erev, 1998; Dorfman et al., 1975). As an example, Mueller and Weidemann's Decision Noise Model (DNM; 2008) describes both yes/no classification and rating responses

in terms of random criterion sampling. For yes/no responses, a central criterion is sampled from a Gaussian distribution against which the stimulus is compared. As usual, the stimulus is classified according to which side of the criterion it falls. Rating responses can further be made by sampling criterion values from either the left- or right-most category, depending on which side the stimulus has fallen. Should the newly sampled criterion value exceed the stimulus value, then sampling ceases as the stimulus is within a confidence category, and so the appropriate rating response is given; otherwise sampling continues.

In some instances, however, the observer may not be provided with feedback at all (e.g., Dusoir, 1980; Schoeffler, 1965; Thomas, 1975). Under such circumstances the observer has only their objective internal representation and their previous response as sources of information, Thomas (1975) suggested an error correction model that assumed shifts were based upon the observer's response, and the deviation between the sensory magnitude and the criterion. That is, in the absence of feedback the observer will use sensory magnitude in updating the criterion. As Thomas (1975) notes "This assumption [error-correction] about criterion adjustment is common to all models that assume the subject computes a standard as a weighted average of the most recent sensory information" (p. 160). That is, the criterion is an updated internal standard, reminiscent of the process assumed by adaption-level theory (Helson, 1947, 1964; see also Durlach & Braida, 1969, and the recent Internal Reference Model: Dyjas, Bausenhart, & Ulrich, 2012). Specifically, assuming that a signal sample is incorrectly labelled as noise, the sample has fallen below the criterion and so the criterion shifts towards x (i.e., is lowered) as Thomas (1975) suggests. Furthermore, while this conceptualisation is assumed to operate only in the absence of TTKR, there is little reason to suggest that it does so exclusively. Evidence that this may be so will be discussed next. In Chapter 3 these issues will be taken up again and discussed more thoroughly.

1.5.1. Exploring Criterion Shifts Using the PCT

A particularly convenient aspect of the PCT is that we may assess more directly the effects of decision noise upon task performance. However, virtually no research has used the

PCT in investigating just how well the ALMs account for trial-by-trial response data. The only study that has considered the abilities of the ALM under such conditions was that of Kubovy and Healy (1977a). This study attempted to compare probabilistic and deterministic response models; specifically, they compared Schoeffler's (1965) model with the general ALM introduced by Dorfman & Biderman (1971) using a numerical analogue to signal detection; that is, a PCT using externally distributed numerical stimuli. Their approach in comparing the models rested upon comparing the data obtained across two experiments in which the response format was expected to influence how the observer responded.

In the first condition observers were free to use whatever decision strategy they preferred. Observers simply underwent the task with no constraints placed upon the way they responded. In the second condition a cut-off report method was used where observers had to explicitly state their criterion location on each trial (i.e., some critical number), and were thus required to respond in accordance with the stated criterion value. If the stimulus value (a number sampled from one of the distributions) exceeds the stated criterion value then the observer must respond "high", for example. Observers were then provided with TTKR, after which they had the chance to update the criterion. In using this method Kubovy and Healy reasoned that they could assess criterion shifts without having to fit the ALM to the data.

To formally compare the models, the expected number of static criterion violations – the number of low (high) responses to stimuli above (below) the optimal criterion– were calculated based upon Schoeffler's asymptotic response curve (cf. Equation 1.18). The numbers of observed static criterion violations from both conditions were then compared to assess whether observers were responding probabilistically. The authors found that the observed proportion of violations from both conditions were more similar to each other than to the number predicted by Schoeffler's model. Observers were determined not to be responding probabilistically which lead to the rejection of Schoeffler's model. The reason driving this decision was that since the observers were forced to respond in a more ALM-like fashion, the similarities in proportion imply that a similar process must have been operating

for the free response task. However, a sequential analysis of the data from their Condition 2, which looked at the direction of criterion shifts following specific decisional outcomes, found the shifts to be incompatible with the expected shift direction specified by the ALM. Critically, shifts were reported in both directions following errors and correct decisions, whereas the ALM prescribes shifts in only a single direction following an outcome. Consequently, the ALM was also rejected.

The major finding from Kubovy and Healy's (1977a) study is that the criterion may not shift in the direction specified by the general ALM. Instead, the criterion appears to shift in either direction following correct decisions. This finding may provide a reason as to why the fits by Dorfman et al. (1975) failed to account for all the variability. It may be that the addition of a drift parameter is unnecessary if we accept that the shifts are not always occurring in the direction specified by the models. This said, a purely weighted average explanation cannot account for the data either, as the criterion would exclusively shift toward the sampled value. Therefore, it appears that neither explanation is completely adequate. Furthermore, the results are difficult to interpret given the forced nature of the cut-off report method (Dorfman, 1977). Additionally, whether the ALM is an adequate model for such data is difficult to decide in the absence of formal model fitting, though the descriptive findings certainly suggest that the validity of the ALM is questionable. These issues will be covered in Chapters 3 and 4.

1.6. Summary and Aim of Thesis

Analogues to SDT where the stimuli are distributed externally to the observer provide opportunities to study observer decision behaviour independently from the sensory contributors to task performance (e.g., Kubovy & Healy, 1977; Lee & Zentall, 1966), providing an opportunity to better understand criterion variability, where such knowledge is precluded in orthodox detection. The present chapter has touched upon both the theoretical issues and developments in addressing criterion variance. One immutable finding common to all is that observer responding is not consistent, and a multiplicity of models may be invoked to explain

trial-by-trial variability in the decision criterion. This said, the validity, or applicability, of a number of models remains in doubt, and so further efforts are required to improve our understanding about how observers approach categorisation/detection-type tasks. It is hoped that this thesis can contribute toward this understanding and draw attention not only to the issues of criterion variance, but also the role that feedback plays in observer performance.

A central goal, then, is to examine more closely how feedback affects observer choice behaviour and whether the current ALMs can account for observer response data generated using the PCT. In doing so, Chapter 2 will return to a study that was conducted for my Masters degree (Taylor, 2010) where a reanalysis of the original thesis data is conducted. The study investigated how different types of feedback may affect observer performance, though did not include any analysis on how observers assigned their responses (i.e., assessment of the psychometric function) nor any sequential-based analysis upon the trial-by-trial criterion shifts (i.e., fitting an ALM or similar models). These limitations are addressed in the hope of developing a fuller appreciation of the effect of feedback. Additionally, while a number of dynamic criterion models may be fit to the data, the analysis will largely focus on simple error-correction models as these have received the most empirical support (cf. Thomas, 1973, 1975; Dusoir, 1980). Though there is certainly equivocation regarding the overall validity of the ALM (e.g., Kubovy & Healy, 1977a) the various models serve as a useful metric for assessing the effects of feedback upon performance. Moreover, they provide an opportunity to directly assess how well such models perform where the nature of the statistical distributions are known; such evaluations have not been made up until this point.

A notable limitation with the ALMs as they presently stand is that they appear to be lacking an explanation in regard to the way in which feedback influences the use of stimulus information. One of the more pertinent issues is the fact that all models must make some assumption, or guess, about the initial criterion value, or mean value around which the criterion varies (or is sampled; see, for example, Turner et al., 2011); however, the models make no suggestion as to how observers store, or integrate, stimulus information with

experience. Criterion placement and maintenance requires some idea of the nature of the stimuli (statistical properties, magnitudes). In most psychophysical studies the stimuli are abstract; this means that observers have very little prior information in facilitating placement of a criterion. As Turner et al. (2011) point out

> "If it is agreed that observers cannot possibly have a useful representation of these stimuli before significant experience with the stimuli themselves, then it must also be agreed that there is no basis for the observer to place a criterion, referent or bound within whatever impoverished representation they may possess" (p. 584).

What is then needed is a model whereby the stimulus information, accumulated across time, influences the criterion. Therefore, one of the goals of this thesis is to unify and improve upon the extant learner models by explicitly considering how feedback might be used or integrated. Chapter 4 introduces a model which seeks to achieve this end. This model is an extension to Kubovy and Healy's (1977a) Ideal Learner Model (ILM) where stimulus information is represented as the trial-by-trial updating of the mean value of each stimulus distribution. This model then allows for stimulus information to be integrated across time, where the criterion is established as the midpoint between the means and is subject to fluctuation as new information arrives. While the model has limitations (discussed later) it does provide a means to explore performance in tasks like those described by Taylor (2010). Furthermore, we may also assess how various feedback manipulations affect the stimulus representations. That is, how does the nature of the feedback affect the storage of stimulus information and the subsequent trial-by-trial shifts in the criterion? This is also investigated in Chapter 4. Finally, Chapter 5 describes a small empirical investigation, the data from which will be used to assess the appropriateness of the modified ILM. The model will also be tested against other ALMs, yet to be discussed, and so a large part of Chapter 5 will focus upon model selection and comparison. It is here that we may draw distinctions between the models in their ability to fit the empirical data. One of the more important factors under investigation is the effects of feedback, and how well the models can account for it.

Chapter 2

Response Inconsistences and the Psychometric Function

In this chapter we will return to Taylor's (2010) results which sought to establish whether various types of feedback had a demonstrable effect upon observer performance in a probabilistic categorisation task. The original analyses suffered from several caveats, solely focusing upon observer sensitivity statistics and not considering the observer choice function, or psychometric function. The analyses undertaken in this chapter remediate these shortcomings. The reason motivating this re-analysis is that the psychometric function facilitates an analysis of how dispersed the signal and noise responses are across the decision axis. Consider the step function that is assumed by SDT. Assuming observers are able to stabilise their decision criterion across trials, they would consistently assign signal responses for all stimuli above the criterion, and noise response for all stimuli below. The dispersion of signal and noise responses is thus confined to the area above and below the criterion, respectively. In the more likely case, where observers are not so consistent, signal and noise responses are dispersed across stimuli that reside both above and below the criterion. Such inconsistency in responding has an obvious effect upon the psychometric function whereby a slope is introduced, which becomes increasingly flatter as response inconsistency worsens. That is, variations in the degree of slope correspond to variations in response inconsistency, where the slope may be used to infer the degree of observer criterion variability (e.g., Zak et al., 2012). It is hoped that a focus on the psychometric function will shed more light upon the effects of feedback, beyond what could be assumed from assessing only SDT sensitivity statistics.

In conducting the analyses upon the psychometric function, a Bayesian approach is taken by fitting hierarchical models to estimate the model parameters. This approach applies a structured individual differences model which views each individual's slope parameter as a draw from a higher population level distribution (see Section 2.2.2, though more thorough introductions to Bayesian inference are available; e.g., see Lee & Wagenmakers, 2013). While differences in the group level slope distributions will be inherently useful, they tell us little about the process through which observer inconsistency, if any, arose. This is to say, while slopes in the choice function can be used to infer the presence of criterion variance, it is

unknown just *how* the criterion is shifting. Variability in the decision criterion has led to various dynamic extensions to the fundamental SDT model (cf. Section 1.5). In reality there may be a number of strategies underpinning observer criterion shifts, and in some cases these shifts may be far from systematic (e.g., lapses of attention, boredom, fatigue). When it comes to modeling these shifts the number of possible approaches becomes an alarming prospect. Instead, motivations toward shifts are generally limited to a small number of sensible strategies which are then amenable to modeling in some tractable way. One of these mechanisms, which was subsumed as a viable shift mechanism by Taylor (2010), is the error-correction process first introduced by Kac (1962, 1969) and later extended by Dorfman and colleagues (1971, 1975). However, the model was never fit to the original data, something this chapter will attempt to remediate. Like the psychometric function, model fitting will proceed by applying a hierarchical Bayesian version of the original Kac (1962) model, allowing for model comparisons to be made between dynamic and static criterion versions of SDT (covered in Section 2.2.2).

2.1. Summary of Taylor's (2010) Original Research

Taylor (2010) investigated the effects of providing different types of feedback to observers who were completing a binary categorization task under two levels of difficulty. The categorization task required observers to classify externally distributed tones as either high or low (cf. Lee & Zentall, 1966; Kubovy & Healy, 1977a; also see Section 1.3.2). Recall that stimuli distributed in such a way means that the evidential values of the stimuli are observable to both the experimenter and the observer. The tonal categories were fashioned such that they approximated, as best as possible, Gaussian distributions. Task difficulty was determined by varying the degree of overlap between the external distributions. Greater degrees of overlap mean that more stimuli can belong to either distribution and so correctly categorizing the stimuli becomes harder. The task is made easier by reducing the degree of overlap and so the stimuli become increasingly representative of their distribution of origin. Two levels of difficulty were used in the experiments, corresponding to a hard ($d' \approx 1$) and

easy ($d' \approx 3$) condition. The known characteristics of the stimulus distributions allowed ideal observer performance to be determined, providing an asymptotic level of performance against which observer performance could be gauged. In other words, the external noise method allows one to assess how far an observer departs from optimality. A further benefit is that the optimal criterion location may be specified. Taylor only considered the equal variance and equal base rate ($\pi = .5$) case which places the optimal criterion at the midpoint between the means of the stimulus distributions ($c_o = d'/2$). The availability of such knowledge (optimal criterion, distributional means), as will become apparent, is immensely useful in manipulating the way feedback is provided.

Taylor (2010) compared two types of trial-by-trial feedback, or what will from hereon be referred to as knowledge of results (KR). The first type represented what is effectively the status quo for KR, where the observer is simply told on each trial which distribution the stimulus had been drawn from (the observers were also informed before beginning that the probability of each stimulus being presented was .5). This type of KR is referred to as trial-by-trial knowledge of results for stimulus events ($TTKR_e$). In a more general sense, $TTKR_e$ amounts to informing the observer whether their decision was correct or incorrect. $TTKR_e$ is also inherently probabilistic and is completely determined by the prior probability of each stimulus event. Consider a sequence of stimuli presented to an observer. On each trial a stimulus is drawn from one of two distributions with equal probability. Over the course of the experiment there are trials where the observer will be presented with identical stimuli (e.g., tone 15), though the sampling distribution for the stimulus may have differed on each occasion. So, despite the stimulus possessing *exactly* the same evidential characteristics, the TTKR associated with the sampled stimulus will differ on each of these occasions. This facet is explicated in Table 2.1 (on the next page) which uses tonal stimuli as an example.

In this contrived example we see that on trial 1 tone 15 is sampled from the high distribution, whereas three trials later tone 15 is presented again, though this time it has been sampled from the low distribution. Another four trials later tone 15 is presented again,

Table 2.1:
Example trial sequences and associated TTKR.

Trial	Tone	Sample Dist.	$TTKR_e$	$TTKR_i$
1	15	High	High	High
2	17	High	High	High
3	11	Low	Low	Low
4	15	Low	Low	**High**
5	13	Low	Low	Low
6	22	High	High	High
7	19	High	High	High
8	15	High	High	High
9	17	Low	Low	**High**
10	12	High	High	Low

where this time it is again sampled from the high distribution. Considering for now only the $TTKR_e$ column, one can see that the TTKR simply repeats the sampled distribution column. In this example the observer is told that tone 15 is a high stimulus, then a low stimulus, and then high again. In contrast, the second type of KR provided observers with optimised feedback which was less influenced by the statistical properties of the stimulus distributions. This type of KR, then, makes use of the optimal criterion location and determines the appropriate stimulus feedback according to which side of the criterion the stimulus had been sampled from. That is, the KR is completely deterministic. Consider the optimal rule for the tone experiments: if any tone exceeds the fixed decision criterion then a high response is made, else a low response is made. Applying this rule to determine what KR to provide to an observer means that *all* stimuli that exceed the optimal criterion, irrespective of which distribution has been sampled, will be considered as a high stimulus. Conversely, all stimuli that fall below the criterion will be considered as low. This type of KR is referred to as trial-by-trial ideal knowledge of results ($TTKR_i$). Using the example in Table 2.1 again, let us assume that the optimal criterion is located between tones 13 and 14 ($c_o = 13.5$). Applying the optimal decision rule, consider the $TTKR_i$ column in Table 2.1. On trial 4 the observer is told tone 15 should be treated as a high stimulus. A similar situation arises on trial 9 where tone 17 is also treated as high though it had been drawn from the low distribution.

$TTKR_e$ can be problematic; the major issue being that across trials the observer may receive information that contradicts earlier information, or violates an expectation the observer has generated about the nature of the stimuli. In situations where the nature of the underlying stimuli is increasingly ambiguous the observer may place greater weight upon the information provided to them on each trial. Changes in the informational value associated with identical stimuli may lead to more frequent switches between responses to that stimulus, an effect tantamount to a criterion shift. Specifically, probabilistic KR should increase the frequency and magnitude of criterion shifts under conditions where the stimuli can be sampled from either distribution with greater frequency. Conversely, where there is little overlap between the stimulus distributions there is less inherent ambiguity, with fewer tones being drawn from the overlapping region. This entails a more consistent stream of KR information and so should all but eliminate any ill effect the $TTKR_e$ engenders.

The assumed antithesis is that an idealised stream of KR information should be impervious to the statistical nature of the stimulus distributions because it is deterministically defined. This should, then, circumvent the problem of (subjective) inconsistency, particularly under conditions where the origin of the stimulus is uncertain. Moreover, $TTKR_i$ should reduce the size and frequency of criterion shifts and reduce the overall degree of decision noise. In easier conditions the expectations for any benefit of $TTKR_i$ mimic those assumed when dealing with $TTKR_e$ because of the minimisation in distributional overlap. It should be noted, however, that while in any real world case, or indeed in conventional detection or discrimination paradigms, $TTKR_i$ is unrealistic (and impossible). Its value is to contrast just how the statistical properties of the underlying model can influence how the observer may process and use the incoming information. It provides another comparison against which changes in performance can be gauged, in addition to comparing it to the expected asymptotic level of performance. In summarizing these expectations, where there is greater distributional overlap between the tonal distributions, $TTKR_i$ was expected to result in better overall performance compared to $TTKR_e$. This effect was expected to be mediated by an increase in criterion variance induced by $TTKR_e$. Finally, where the

Table 2.2:

Mean Summary Estimates from Taylor's (2010) Experiment.

TTKR	Diff		HR	FAR	d'	c	η
Events	Hard	*M*	0.68	0.34	0.88	-0.04	0.88
		SD	0.03	0.04	0.11	0.07	0.11
Ideal	Hard	*M*	0.68	0.34	0.88	-0.02	0.88
		SD	0.03	0.02	0.10	0.06	0.10
Events	Easy	*M*	0.88	0.14	2.32	-0.05	0.77
		SD	0.04	0.04	0.27	0.13	0.09
Ideal	Easy	*M*	0.88	0.15	2.25	-0.05	0.75
		SD	0.04	0.04	0.24	0.13	0.08

Note: Diff = Difficulty; M = Mean; SD = Standard Deviation.

distributional overlap is reduced the effect of either type of TTKR is expected to be largely eliminated.

Taylor (2010) investigated these hypotheses the results of which are contained in Table 2.2. Considering first the hard condition, when only the sensitivity statistic (d') is evaluated we see no ostensible difference in performance between the $TTKR_e$ (M = .88) and $TTKR_i$ (M = .88) groups. However, the converse appeared to be the case in the easy condition, where the $TTKR_e$ group performed marginally better (M = 2.32) than the $TTKR_i$ group (M = 2.25). The original analysis proceeded by conducting a factorial 2x2 ANOVA on the estimated sensitivity values. Unsurprisingly, statistical analyses confirmed null effects for the interaction between task difficulty and TTKR type, $F(1, 39)$ = .39, p = .54, and for the main effect of TTKR type, $F(1, 39)$ = .38, p = .54. The effect for task difficulty was statistically significant, $F(1, 39)$ = 557.75, p < .001; however, this is of no theoretical significance because the design of the study was such that observers would necessarily perform differently across levels of difficulty. Perhaps a better analysis might have been to consider the efficiency measure, η, which places performance on the same scale and is simply the ratio between $d'_{observed}$ and d'_{actual}. Though this transformation does nothing to change the overall results - and nor should it - it does make comparisons between groups more interpretable. One can now speak about how efficient observers were across groups. So, when an ANOVA is

conducted using η as the dependent variable, while we achieve the same pattern of results, the effect for task difficulty, $F(1,39) = 17.45$, $p < .001$, is now interpretable and we may say that observers in the easy condition were less efficient than observers in the hard condition.

On the whole, the relationship between TTKR, criterion variance, and sensitivity does not appear to hold. Given that increased criterion variance is expected to suppress estimates of sensitivity, the negligible differences across these values makes it difficult to determine whether the TTKR had any genuine effect at all. While the presence of criterion variance *may* be inferred via comparisons between the mean sensitivity values, this is by no means a sensitive analysis. Given the data and the analyses conducted to this point, the evidence for an effect of TTKR is tenuous. However, the purpose here is to re-evaluate the data and subject them to more thorough analyses. To facilitate these analyses the following section will briefly discuss the analytic approaches that will be used.

2.2. Re-analysis: Evaluating the Psychometric Function

Signal detection theory tells us that the criterion with which observers make decisions is fixed across trials. In terms of evaluating observer choice probabilities across a set of stimuli that are increasing in magnitude, the tenets of SDT suggest that we should observe a clear delineation in responses whereby responses of one type (e.g., high responses) are exclusively found on one side of the criterion while all cases of the complementary responses are found on the other. Such human data are almost never observed. One exception was a study conducted by Kubovy et al., (1971) where observers completed a large number of binary choice trials with externally distributed stimuli. While the obtained psychometric functions very nearly approximated a step function, there were still some violations of the static criterion hypothesis (observing low responses for stimuli above the criterion and vice versa) which lead the authors to reject the static criterion assumption. In most cases, however, human observers depart drastically from optimality and produce psychometric functions that are sigmoidal, not stepped (e.g., Lee & Zentall, 1966).

A critical utility of the psychometric function is that it allows one to assess how dispersed an observer's responses are; it can also be viewed as the CDF of a distribution of trial-by-trial criterion values (cf. Schoeffler, 1965; Wickelgren, 1968), denoting the probability that the criterion was located at a particular stimulus value across time. It also allows one to assess how various factors might influence observers' responding, as in the case of Taylor (2010). Specifically, factors that are expected to improve observer responding should have an associated effect upon the dispersion of responses, and thereby the slope of the function. Applying this reasoning to Taylor's (2010) experiment, in difficult conditions one would expect $TTKR_i$ to have a positive effect upon responding in that it limits response dispersion. If the TTKR manipulations are having an effect upon the criterion, then $TTKR_i$ should *increase* the slope of the function, which is tantamount to a reduction in trial-by-trial criterion shifts. The choice function should begin to resemble more of a step function under conditions where the degree of decision noise is assumed to be mitigated. By this reasoning, where $TTKR_e$ is provided response dispersion is expected to be greater and the slope of the function should *decrease.* In the absence of any differential TTKR effect, as is the case in the easy condition ($d' = 3$), one would expect similar slopes across both TTKR groups. These are the expectations for the present analysis and provide a means of inferring variance in the observer decision criterion.

2.2.1. The Psychometric Function

In the most general sense, the psychometric function is a monotonically increasing function that maps the probability of a response conditional upon increasing values of stimulus magnitude. There are a number of ways the function may be modeled, though all require the specification of a location and scale parameter that is estimated when fitted to data. Accordingly, it can be based upon any number of distributional assumptions (e.g., the Normal, Weibull, or Cauchy; for examples see Knoblauch & Maloney, 2012), though for present purposes the logistic function will be used. In practice, what we wish to model is the proportion of high responses made for each tone value. Formally, for the i^{th} participant

responding to the j^{th} stimulus, denoting each tone by x_j, the probability of making a high response for each increasing tonal frequency is modeled as

$$P(R_{ij} = High \mid x_{ij}) = \frac{1}{1 + e^{-[\alpha_i + \beta_i x_{ij}]}} \quad \text{Eq. 2.1}$$

where α_i and β_i are free parameters that correspond to the intercept and slope of the logistic function. To ease the notational load for the remainder of the analysis, the term on the left-hand side of the equality can be treated as the rate of high responses across all tones for each observer, and is denoted θ_{ij}. The model may be linearized by applying a logit transform to the sigmoidal logistic function and can be fit as a Generalized Linear Model (GLM) with a logit link function, where

$$\text{logit}(\theta_{ij}) = \alpha_i + \beta_i(x_{ij}) . \quad \text{Eq. 2.2}$$

In practice, to fit the model is trivial and can be done using R's `glm()` function[6], though for the present analysis we will use a hierarchical Bayesian model to estimate the psychometric function parameters (details will be discussed in the next section). The slope and intercept parameters may also be transformed so that they can be interpreted as scale and location parameters, respectively; in other words, the mean and standard deviation of the assumed criterion distribution. The mean is $\mu_i = -(\alpha_i/\beta_i)$, and the standard deviation is found by taking the reciprocal of the slope, $\sigma_i = 1/\beta_i$. These transforms place the parameters on the stimulus scale (e.g., tonal magnitude) meaning the parameters are directly interpretable across groups. The standard deviation provides a measure of criterion variability and so is expected to vary in response to the TTKR manipulations (Zak et al., 2012, used a similar approach in inferring criterion variance is a dual detection task). The primary dependent variable is now the slope estimates. There is no *a priori* reason to expect the intercept (location parameter) to vary across TTKR groups, though, of course, it will change across levels of difficulty as the distributional means separate. Before moving onto the results of the

[6] R is a programming language used for statistical analysis and modeling.

re-analysis, the next section will briefly cover Bayesian parameter estimation and model comparison.

2.2.2. Bayesian Parameter Estimation[7]

The goal of Bayesian parameter estimation does not fundamentally differ from the conventional maximum likelihood estimation (MLE) approach; that is, we wish to find the value of a parameter, θ, or set of parameters, that are currently unknown. However, Bayesian estimation approaches the problem in a different way. The core difference lies in the Bayesian's explicit consideration of one's prior uncertainty about the possible value(s) the parameter(s) of interest may take. To compare, the MLE approach derives parameter estimates by evaluating an objective function (usually the log of the likelihood function) at various parameter values, or combination of parameter values. The best parameter estimate(s) is the value(s) that maximise (hence maximum likelihood estimation) the value of the objective function. The downfall of this approach is that uncertainty is not built into the model we are fitting. While parameter uncertainty can be inferred, it requires *post hoc* methods, such as bootstrapping, which generate samples based upon the already maximised parameter values.

In contrast, in Bayesian parameter estimation one specifies, *a priori*, the degree of uncertainty regarding the parameters before any data are observed. This is achieved by specifying a probability distribution of some form, called the *prior distribution* – denoted as $p(\theta)$ - which assigns probabilities to a range of parameter values. The canonical example is the uniform prior which assigns equal probability to all possible parameter values; this is usually called an *uninformative prior*. Parameter uncertainty is thus an integral part of the Bayesian statistical model, which is quite apart from conventional MLE. Moreover, Bayesian parameter estimation generates a distribution of likely parameter values, referred to as the *posterior distribution* – denoted $p(\theta \mid D)$ – rather than single values (D refers to data). The

[7] For a more thorough treatment on Bayesian statistics and modeling the reader should consult Lee and Wagenmakers (2013), on which this section is largely based.

posterior engenders both our prior degree of uncertainty and our now updated belief about θ having just observed some data. The transition from the prior distribution to the posterior distribution proceeds via Bayes' Theorem,

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}, \qquad \text{Eq. 2.3}$$

where $p(D \mid \theta)$ is the likelihood of the data and $p(D)$ is the marginal likelihood. For all practical purposes, when estimating parameters, the marginal likelihood, which does not involve the parameter of interest, can be excluded without any loss of generality. This then reduces Equation 2.3 to the Bayesian mantra which states that the "posterior is proportional to the likelihood times the prior", or $p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$.

When the prior and posterior distributions belong to the same distributional family (e.g., Beta distribution), the distributions are said to be *conjugate*, and in simple cases of conjugacy the transition from prior to posterior can be solved analytically (i.e., via closed form solutions). However, this is often only possible for very simple models; for more complex models we must derive the posterior via different means. This restriction inhibited the use of Bayesian inference prior to the advent of computer-based sampling methods (e.g., Markov Chain Monte Carlo sampling, or MCMC). Rather than deriving analytic solutions, sampling algorithms (e.g., Metropolis-Hastings or Gibbs sampling) approximate the posterior distribution by recursively generating sequences (chains) of values based upon the model specifications (i.e., the prior, and likelihood function). When the sampling becomes stationary (varies around a single numerical point) it implies the values are being drawn from the appropriate posterior distribution, where the mean (or sometimes median) of the posterior distribution reflects the most likely parameter value. Given the computational efficiency of MCMC sampling, it now means that many cognitive models of varying complexity may be fit using Bayesian approaches. Furthermore, there are a number of freely available programs that facilitate Bayesian parameter estimation and model fitting (e.g., WinBUGS and JAGS[8]).

[8] In this thesis only JAGS is used and be freely obtained from mcmc-jags.sourceforge.net.

2.2.3. Bayesian Model Comparison

It is often the case that we wish to make a judgment about which of a number of available models best describes a particular cognitive process. When we want to compare models of cognition we first must estimate the parameters for each model. Once we have the parameters the model is said to be fit, where the parameter values should tell us something substantive about the data we have. We must also consider how well each model has fit the data and so model deviance is a central component in model comparison. There are various avenues that can be taken in comparing models (e.g., comparing fit indices like the Akaike Information Criterion, AIC), though none consider the relative evidence of each model.

Bayesian model comparison, in an identical fashion to Bayesian parameter estimation, proceeds via Bayes' theorem, though now the parameters we are estimating are dependent upon some model of interest. Formally,

$$p(\theta \mid D, M) = \frac{p(D \mid \theta, M)\, p(\theta \mid M)}{p(D \mid M)}, \qquad \text{Eq. 2.4}$$

where M is the model of interest. The likelihood, $p(D \mid \theta, M)$, is usually used as a goodness-of-fit index and is used to assess the relative deviance of each model. Instead of simply comparing the deviance of either model, it is possible to compare the relative *evidence* of each model, and so unlike Bayesian parameter estimation, the marginal likelihood, $p(D \mid M)$, plays an important role. The marginal likelihood is a weighted average of the likelihood, evaluated at every point across the parameter space, and weighted by a prior probability for each parameter value. Mathematically, what we are doing when we average across all possible likelihood values is integrating out the model parameters. Assuming, for the sake of convenience, that the parameters are continuously defined, the marginal likelihood can be expressed as

$$p(D \mid M) = \int p(D \mid \theta, M) p(\theta \mid M) d\theta\,, \qquad \text{Eq. 2.5}$$

where θ represents the parameters of interest. In other words, we want to know the probability of the observed data given the parameterisation of the model. The hallmark of a good model is its ability to make a large number of quality predictions out of a much larger set of possible predictions. So, while more complex models may be able to make more predictions than a simpler model, a large number of these predictions may in fact be uninformative. The strength of evidence for any model, then, rests with its predictive ability.

The marginal likelihood quantifies the predictive performance for each model; the better the model is at predicting data, the stronger the evidence is for that model. To compare the relative evidence between two models, all one needs to do is divide the marginal likelihoods, which yields a value known as the Bayes Factor,

$$BF_{12} = \frac{p(D \mid M_1)}{p(D \mid M_2)}. \qquad \text{Eq. 2.6}$$

The Bayes factor determines to what degree the evidence favours one model over the other. For example, if we obtain a $BF_{12} = 10$ we can say that the data are 10 times more likely under M_1 than under M_2. Conversely, if $BF_{12} = 0.1$ then the data are 10 times more likely under M_2. It is generally considered that a $BF_{12} > 10$ indicates strong evidence in favour of the model in the numerator (noting that BF_{21} can be found by flipping Equation 2.6; see Jeffreys, 1961).

2.2.4. Fitting the Psychometric Function

This section introduces the Bayesian model that will allow estimation of the psychometric function's parameters. To reiterate, we first assume there exists a rate, θ_{ij}, that corresponds to the proportion of responses the i^{th} observer has made to the j^{th} stimulus, here denoted r_{ij}, out of a total number of stimulus trials, n_{ij}, noting that x_j is a stimulus of increasing magnitude. What we wish to model is the process that generates these data, which means we need to specify the likelihood of the data. Accordingly, for the i^{th} observer we assume that their responses are binomially distributed, $r_{ij} \sim Binomial(\theta_{ij},\ n_{ij})$, where the rate parameter, θ_{ij}, is in turn determined by the psychometric function at each stimulus level, x_{ij}. Assuming a logistic psychometric function, evaluation of the rate parameter at each level

of stimulus intensity is contingent upon the intercept, α_i, and slope, β_i, parameters that define the psychometric function. These are estimable by applying a logit transform on the rate parameter (cf. Equation 2.2).

Recall that in Bayesian estimation a prior distribution for each parameter of interest must be specified. The likelihood of the response data is assumed to be a binomial process governed by a binomial rate parameter, which is itself determined by a psychometric function defined by an intercept and slope parameter. So, in this case we require a prior on both the intercept and slope parameters in order to arrive at the posterior distribution for each. The first, and perhaps more common, approach assumes that all observers within a specific group share a common intercept, α (note the dropping if the subscript, i), and slope, β, parameter; that is, all observers share *common rates*. To achieve this, priors need only be specified on the parameter values themselves; for example, we might simply place Gaussian priors on both the intercept and slope parameters for each experimental group. In effect, what we try to find are the parameter values that best fit *all* the observers within the group - the values we would most likely obtain if we simply averaged across all observers. However, the weakness is that in doing so we are treating all observers as a fixed effect (no individual differences). Instead, we will consider a hierarchical Bayesian model, or what is often referred to as a structured individual differences model.

The promise of the hierarchical approach is that each observer has their own intercept and slope parameter (now α_i and β_i, respectively); the observer is treated as a random rather than a fixed effect. Furthermore, each α_i and β_i are now assumed to have been sampled from a respective higher population, or group, level distribution. Each group now has a population level distribution for each of the subject level parameters, α_i and β_i. The population level distributions, which are usually assumed to be Gaussian, are parameterized by a mean and a standard deviation, and are referred to as the model *hyper-parameters*. The hyper-parameters thus influence the subject level parameter values within each group. So, while each observer's parameters are independently sampled, and therefore can certainly be different, observers are also related to each other because they are assumed to have come

from the same population distribution. As a consequence, inferences made about one observer will influence inferences made about another (see Lee & Wagenmakers, 2013, Chapter 10). A further difference is that we no longer explicitly place priors on the subject level intercept and slope parameters. While there *are* priors on α_i and β_i, the subject level priors are determined by priors we must now place on the population-level distributions. That is, we must specify priors for the mean and standard deviation of the intercept ($\mu_\alpha, \sigma_\alpha$) and slope ($\mu_\beta, \sigma_\beta$) population level distributions across each experimental group.

In now making the hierarchical model explicit, we assume the population level distributions for the subject level parameters are normally distributed as follows:

$$\alpha_i \sim Gaussian(\mu_\alpha, \sigma_\alpha) \quad \text{Eq. 2.7}$$

$$\beta_i \sim Gaussian(\mu_\beta, \sigma_\beta). \quad \text{Eq. 2.8}$$

Of most interest are the differences between the population level parameters, and in particular the slope estimates, and so we need to specify priors on the population level distributions. The following priors were used for the fitting data reported here (see also Lee & Wagenmakers, 2013, Chapter 12):

$$\mu_\alpha \sim Gaussian(0, .01) \quad \text{Eq. 2.9}$$

$$\mu_\beta \sim Gaussian(0, .01) \quad \text{Eq. 2.10}$$

$$\sigma_\alpha \sim Uniform(0, 10) \quad \text{Eq. 2.11}$$

$$\sigma_\beta \sim Uniform(0, 10). \quad \text{Eq. 2.12}$$

The priors placed upon the hyper-parameters are considered uninformative. That is, they are non-specific in terms of any form of expectation regarding the likely, or intuited, parameter value. Conversely, informative priors would place the mass of the prior distribution over an expected parameter range, a range generated from past experiments and/or experience. It is also possible to use posterior distributions as prior distributions, where in iterative approaches the posterior distributions from previous experiments may be used as priors for

subsequent experiments (Kary, Taylor, & Donkin, in press). As the oft cited phrase goes, "today's posterior is tomorrow's prior" (Lindley, 1972, p. 2).

2.2.5. Fitting the Error-Correction Model

To fit the error-correction model (ECM; Kac, 1962) we require the likelihood of an observer making either a hit, h, or false alarm, f, on each and every trial. For the j^{th} participant on the i^{th} trial, h_{ij} and f_{ij} responses are the result of a binomial process that are determined by rates H_{ij} and F_{ij}, respectively; thus,

$$h_{ij} \sim Binomial(H_{ij}, n_{ij}^1) \quad \text{Eq. 2.13}$$

$$f_{ij} \sim Binomial(F_{ij}, n_{ij}^2) \quad \text{Eq. 2.14}$$

$$H_{ij} = 1 - \mathbf{\Phi}(c_{ij} - d'_j) \quad \text{Eq. 2.15}$$

$$F_{ij} = 1 - \mathbf{\Phi}(c_{ij}) \quad \text{Eq. 2.16}$$

where d'_j is the sensitivity index for the j^{th} participant, n_1 and n_2 denote high and low trials, respectively, and $\mathbf{\Phi}$ is the CDF for the normal distribution . The criterion, c_{ij} is deterministically updated on each trial according to the axioms of the ECM, and so

$$c_{i+1j} = c_{ij} + \Delta\delta_j \,, \quad \text{Eq. 2.17}$$

where δ_j is the magnitude of the criterion shift, and Δ specifies a shift matrix. For the error correction model shifts are determined by an S x KR matrix (Stimulus: ∈ {0,1} by KR: ∈ {0,1}, where 0 and 1 indicate low and high tones, respectively)

$$\Delta_{sk} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{Eq. 2.18}$$

Shifts only occur on trials where errors have occurred, , Δ_{01}, Δ_{10}, and in the direction specified. Note that we model the process according to the TTKR the observer received. It should be the case that running the $TTKR_i$ sequence through the shift matrix reduces how often the criterion shifts, as observers are expected to commit fewer errors. In fitting this model it is assumed that the shift magnitude is the same for both types of error, $\Delta_{01} = \Delta_{10} =$

δ_{ij}. It has been shown that this parameterization necessarily leads to probability matching (see Dorfman & Biderman, 1971; Thomas, 1975), and while it is certainly reasonable to assume that not all observers will probability match, the constraint should not be problematic for the simple purposes of evaluating shift differences across TTKR conditions. The model, then, requires three parameters to be estimated: d'_j, δ_j, and c_{0j}. The c_{0j} parameter denotes the initial criterion value, the starting point from which the criterion shift process starts. All individual level parameters are assumed to be draws from normally distributed population level distributions,

$$d'_j \sim Gaussian(\mu_d, \sigma_d) \quad \text{Eq. 2.19}$$

$$\delta_j \sim Gaussian(\mu_\delta, \sigma_\delta) \quad \text{Eq. 2.20}$$

$$c_{1j} \sim Gaussian(\mu_c, \sigma_c), \quad \text{Eq. 2.21}$$

where the priors placed on the population level mean parameters are:

$$\mu_d \sim Uniform(0, 4) \quad \text{Eq. 2.22}$$

$$\mu_\delta \sim Uniform(-1, 1) \quad \text{Eq. 2.23}$$

$$\mu_c \sim Uniform(0, 4) \quad \text{Eq. 2.24}$$

The priors placed upon the population level standard deviation are all identical, $Uniform(0, 10)$. All priors are considered uninformative; however, the specified priors possess restricted ranges that reflect some intuition about where the parameter values should fall. Therefore, they provide some form of information. However, because equal probabilities are assigned to all values in these ranges, there is no preference for any particular value; in this sense, these priors are uninformative (see Appendix A for model code).

2.2.6. Results

To get a general idea of the variability in the group psychometric curves each observer's raw response proportions (r_{ij}/n_{ij}) across stimuli are plotted in Figure 2.1. From eyeballing Figure 2.1 for the difficult condition it is immediately apparent that there is much

more variability in the curves for the $TTKR_e$ group compared to the $TTKR_i$ group. In fact, the observers in the $TTKR_i$ group appear to be very consistent. Critically, there does appear to be a difference in the slopes of the functions between the TTKR groups, with the $TTKR_i$ slope appearing steeper than the $TTKR_e$ slope. Considering now the easy condition, there is an evident degree of variability between the observers in each TTKR group, though on the whole the slopes appear to be fairly consistent across both the TTKR groups. This pattern of results is consistent with the expectations outlined earlier. Turning to the results from the psychometric fits, the population level posterior distributions were generated by running three MCMC chains and taking 10,000 samples from each (after 500 samples burn-in[9]).

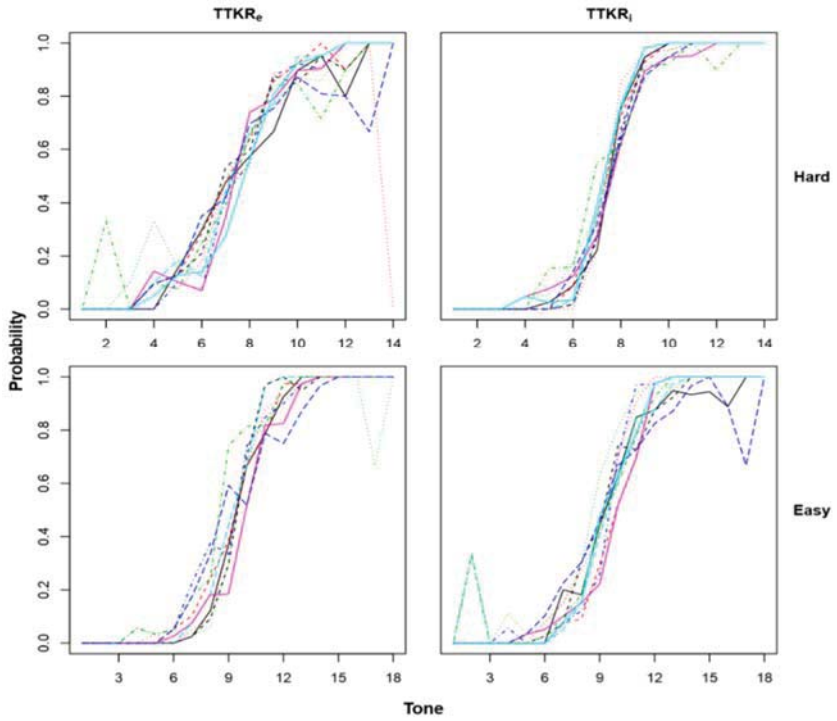

Figure 2.1. Plots of observer psychometric curves for each experimental conditions; N = 11 across all conditions. The proportion of high responses is plotted as a function of stimulus magnitude.

[9] Burn-in refers to the discarding of a nominal portion of the incipient MCMC samples so that any autocorrelation in the sampling is mitigated.

Table 2.3:

Psychometric Function Maximum A Posteriori (MAP) Estimates for Population Level Posterior Distributions.

Group	α	2.5%	97.5%	β	2.5%	97.5%	μ	σ
1	-5.79	-6.18	-5.42	.78	.74	.83	7.40	1.28
2	-11.93	-14.22	-10.67	1.60	1.43	1.90	7.45	.62
3	-9.88	-11.50	-8.42	1.06	.90	1.23	9.36	.95
4	-9.09	-10.25	-8.09	.97	.87	1.09	9.37	1.03

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy

For the analyses reported here sampling was conducted using JAGS (model code can be found in Appendix A). Parameter estimates are typically based upon the mode of the parameter posterior distribution. These are the values considered to be most likely based upon our prior uncertainty and are referred to as the maximum a posteriori (MAP) estimates; these are displayed in Table 2.3. The percentage values contained in Table 2.3 refer to the Bayesian 95% credible interval. The credible interval denotes the range of the posterior distribution between the 2.5 and 97.5 quantiles. That is, we can say we are 95% confident that the parameter value lies in this range.

The μ and σ values pertain to the mean and standard deviation of the criterion distribution, and are found by transforming the intercept and slope parameter as discussed previously in Section 2.2.1. To facilitate future discussion, note that the optimal criterion in the difficult condition is located between tones 7 and 8 ($c_{opt,hard} = 7.5$), whereas in the easy condition it is located between tones 9 and 10 ($c_{opt,easy} = 9.5$). Additionally, note that unit standard deviation is approximately equal to two tones. The location parameter estimates the midpoint of the function, corresponding to the mean criterion location. It can be seen that there is no great difference in criterion location between the TTKR groups across the respective levels of difficulty. It can also be seen that all groups are slightly lax in their criterion placements, though the $TTKR_i$ group in the hard condition is slightly more optimal.

There is an obvious difference in the scale parameters between the $TTKR_e$ (σ = 1.28) and the $TTKR_i$ (σ = .62) groups in the hard condition. There is also a difference between the scale values for the $TTKR_e$ (σ = .95) and the $TTKR_i$ (σ = 1.03) in the easy condition, though the magnitude of this difference is much smaller. What these values imply is that observers who receive $TTKR_e$ in the hard condition shift their criterion by as much as 1.28 tones, whereas observers receiving $TTKR_i$ only shift their criterion by .62 tones; $TTKR_e$ observers have twice as much variance as $TTKR_i$ observers. In the easy condition both groups are shifting their criterion by approximately one tone. The differences between the TTKR groups in the hard condition are more manifest when the posterior distributions are considered. The complete population level posterior distributions for the slope parameters were also transformed so that they are expressed on the stimulus scale, where they are displayed as violin plots in Figure 2.2.

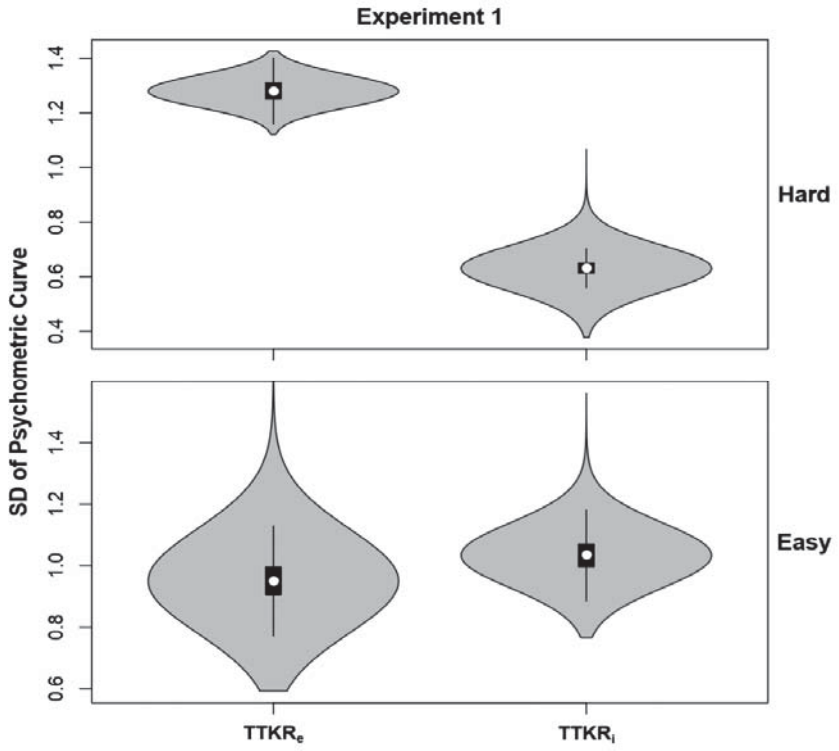

Figure 2.2: Population-level posterior distributions for the transformed slope so that that they are expressed on the stimulus scale. The critical feature is the difference between the distributions across the $TTKR_e$ and $TTKR_i$ groups in the hard condition, where the average deviance in the $TTKR_i$ condition is lower than the $TTKR_e$ deviance.

A violin plot is essentially a boxplot surrounded by a smoothed histogram which denotes the density of the distribution at various parameter values. From the figure it is immediately apparent that there exists a difference in the posterior distributions between the TTKR groups in the hard condition, suggesting that differences that appeared to be present in Figure 2.1 are genuine, and that $TTKR_i$ improved the overall ability for observer to categorise their responses. These results from the re-analysis suggest that variability in the criterion was at its greatest where observers received $TTKR_e$ in the hard condition. As for the TTKR groups in the easy condition, the amount of variability did not appear to be differentially affected by the type of TTKR. From the analysis on the slope of the psychometric functions we are left with the impression that there was more decision noise for the observers receiving $TTKR_e$. However, changes in the slope of the function cannot provide information regarding how the criterion shifts. Next, we will see whether this effect is captured by the fits from the error-correction model. The posterior distributions for the ECM fits were generated from four chains, each consisting of 10,000 samples, with a burn-in of 1000 samples. The MAP estimates from the ECM fits are contained in Table 2.4. The parameters of most interest are the population level shift magnitudes, δ, which estimates the size of the shift for each group of interest. While there is evidence that the criterion is shifting following errors, the size of the shifts is quite negligible. However, there is precedence for such small shift values (see Dorfman & Biderman, 1971; Dorfman et al., 1975; Dusoir, 1980), and when one considers the scale on which the magnitude of shift lies there does appear to be some effect.

Table 2.4:

MAP Estimates from Population Level Posterior Distributions for the ECM.

Group	c_1	2.5%	97.5%	d'	2.5%	97.5%	δ	2.5%	97.5%
1	.39	.31	.47	.88	.79	.98	.009	-.002	.024
2	.49	.40	.60	.89	.79	.98	-.001	-.013	.011
3	1.10	1.00	1.20	2.30	2.19	2.41	.003	-.013	.025
4	1.08	.99	1.16	2.23	2.12	2.34	.008	-.009	.029

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy; MAP = Maximum A Posteriori.

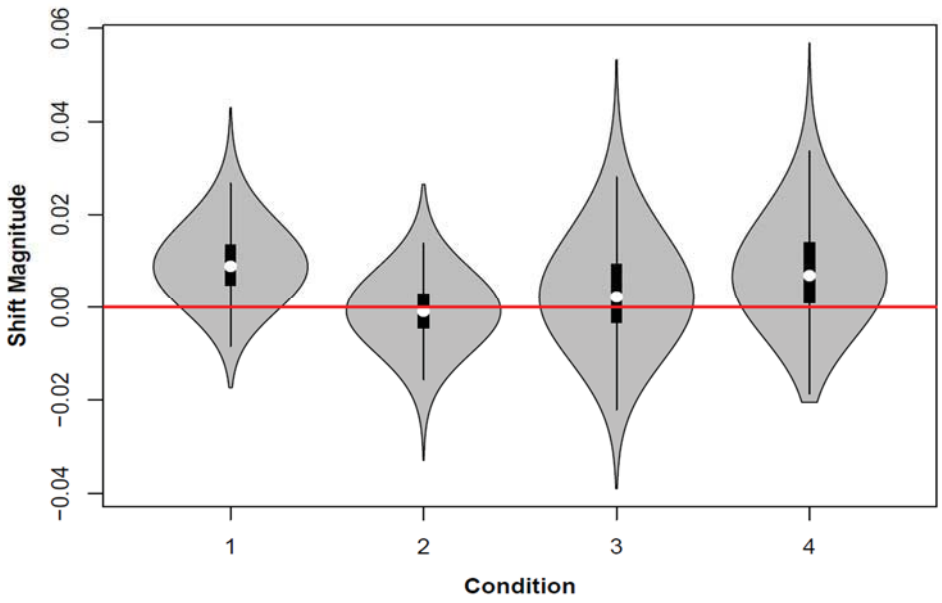

Figure 2.3: Violin plots of the population level posterior distributions for the shift parameter, δ, across groups.

The population level posterior distributions for δ across each condition are displayed in Figure 2.3. From a qualitative perspective, we might take it as encouraging that the shift parameters in the hard condition are in the expected direction, and that the $TTKR_e$ exhibited greater criterion variance than the $TTKR_i$ group. Interestingly, the effect reverses for the TTKR groups in the easy condition and it appears that the observers in the $TTKR_i$ group exhibit greater variance in their criterion. In an ideal world the shift parameters would be very similar between these two groups as the type of TTKR is not expected to influence the criterion in any differential way. What this might be is stochastic variation around the zero point. But how reliable are these differences? Are the differences meaningful? Do they sufficiently depart from zero?

We can approximate the posterior effect size by taking the difference between the shift magnitude posterior distributions for each TTKR group (i.e., between groups 1 and 2 in the hard condition and groups 3 and 4 in the easy condition). The resulting posterior distribution of differences is then compared with a zero-centred prior distribution which reflects null difference. That is, we take the probability density for the prior distribution at

the value of zero and compare this with the probability density for the posterior difference, also at the value of zero. The ratio of these two values is the Bayes factor and indicates the relative evidence for the difference under each hypothesis (the null and the alternative). This test is referred to as the Savage-Dickey Density Ratio test (see Lee & Wagenmakers, 2013). In testing the difference between the shift magnitudes, a zero-centred Cauchy distribution was used as the prior and compared to the posterior differences. In demonstrating this procedure, note that the density of the prior distribution at zero is .32. When the difference between the shift magnitudes are considered for the hard and easy condition, we find that the density of the differences at zero are approximately 24.3 and 38.1, respectively. We obtain the Bayes factor by dividing the posterior difference by the prior, where we find that the differences displayed in Figure 2.3 are $BF_{01} \approx 76$ and $BF_{01} \approx 120$ times more likely under the null hypothesis. We might also repeat the same procedure for each shift magnitude posterior distribution to assess whether the shift magnitude for each condition reliably differs from zero. Unsurprisingly, the differences are $BF_{01} \approx 66$, $BF_{01} \approx 233$, $BF_{01} \approx 143$, and $BF_{01} \approx 118$ more likely under the null for Groups 1-4, respectively. What these parameters tell us is that a model where the criterion, in effect, is not shifting is more consistent with the data. So the predictions made by the error-correction model differ very little from the predictions made by the normative signal detection model where the decision criterion is, of course, fixed.

We might also consider the sensitivity parameter estimated from the error-correction model fits in assessing the utility of the model (cf. Table 2.4). If there is additional noise in the data that the model is expected to account for, then we would expect an improvement in the estimated sensitivity values across each group. This expectation exists because decision noise spuriously decreases sensitivity estimates. However, when the d' estimates in Tables 2.2 (the standard SDT estimates) and 2.4 (the ECM estimates) are compared there are only trivial differences between them; Group 1: .88 vs .88; Group 2: .88 vs. .89; Group 3: 2.32 vs. 2.30; and Group 4: 2.25 vs. 2.23. So, in fact, we see that the estimates become slightly *worse* for Groups 3 and 4. This finding agrees with the conjecture that the fitted error-correction model is

making predictions that are consistent with a static criterion. Before drawing any further conclusions it will be best to formally compare the static criterion model (SCM) – the stock standard SDT model - and the error-correction model (ECM). But before doing that, the parameters for the SCM will need to be estimated using the same hierarchical Bayesian framework used to estimate the ECM parameters. This is fairly straightforward, requiring only a simple restriction to be placed upon the model discussed in Section 2.2.5. The restriction is that the criterion remains the same across all trials, facilitated by setting all shift magnitudes to zero. All other priors remain unchanged. The resulting parameter estimates were identical to those contained in Table 2.2, though now we have posterior distributions for the sensitivity and criterion estimates across all groups. In terms of formal model comparisons, there are two approaches one could take.

The first is analogous to conventional model comparisons and involves an assessment of the relative model deviances; that is, the relative discrepancy between the empirical data and the corresponding model estimates (the *goodness-of-fit*). This is usually quantified by the likelihood, $P(D \mid \theta, M)$ – the probability of the data given the model parameters. The smaller the deviance is the better the model is assumed to be; however, not all models are created equally, and often models will differ in their complexity. Complexity usually refers to the number of free parameters a model has. If a model has few free parameters the model is constrained and not overly flexible. Conversely, a model that has more free parameters is less constrained and can be very flexible. The issue is that flexible models may be hard to falsify and can potentially fit a wide range of data, to the point that the model overfits the data. On the other hand, some models may be too constrained and unable to account for broader data patterns. The idea, then, is to find a model with the least number of parameters that can also account for the data adequately. In pursuing this goal models are usually penalized according to the number of free parameters required to fit the model, for example, using the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). While these serve as a general arbiter of parsimony, they are not appropriate for hierarchical Bayesian models.

The Deviance Information Criterion (DIC) is a model selection tool that is particularly useful for hierarchical models where the posterior distributions have been obtained through MCMC sampling. Much like the AIC and BIC, the DIC first assesses the deviance between the data and the model, and then penalizes the model according to its relative complexity. The way the DIC does this, however, is slightly different. Where the AIC and BIC require the actual number of parameters in penalizing the model, the DIC applies a penalty based upon the *effective* number of parameters, denoted as pD. This can be calculated in one of two ways, though when JAGS is used in generating posterior distributions the effective number of parameters is calculated as $pD = \mathrm{var}(D)/2$, where D is the model deviance. The penalty term is usually two times the number of effective parameters, and this is added to the model deviance; that is, $DIC = D + 2pD$. The general criterion in model selection is that the model that returns the smallest DIC provides the most parsimonious account and should therefore be preferred. The DIC provides a useful heuristic for model selection, and is cases where obtaining model marginal likelihoods is intractable it may be the only tool available. Nevertheless, the DIC is also susceptible to the same model comparison problems as both the AIC and BIC inasmuch as simpler models may be unfairly favored. Fortunately, the marginal likelihoods for both the ECM and SCM can be obtained and so model comparisons can proceed by the second, and preferred, approach by comparing the relative evidence for each model using Bayes factors.

For the case considered here we are comparing the SCM, denoted as M_1, with the ECM, denoted M_2. As detailed in Section 2.2.3., the Bayes factor is obtained by taking the ratio of the marginal likelihoods for each model: $BF_{12} = p(D|M_1) \,/\, p(D|M_2)$[10]. While comparing the relative evidence is very straightforward, obtaining the marginal likelihood for each model is not as easy. This first step requires sampling a large number of values (100,000 were used

[10] One may note that when the shift magnitudes were being compared the subscript for the Bayes factor was 01, where now the subscript is 12. When a null and alternative hypotheses are being considered the null is conventionally assigned a zero and the alternative a 1. Conversely, when cognitive models are being compared we arbitrarily assign one model as Model 1 and the other as Model 2. There is no reason why one model cannot be considered as the "null" and the other the "alternative", though these are the conventions used throughout the literature.

here)[11] from the specified prior distribution for each parameter of interest (see Section 2.2.5). Next, we assess the likelihood of each observer's trial data for each possible parameter combination by plugging the sampled values into Equations 2.15 and 2.16, and do so for both models under consideration. Therefore, a likelihood is calculated for the i^{th} observer, across n trials, for 100,000 unique parameter combinations. Usually the product of the hit and false alarm likelihoods are obtained, though with such large numbers it is often easier to deal with log-likelihoods which require the summation of likelihoods. Finally, the trial-by-trial hit and false alarm likelihoods are summed and the average taken across the resulting 100,000 values (i.e., the average likelihood of the data under 100,000 different parameterisations), which provides us with the marginal likelihood for each observer (see Kary et al., 2015; code for these calculations can be found in Appendix A). The observer Bayes factors are displayed in Figure 2.4.

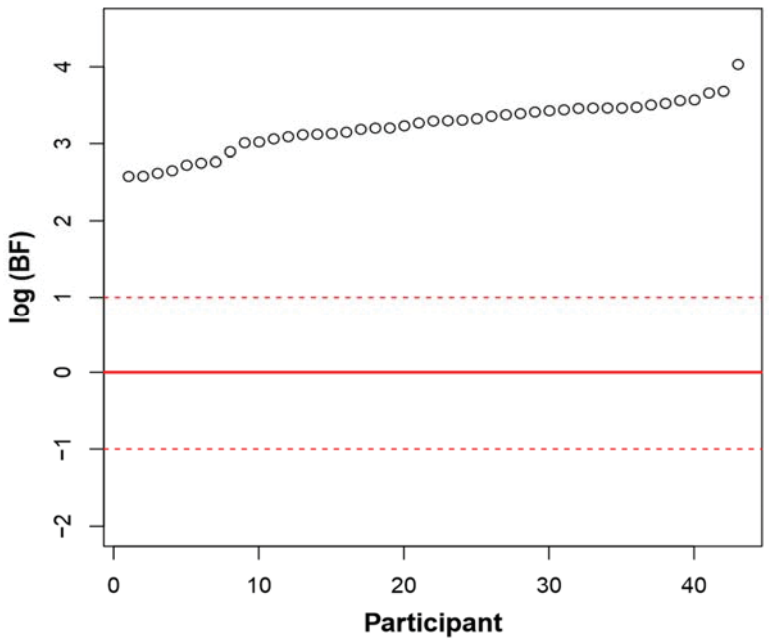

Figure 2.4: A plot of the logarithm of the BF_{12} for all observers. The solid line at zero denotes the switch point for evidence in favour of either model, where all values > 0 provide evidence in favour of M_1: SCM. The area within the dotted line denotes the range of values that would provide little differential evidence for either model. It can be seen that all observers were favoured by the static criterion model.

[11] Which really is not enough; ideally samples should be in the millions in order to sufficiently cover the parameter space. However, the sampling and storage of such large numbers was computationally prohibitive in the present study and exceeded the available RAM.

It is not uncommon for BF values to be very large, so Figure 2.4 displays the logarithm of the BF. Additionally, the values have been ordered, despite the points corresponding to observers across all groups. The reason why this has been done is that there were no differences in model preference across groups, and it is clear that the evidence is unequivocally in favour of the SCM model. The dotted lines in Figure 2.4 denote the region where the evidence would be considered as agnostic toward any particular model; such agnosticism is clearly absent given the current results. Across all participants, the Bayes factors ranged from approximately 13 to 56. What this result implies is that the ECM either made predictions that were similar to the SCM, though ostensibly still poorer (the lower Bayes factor range), or predictions that were just completely inconsistent (the higher Bayes factor range). On the whole, the evidence in favour of the ECM is non-existent.

2.3. Discussion

The foregoing re-analysed data from a probabilistic categorisation task, focusing upon how observers distributed their responses for increasing stimulus magnitudes, and whether types of feedback had an appreciable effect upon observer responding. Inconsistencies in observer responding were evaluated by fitting a psychometric function to the response data. Under the assumption of consistent responding, the psychometric function should approximate a step function, and is the form assumed under SDT. However, in all cases the empirical response function departed from a step function, implying that observers were not consistent with their responding. Furthermore, the slope of the psychometric function was sensitive to manipulations in the type of feedback, particularly under hard conditions. The critical finding was that, under conditions of difficulty, the psychometric slopes were steeper and more like a step function when $TTKR_i$ was provided compared to $TTKR_e$. This change in slope between the TTKR groups suggests a reduction in criterion variability. Conversely, in the easy condition the type of feedback had no differential effect upon the psychometric slopes, estimates agreeing with the TTKR conjecture; that is, $TTKR_i$ reduces variability in observer responding where the stimuli are increasingly ambiguous. The results reported

here also converge with previous research that has investigated decision behaviour using externally distributed stimuli. Lee and Zentall (1966) also found that the slope of their observer choice function changed in response to feedback manipulations; though in their study TTKR was either presented or it was not. Lee and Zentall found that observers performed better in a difficult categorisation task when TTKR was removed (see also Schoeffler, 1965). Their analysis is limited insofar as the change was inferred by visual analysis, though there were reliable changes in performance statistics (d') across difficulty levels which lends weight to the inference that TTKR influences performance. One clear point of enquiry for Taylor's (2010) data is the fact that, if observers were shifting their criterion twice as much as ideal observers, why was there no concurrent change in sensitivity between the groups?

Mathematically, criterion variance is expected to reduce the estimated d' by a factor of $\sqrt{1+\sigma_c^2}$, where σ_c^2 denotes criterion variance (see McNicol, 1972; Schoeffler, 1965; Wagenaar, 1973). Therefore, what one estimates is

$$d'_{obtained} = \frac{d'_{actual}}{\sqrt{1+\sigma_c^2}}. \qquad \text{Eq. 2.25}$$

One possible explanation of why this result did not pan out is due to sampling variability. Had more participants been run, then perhaps the effect may have presented itself in the data. One way to address this problem is to simulate observer performance based upon the estimated values of decision noise (in Table 2.3). A small simulation was run in which 10,000 observers for each TTKR were simulated, where each observer completed 200 high and 200 low trials. On each trial 'high' responses were generated using the following formulae:

$$HR = 1 - \mathbf{\Phi}(c - d') \qquad \text{Eq. 2.26}$$

$$FAR = 1 - \mathbf{\Phi}(c), \qquad \text{Eq. 2.27}$$

where $d' = d'_{actual}/\sqrt{1+\sigma_c^2}$, and σ were the slope estimates contained in Table 2.3 (for code see Appendix A). The simulated proportion of hits and false alarms were then

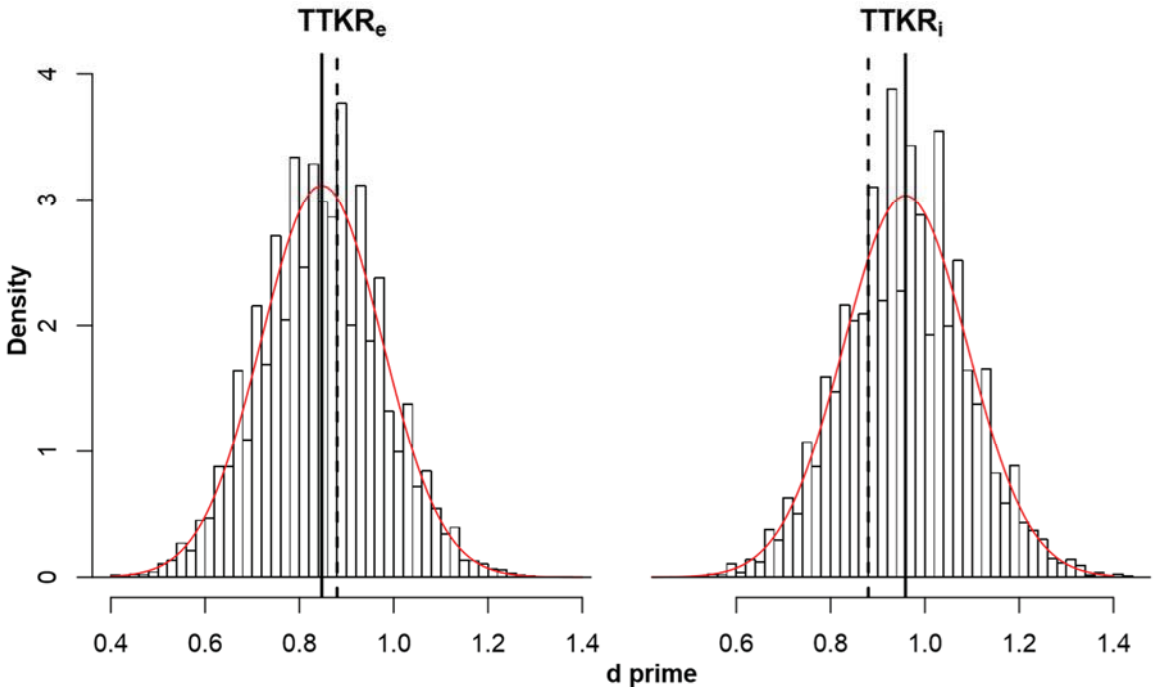

Figure 2.5: Estimated d' values from 10,000 simulated observers per TTKR group. The distributions reflect sampling distributions for d' as a function of KR type. The solid black lines indicate the means of the distributions, while the dotted lines indicate the estimated d' from Taylor's (2010) data.

divided by the total number of event trials and recorded as the estimated observer rates, which can then be used to estimate d'. Figure 2.5 displays the simulated sensitivity values. The simulated mean for the $TTKR_e$ group ($d' = .86$) is close to the group mean of .88, whereas the simulated mean for the $TTKR_i$ group ($d' = .97$) is quite different from the group mean of .88. Moreover, the simulated mean more appropriately reflects what would be expected if observers were exhibiting less inconsistency in their decision making, and implies that had more real observers been sampled the effect would manifest itself in the sensitivity estimates.

While the differences between the slopes point to an effect of TTKR in the hard condition, the question arises as to what possible processes could give rise to the data, in this case the slopes of the psychometric functions. From a cognitive perspective, what could the observer be doing across trials that forces shifts to occur in their decision criterion? Taylor (2010) surmised that the observers may be error-correcting, though this remained

speculative in the absence of model fitting. This caveat was addressed in the present re-analysis by fitting an error-correction model to the observer response data. On a qualitative level the direction of the shift parameters, at least across TTKR groups in the hard condition, agreed with the putative effects of TTKR. However, the differences were found to be trivial and of no substantive value. This, too, was the case for the differences between the shift parameters across TTKR groups in the easy condition. Furthermore, when the ECM was formally compared with the conventional SDT model the evidential balance swung heavily toward the SCM.

The Bayes factor, in effect, summarises the relationship, or consistency, between model predictions and observed data. The ECM assumes that the criterion shifts following errors and so includes an additional parameter to facilitate this process. However, this increases the complexity of the model and allows the model to make more flexible predictions. It would seem, then, that the additional shift parameter provided no improvement in predictive ability and is essentially redundant. Moreover, the ECM made predictions that were less consistent with the observed data when compared to the SCM. In light of such apparent differences between the slopes of the psychometric functions this result might seem counterintuitive, though this "null" result speaks more about model specification rather than whether or not the criterion was shifting. Put another way, the failure of the ECM to account for the observed data is not evidence that the criterion is invariant, but instead suggests the ECM is not specified correctly, or is just not appropriate. This conclusion is bolstered when one realises that the noise that does exist in the data can only be ascribed to decision noise. What can we conclude from the analyses undertaken so far, then? It is certainly clear that the observers were inconsistent in their responding, though the processes through which these inconsistencies arise were not consistent with the error-correction model as it is currently specified. The question remains, what type of model might provide a better description of the response data we have? This question provides the focus for the following chapter.

Chapter 3

Bayesian Model Comparisons of Dynamic Criterion Models

In tasks that require observers to detect or categorize perceptual stimuli they are often provided with feedback, or TTKR, to facilitate the learning process, and thereby performance. From the re-analysis of Taylor's (2010) data it is apparent that the type of TTKR that is provided during the task can affect how the observer responds across trials. This was the case in conditions where the confusability between the stimuli was high, whereas the effect was negligible in the easier task. The effect was seen most clearly when the observer's psychometric function, or choice function, was assessed. However, while there were differences between the slope parameters for the TTKR groups in the hard condition, a more complete understanding requires an investigation of the cognitive processes that most likely generated the data. We can infer that the $TTKR_e$ exhibited greater variance in the criterion compared to the $TTKR_i$ group, but what is driving the trial-by-trial criterion changes is unknown. To shed some light on this issue an error-correction model (Kac, 1962) was fit to the data. This model assumed that the differences between the TTKR group data arose through variations in the error-correction process as a function of KR type, where the $TTKR_e$ observers in the hard condition were expected to shift their criterion more often following errors. Conversely, the $TTKR_i$ observers were expected to shift their criterion less. The parameter estimates were qualitatively sensible with respect to expectations under the influence of TTKR, though the model did not fit the data well and made predictions that were not consistent with the data. The results from model comparisons saw, then, the static criterion model win out.

Part of the reason why the error-correction model may not have been able to account for the data is that the model is not specified with respect to the task demands of probabilistic categorization. If this were an orthodox detection task (i.e., the canonical tone in noise experiment) no more could be said, and we might simply conclude that the decision criterion was best modeled as a fixed value. However, we have only considered some of the information available to the observer, and treated the data as if they were conventional detection data. In doing so, we worked only with the stimulus and response sequences – which are, of course, the only pieces of information available after a tone in noise experiment;

however, this means that the model is constrained and assumptions must be made regarding how the criterion shifts should be characterized. For example, the usual assumption is that the criterion shifts by a constant amount in both directions following errors (and by a smaller constant amount following correct decisions in the general case; see Dorfman & Biderman, 1971), the direction of which is determined by a shift matrix. These restrictions are necessary, given the data, though some considered the idea that the internal stimulus effect might also affect how the decision criterion shifts (e.g., Thomas, 1975; see also Treisman & Williams, 1984). A hypothesis like this is certainly reasonable, though making any substantive claim regarding it is usually precluded because one cannot measure, and so cannot model, the internal sensory effect.

The probabilistic categorization task, however, makes available the evidential values upon which responses are based on each and every trial, and serves as a proxy for the internal sensory effect. The criterion may then be modeled in ways that allow the stimulus values to influence both the direction and magnitude of criterion shifts without having to introduce additional parameters. Moreover, on intuitive grounds it seems reasonable to expect that observers are using the stimuli in order to maintain some criterial value. One way in which this can be facilitated is by treating the previous stimulus value (the stimulus presented on trial $n-1$) as the criterion for trial n. The observer now merely decides whether the current stimulus is higher, or lower, relative to the previous stimulus. Having stimulus magnitudes available affords the error-correction model more information to work with, which may improve the predictive ability of the model. Furthermore, if it turns out that the model can predict the data better than the error-correction model covered in the previous chapter, we have learnt something about the cognitive processes underlying the shifts in the criterion – at least for the case of probabilistic categorization.

This chapter will present the results of a follow-up study to Taylor's (2010) original study. The aim of study was largely to replicate the effects described in Chapter 2 and further substantiate the effect of TTKR. Additionally, it provides new data with which model comparisons can be made. In the section that follows, how the ECM may be modified to allow

for the inclusion of the stimulus information will be discussed. Then the details of the study will be covered, after which the analyses and model comparisons will be discussed.

3.1. Mathematical Models of Criterion Shifts

The major point of departure between the dynamic criterion models discussed previously and the ones presented in this chapter is that the shift mechanism is determined, in part, by the stimulus information on each trial. The basic form of the dynamic criterion model (or additive learner model: ALM) remains unchanged; that is, the criterion is shifted across trials, typically following errors, by some amount. However, by allowing the stimulus information to be included we achieve an element of flexibility in the way the criterion shifts, and this is gained without adding any extra parameters. How the stimulus information is used is relatively straightforward. On each trial the stimulus is compared with the current criterion, where some distance exists between the stimulus and the criterion, ($x_n - c_n$). Instead of assuming the criterion shifts by some fixed amount, we instead assume that the criterion shifts by some amount proportional to the total distance between the current stimulus value and criterion. The trial-by-trial addition of some, or all, of this value in effect changes how the criterion behaves, or is characterised. The original characterisation views the criterion as a stationary Markov process; including stimulus information sees the criterion as more of a weighted average, or standard (e.g., Helson, 1947; Durlach & Braida, 1969).

Thomas (1975) refers to the weighted standard as the "pure error correction" model, the reason being that any criterion shift that is a function of the difference between a comparison (the stimulus) and a standard (the criterion) will naturally shift toward the stimulus itself (c shifts toward x), where if an error has occurred the shift is in accordance with the expectations for error correction. With this in mind, consider the following pure error correction model (PECM),

$$c_{i+1} = c_n + \Delta\delta(x_n - c_n)\,, \qquad \text{Eq. 3.1}$$

Table 3.1:

A Summary of the Dynamic Criterion Models Evaluated in this Chapter.

	Parameters		
Model	No.	Fixed	Criterion Shift
SCM	2	$\delta = 0$	NA
ECM	3	NA	$c_n + \Delta\delta$
PECM	3	NA	$c_n + \Delta\delta(x_n - c_n)$

where x_n is the stimulus value on $the\ n^{th}$ trial. The form of the model is very similar to the ECM introduced earlier, though there are two major differences. The first is that the δ parameter no longer corresponds to shift magnitude but is instead a constant of proportionality. This means that the criterion shift is some portion of the total difference between the stimulus and criterion. Thus, the second difference is that shifts are no longer constant across trials. The direction of the shifts are further determined by the sign of the difference, where the matrix, Δ, no longer determines the direction of shifts, but rather constrains the model so that shifts can only occur following errors. It then only takes on positive values for trials where errors have occurred, and so

$$\Delta_{sk} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \qquad \text{Eq. 3.2}$$

With this in mind, we now have three substantive models of interest: the standard signal detection model, or static criterion model (SCM); the standard error-correction model (ECM: cf. Kac, 1962); and the pure-error correction model (PECM: cf. Thomas, 1975). The use of acronyms in delineating the different types of dynamic processes may seem a little overwhelming, though hopefully the terms are relatively intuitive. To ease the load on memory the models have been summarized in Table 3.1.

3.1.1. Prior Distributions

While the details involved in the fitting of the models remain largely unchanged, the prior distributions used in this chapter will be slightly different. The differences are fairly

small, though they aid in constraining the models. In Chapter 2 the priors placed on the population-level parameters were broad, which means that values that are unlikely to occur fall within the model's prediction space. This is not particularly useful and so models may be unfairly penalised by virtue of the prior distributions that were specified. With this in mind, in this chapter we "reign in" our prior distributions, which are guided by the following intuitions.

Across both conditions of difficulty a uniform prior was specified for the mean sensitivity parameter which ranged from 0 to 4. In neither condition would we expect sensitivity to reach 4; both are bounded by an asymptote that precludes performance extending beyond the limit. Accordingly, the prior distribution should also be bounded by the same limit. It is also reasonable to assume that observers will always have better than chance sensitivity ($d' > 0$), and so the lower bound on the prior should reflect this. If we consider the individual data from Chapter 2, we may reasonably expect that performance in the hard condition falls comfortably between 0.5 and 1, whereas performance in the easy condition falls somewhere between 2 and 3. These assumptions and expectations are then modeled by placing uniform priors across this range; that is, our priors now are $\mu_d \sim Uniform(0.5, 1)$ for the hard condition and $\mu_d \sim Uniform(2, 3)$ for the easy condition.

Modifications were also made to the priors for the population-level initial criterion mean. Previously, the uniform prior for this parameter ranged from -2 to 2, which covers a significant proportion of the decision axis. However, considering that each event occurs with equal probability it would seem unnecessary to expect that the initial criterion value would extend far beyond the optimal criterion location (i.e., $c_1 \approx 0.5$ for the $d' = 1$ condition and $c_1 \approx 1.5$ for the $d' = 3$ condition). We model these assumption by now placing a constrained uniform prior on the initial criterion parameter, where $\mu_c \sim Uniform(0, 1)$ is assigned for the hard condition and $\mu_c \sim Uniform(1, 2)$ for the easy condition.

Finally, the ECM requires a shift parameter that determines the increment by which the criterion is adjusted. The prior originally used ranged from -1 to 1, though in light of the estimates obtained from the model fits in Chapter 2 it seems that this range is unnecessarily

large. Accordingly, the population-level shift prior is constrained to lie between -.5 and .5 and is uniform across this range, $\mu_\delta \sim Uniform(-.5, .5)$. With regard to the PECM, all priors described in this section apply, though instead of a shift magnitude parameter we require a shift proportion parameter. Therefore, it must take on values between 0 and 1. The population-level mean for the shift proportion parameter is assigned the following uniform prior, $\mu_\delta \sim Uniform(0, 1)$. The only other point to note is that all parameters take on standardized values (i.e., the sensitivity and criterion parameters). The tone evidence values, x_n, were thus standardised according to the distributional properties for high and low tones. All computer code for these models can be found in Appendix A.

3.2. Tone Discrimination Analysis

The original study covered in Chapter 2 presented high and low tones to observers with equal probability. The distributions from which the tones were sampled spanned a total of 11 tones, with each distribution consisting of 200 samples. That is, the experiment presented 200 high and 200 low tones to each observer (N = 400). The methodology for the current experiment does not differ from that of the first, though the number of stimuli were doubled. The reason behind this was to increase the overall difficulty of the task, and to encourage observers to place more reliance upon the TTKR. However, the increase in number of tones means that the number of JNDs separating adjacent tones decreases. It is important that all tones are sufficiently discriminable from each other; to ensure the assumption of sensory control was still met, a small study was undertaken in which the pair-wise discriminability of adjacent tones was investigated. A 2AFC task was devised in which all observers were required to identify whether the tone in the second interval was higher or lower than the tone in the first interval.

3.2.1. Observers

A total of 5 observers (all males) completed the tonal discrimination task. All observers were recruited from Massey University and had normal hearing.

3.2.2. Stimuli

The stimuli were a series of 34 pure tones generated in MATLAB with a Hanning filter applied (see Appendix B for stimuli list). Tones ranged from 445.4 Hz to 1145.0 Hz, separated by 21.2 Hz steps. The tones were exported as 16 bit wave files for use (sample rate = 44100 Hz) and played at a sound level set to be comfortable for the observer through Panasonic RP-HT161E-K stereo headphones. Tones were separated by 7 to 17 JNDs (Shower & Biddulph, 1931; see also Appendix C). All adjacent tones were played as ascending and descending pairs (e.g., 1-2, 2-1, 2-3, 3-2, 3-4, 4-3, and so on) which resulted in 66 tone pairs.

3.2.3. Procedure

Observers used a response box to convey their decision as to whether the second tone was higher or lower than the first. Two buttons were provided, one each for higher and lower responses. Within each interval the tone was presented for 200 ms, which included rise/fall time. Observer responding was self-paced. Once a decision had been made the next pair of tones was presented. Tone pairs were randomly ordered for all combinations.

3.2.4. Analysis

The percentage of correct decisions was the primary dependent variable of interest. In order to satisfy the sensory control assumption, a criterion of 95% accuracy was adopted. In accepting that Weber's fraction was diminishing toward the higher tone range (see Appendix C), some focus was placed upon the percentage correct for these pairs.

3.2.5. Results

The mean accuracy across observers was 97.4%, indicating the tones were sufficiently discriminable. Only one observer failed to achieve perfect accuracy and performed well below the criterion mark (performance was at 87%). The distribution of errors for this observer were spread across the range of the pairs and not solely confined to the high tonal region. This might suggest that lapses in concentration contributed toward the errors rather than any difficultly in discriminating between the tones, or that the observer

did not attempt the task correctly. All other observers seemed to have no trouble in completing the task and so the tones were deemed to be sufficiently discriminable.

3.3. Method

3.3.1. Observers

A total of 38 observers participated in Experiment 2. The sample consisted of 15 women (39.5%) and 23 men (60.5%) with an age range of 18 to 34 years (*M = 24.5; SD = 3.8*). Observers were Massey University undergraduates who were reimbursed for their time.

3.3.2. Stimuli

The experimental stimuli were 34 discriminable tones that ranged from 445.4 Hz to 1145.0 Hz, separated by 33 equal steps of 21.2 Hz. Stimulus distributions were created where the frequency of each tone was calculated so as to best approximate a normal distribution with unit variance. Tones from each stimulus class were presented with equal probability. Each distribution, then, were built from a total of n = 200 samples spread across a range of 22 tonal magnitudes (i.e., N = 400). The stimuli were divided into three sample groups: low (L); high hard (HH), and high easy (HE), where the means for each distribution were M = 11.5, 15.5, and 23.5, respectively (see also Appendix B). All distributions had a standard deviation of approximately 4 tones, which corresponds to $d' \approx 1$ and $d' \approx 3$ for the hard and easy conditions, respectively (see Table 3.2). The optimal criteria for this experiment, for each level of difficulty, were located at $c_o = 13.5$ (the midpoint between the L and HH distributions) and $c_o = 17.5$ (the midpoint between the L and HE distributions).

Table 3.2:
Distributional Properties of the Experimental Stimuli.

Distribution	N	Range (Hz)	Lower	Upper	Mean	SD	d'_{actual}
Low	200	445.4 - 890.6	1	22	11.5	4	0
High-Hard	200	530.2 - 975.4	5	26	15.5	4	1
High-Easy	200	699.8 - 1145.0	13	34	23.5	4	3

Tones were presented binaurally using Panasonic RP-HT161E-K stereo headphones with a sound level set to be comfortable for the observer, and were greater than 7 JNDs but less than 17 JNDs apart (50% level; Shower & Biddulph, 1931). Stimulus sequences were then determined by sampling from each distribution without replacement.

3.3.3. Procedure

All observers were told that they would hear a range of tones and that their task was to categorise the tones as either high or low. They were also told that high and low tones would occur with a 50% probability. Additionally, observers were informed that they would receive feedback on each trial, and that they should use the feedback as best as possible in order to maximise their performance. Prior to training the full range of tones was played, after which all observers completed a block of 50 training trials (with no feedback). Once the training trials had been completed, observers then commenced with 400 main experimental trials (a break was given after 200 trials). Each trial had a maximum length of 3500 ms and consisted of the following sequence: tone presentation (200 ms including rise-fall time), response interval (2000 ms), feedback (300 ms), and inter-trial interval (1000 ms). All stimuli were presented via a custom program that utilised the system's default media player, and tones were presented with no added background noise. After each tone presentation observers were presented with a response box that was located in the centre of the screen. The box contained two buttons, one for high responses, and one for low responses. Observers used the computer mouse to make a response. After each response, feedback was provided immediately. Feedback was a green light that flashed above the button that should have been selected. That is, if the feedback light flashed green above the chosen button, this indicated a correct decision; if it flashed green above the alternate button then the decision was incorrect. Participants also completed an additional session where no feedback was provided during the main trials. During the no feedback sessions no light appeared above any of the response buttons after a response was lodged. The order in which observers completed the KR and no KR sessions were randomised and each session was separated by a minimum of

one hour. Sessions lasted between 40-60 minutes and observers were reimbursed for their time.

3.3.4. Analysis

The analytical approach taken here is identical to that taken in Chapter 2. SDT summary statistics were obtained and a psychometric function was fit to each observer's data using the same hierarchical Bayesian model specified in Section 2.2.4. All prior distributions for the psychometric function remained unchanged, as did the number of chains, samples, and burn-in. With regard to the dynamic criterion model comparisons, apart from the changes to the prior distribution outlined above, the number of chains, samples, and burn-in also remain unchanged. Model comparisons proceeded in the same way as Chapter 2 (using Bayes Factors) where the marginal likelihoods for all models followed the process described in section 2.2.6.

3.4. Results

First, signal detection parameters were estimate by fitting the static criterion model across each group. The SDT parameters are displayed in Table 3.3. The first point to note is that there is a difference in sensitivity between the $TTKR_e$ ($d' = .66$) and the $TTKR_i$ ($d' = .95$) groups in the hard condition. This effect was absent in the Taylor's (2010) original study, and it is likely that the increase in overall difficulty (through the increase in stimuli numbers) forced observers to rely more upon the TTKR. The reliability of this effect was

Table 3.3:
MAP Estimates for Signal Detection Indices.

Group	c	2.5%	97.5%	d'	2.5%	97.5%
1	.30	.25	.36	.66	.58	.74
2	.51	.45	.57	.95	.86	1.03
3	1.12	1.05	1.20	2.41	2.30	2.52
4	1.18	1.11	1.25	2.43	2.32	2.53

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

Table 3.4:

Psychometric Function MAP Estimates for Population Level Posterior Distributions.

Group	α	2.5%	97.5%	β	2.5%	97.5%	μ	σ
1	-5.95	-7.71	-4.22	.45	.32	.58	13.26	2.23
2	-11.30	-12.02	-10.64	.83	.78	.88	13.64	1.21
3	-9.16	-9.82	-8.53	.54	.50	.58	17.06	1.86
4	-9.34	-9.98	-8.69	.54	.50	.58	17.33	1.86

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

assessed by performing the Savage-Dickey density ratio test on the posterior distributions for the population level d' values (this method was described in Chapter 2[12]). The difference between Groups 1 and 2 was 4.33 x 10^4 times more likely under the alternative hypothesis then the null. Conversely, the differences in d' between Groups 3 and 4 were 15 times more likely under the null. These results agree with the expectations that TTKR type should differentially affect performance in the hard condition and have no differential affect in the easy condition. The changes in sensitivity also corroborate with the estimated slopes of the psychometric function (Table 3.4). There is a clear difference in the slopes between the $TTKR_e$ ($\beta = .45$) and the $TTKR_i$ ($\beta = .83$) groups in the hard condition. These differences in slopes further translate to larger estimates of criterion variance for the $TTKR_e$ group ($\sigma = 2.23$) compared with the $TTKR_i$ group ($\sigma = 1.21$). Note that one standard deviation is now approximately 4 tones and not 2 as per the previous study.

The posterior distributions (again transformed so that they are expressed on the stimulus scale) are displayed as violin plots in Figure 3.1. The most notable feature is the $TTKR_i$ posterior in the hard condition. There is very little variability in the posterior, implying that observers in the group were quite consistent. It is fairly impressive that observers in this group were able to be so consistent, and serves to further highlight the amount of noise in the posterior for the $TTKR_e$ group.

[12] For all Savage-Dickey tests a zero-centered Cauchy distribution is used as the default prior for the null hypothesis.

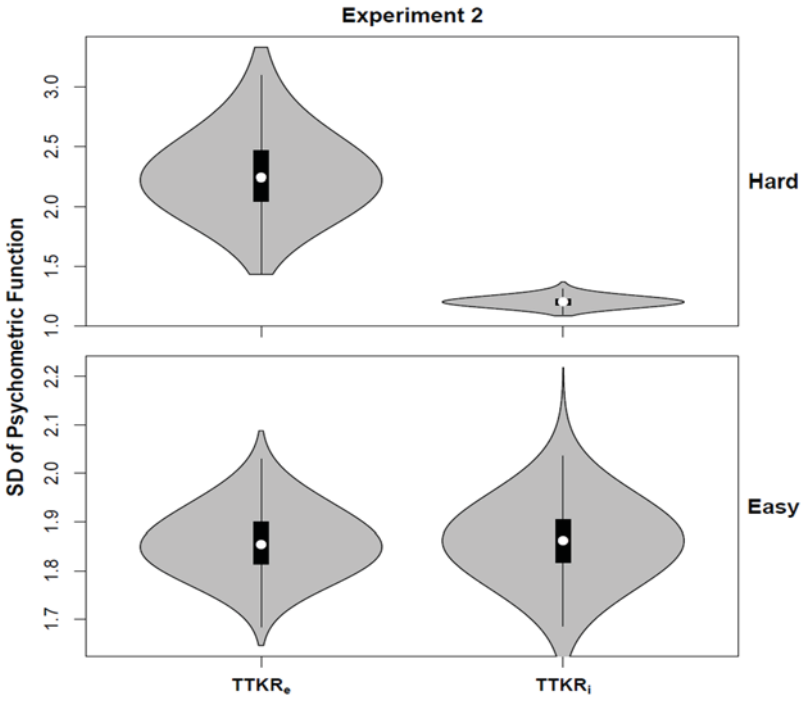

Figure 3.1: Posterior distributions for the population level slope parameter, transformed so that the values are expressed on the stimulus scale, σ.

As expected, the posteriors for the TTKR groups in the easy condition are very similar and appear to be unaffected by KR type. On the whole the results appear to show that the type of TTKR had a substantial effect on response consistency.

3.5. Model Comparisons

The parameter estimates for each model are displayed in Table 3.5, where the population level means and 95% Bayesian Credible Intervals are also reported. The first point of interest is the shift parameters for the ECM. While the values for both the $TTKR_e$ conditions (Groups 1 and 3) are both in the expected positive direction, the $TTKR_i$ estimates (Groups 2 and 4) are in the contrary direction (see also Figure 3.2). That is, assuming the parameters to be accurate, observers in the $TTKR_i$ groups are shifting their criterion in the direction that makes an error *more* likely. This does not seem like a sensible prediction. The shift magnitudes, however, do agree in terms of the absolute values of the shifts. The size of

Table 3.5:

MAP Estimates for Population Level Posterior Distributions across Dynamic Criterion Models.

Model	Group	c_0	2.5%	97.5%	d'	2.5%	97.5%	δ	2.5%	97.5%
ECM	1	.25	.16	.33	.67	.59	.76	.019	.003	.043
	2	.51	.41	.57	.95	.86	1.03	-.008	-.018	.004
	3	1.11	1.02	1.20	2.42	2.31	2.53	.006	-.010	.023
	4	1.17	1.09	1.26	2.43	2.32	2.54	-.003	-.017	.015
PECM	1	.19	.07	.28	.68	.58	.75	.011	.004	.018
	2	.43	.28	.58	.93	.85	1.01	.090	.018	.104
	3	1.10	.99	1.19	2.42	2.29	2.53	.005	.000	.019
	4	1.14	1.05	1.23	2.42	2.30	2.54	.007	.000	.018

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

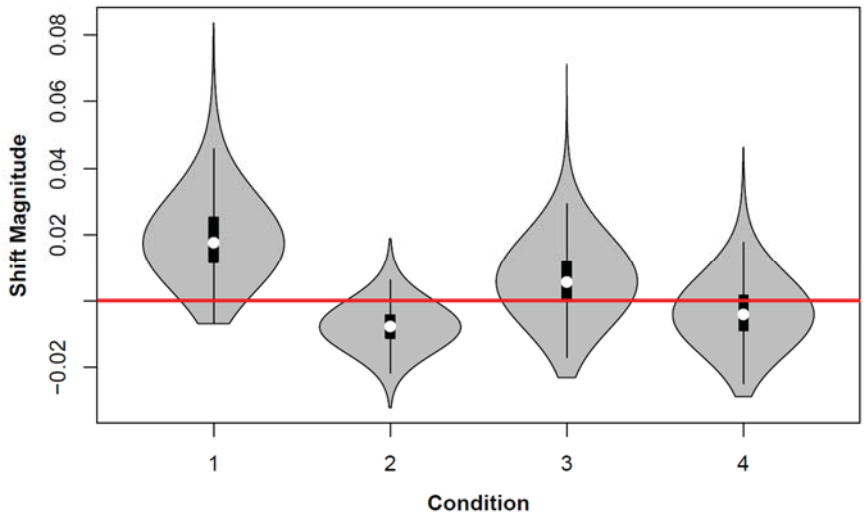

Figure 3.2: ECM population-level posterior distributions for the shift parameter, δ. Conditions: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

the shift was greater for the $TTKR_e$ group (δ = .019) compared to the $TTKR_i$ group ($|\delta|$ = .008) in the hard conditions.Interestingly, the effect also appears to be present in the easy conditions, though to a lesser extent. In assessing the reliability of the differences between the shift parameter posteriors the absolute values of the $TTKR_i$ distributions were compared with the $TTKR_e$ distributions across both levels of difficulty. Much like the results reported

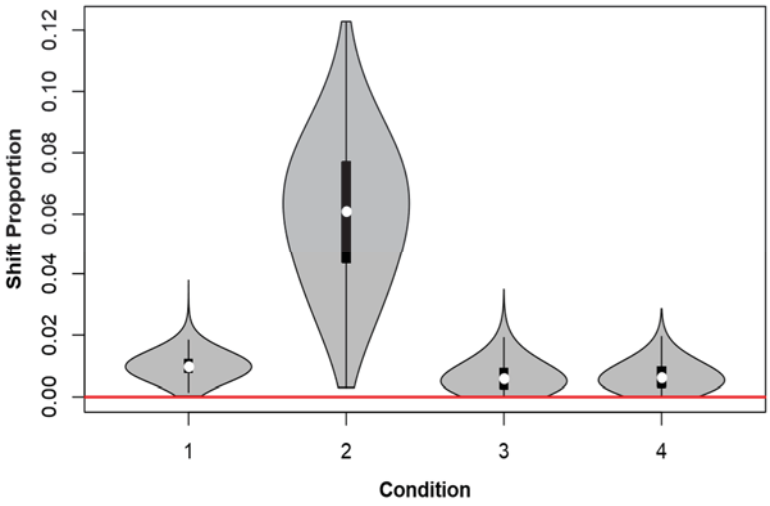

Figure 3.3: PECM population-level posterior distributions for the shift parameter, δ. Conditions: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

in Chapter 2, the differences were 76 and 120 times more likely under the null for the hard and easy condition, respectively. If we consider the individual group distributions we again see that they do not reliably depart from zero and are more likely under the null: $BF_{01} = 5$, $BF_{01} = 61$, $BF_{01} = 122$, and $BF_{01} = 102$ for Groups 1 to 4, respectively.

Turning next to the PECM, there appears to be an issue where the MCMC chains have not appropriately mixed for the Group 2 data which has resulted in quite a noisy posterior distribution for the shift proportion parameter (Figure 3.3). Moreover, the differences between the Group 1 and 2 estimates are in the wrong direction. The difference between the Group 1 and 2 posterior distributions was approximately 3 times more likely under the null, though the differences may be entirely driven by the lack of chain convergence. The actual PECM parameter estimates themselves are also inconsistent with the slope estimates. Seemingly, the proportion of the difference used in shifting the criterion was larger for the $TTKR_i$ group (Group 2; δ = .09) than the $TTKR_i$ group (Group 1; δ = .011). While the ostensible lack of difference between the posterior distributions for Groups 3 and 4 was expected (the difference is 195 times more likely under the null), the shift proportions do not

reliably differ from zero for either Group 3 ($BF_{01} = 51$) or Group 4 ($BF_{01} = 31$) and so virtually model the criterion as fixed. Based upon the parameter estimates alone, neither model appears to provide interpretable, or useful, estimates. All shift estimates are negligibly different from zero and make predictions that are very similar to the static criterion model. This occurs despite both models possessing an additional parameter. This point is further driven home when we acknowledge that the estimates of sensitivity displayed in Table 3.5 do not differ from those displayed in Table 3.3. That is, neither model is capturing any features of the data that the SCM cannot already account for. Both models fail to capture any variability that would lead to improvements in sensitivity estimates.

To firmly establish that the ECM and PECM models add little or nothing to the SCM predictions, Bayesian model comparisons between each of the dynamic criterion models and the standard signal detection model were conducted. In Chapter 2, the model predictions for each and every observer were compared. In this Chapter, however, we will treat the Bayes Factor as a summary statistic in its own right, the details of which are displayed in Table 3.6 for each model. The logarithms of the individual Bayes factors are also plotted in Figures 3.4 and 3.5 for the ECM and PECM model comparisons, respectively. In treating the Bayes factors as a summary statistic we may obtain a sense of how variable the evidence in favour of either model is. When we consider the mean Bayes factors across all observers we see that, on average, the data were 39.6 times and 29.3 times more likely under the SCM when compared to the ECM and PECM, respectively. Note also that the lower average value for the PECM does not imply that it fit the data better. Both the range and standard error for the PECM Bayes

Table 3.6:
A Summary of the Dynamic Criterion Models and Bayes Factors across all Participants.

	Parameters			BF_{12}		
Model	No.	Fixed	Criterion Shift	Range	Mean	SE
ECM	3	NA	$c_n + \Delta\delta$	9 – 87	39.6	3.44
PECM	3	NA	$c_n + \Delta\delta(x_n - c_n)$	.6 - 197	29.3	7.50

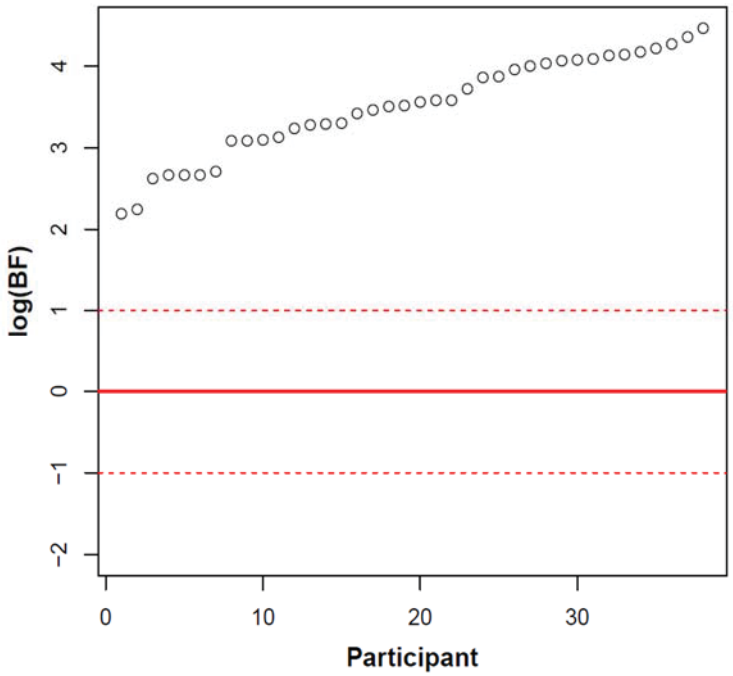

Figure 3.4: Individual level log Bayes Factors as evidence for the ECM across all experimental conditions. The values are ordered without regard for participant group because in all cases the evidence strongly favours the SCM. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

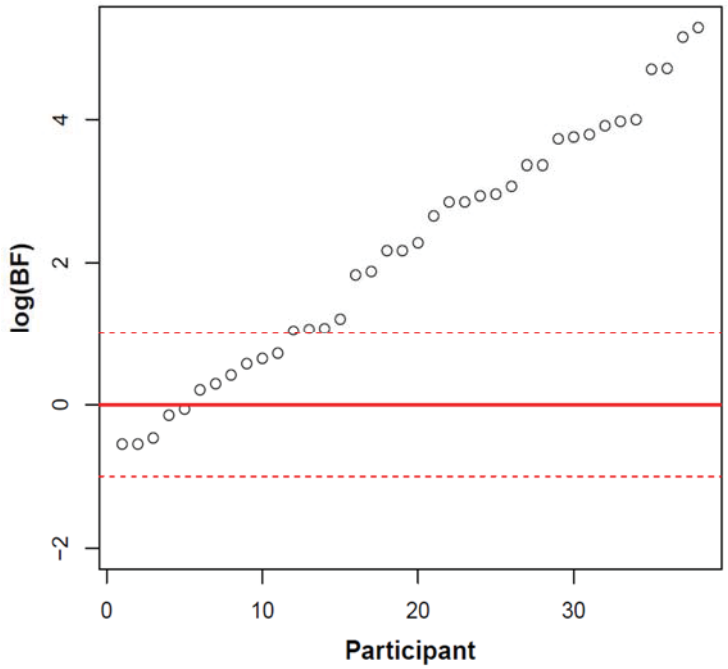

Figure 3.5: Individual level log Bayes Factors as evidence for the PECM across all experimental conditions. The values are ordered without regard for participant group because in all cases the evidence strongly favours the SCM. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

factors was larger when compared to the ECM Bayes factors. This information may also be gleaned by assessing the relative slopes of the plotted values. One can appreciate that the values for the PECM cover a wider range of values when compared to the ECM. What these data serve to illustrate is that there is very little consistency between the predictions made by both dynamic criterion models and the empirical data. Furthermore, these findings replicate the results reported in Chapter 2 despite the prior distributions for the dynamic criterion models being slightly more informative.

3.6. Evaluating the Error-Correction Models

In summarising the foregoing, it is clearly evident that both the error-correction models considered here and in Chapter 2 are unable to make useful connections with the empirical data. Across all model comparisons the SCM was unanimously favoured. The current working hypothesis has been that shifts in the criterion result from an error-correction process, though this need not necessarily be the case, and one might wonder at this point whether the error-correction models are too restrictive. These are by no means novel considerations; Chapter 1 discussed generalisations that were made to Kac's error-correction model that allowed shifts to occur following correct decisions (Dorfman & Biderman, 1971). Similar generalisations may also be afforded to models where the stimulus value contributes toward the criterion shifts on each trial (this is to be discussed in the next section). Specifically, stimulus values are only assumed to have an effect upon the criterion following errors; on all other trials the stimulus is assumed to have no effect upon the decision criterion.

It is unclear how one would apply error-correction models in situations where feedback is not provided, or is unavailable. There are many scenarios, both in the lab and in general, where decisions must be made in the absence of any feedback. However, the absence of KR means we cannot specify a shift matrix, which raises difficulties in prescribing ways in which the criterion should shift, if indeed it does. One solution might be to use the observer's response as a proxy for feedback, thereby modeling the shifts using a stimulus-by-response

matrix; but this is no different to how the shifts are modeled in the $TTKR_e$ case. That is, what is being modeled is not true of the conditions under which the data were generated. A response-by-response matrix cannot be used either, because observers would always be correct, meaning the criterion would never shift. The problem is that observers have no firm idea of the event that has occurred on each trial, which makes intuiting just how the criterion should shift based upon their responses very difficult. Moreover, because the observer has no knowledge of the correctness of their decision it makes the idea of error-correction rather presumptuous. In this case their response *is* the only information about the event. The error-correction model, from a process point of view, is therefore flawed for such cases.

There are also practical limitations that prohibit prior predictive distributions being generated for error-correction models. Prior predictive distributions are obtained in a fashion not too dissimilar to how marginal likelihoods are obtained. It requires sampling from prior distributions and plugging the sampled values into the model equations, though rather than evaluating the likelihood of the data with the model predictions, interest solely focusses upon the model predictions themselves. In other words, prior predictive distributions provide a sense of the type of predictions a model makes prior to any actual data being observed. The Bayes factor, then, simply summarises the consistency between the model predictions and the observed data. However, a difficulty arises insofar as the error-correction model requires knowledge of the data (observer responses on each trial) in order to determine the next criterion value. To generate prior predictives we would then have to consider, via simulation, every possible response sequence (which would include sequences that are unlikely to be observed), given the current stimulus sequence, to derive the predictions. Such an approach would be prohibitive from a computational standpoint. One could, in practice, generate prior predictions for the error-correction model that *are* based upon the observer's response sequence, and in fact this is essentially what is done when the model is fit to the data. However, what one has now is a subset of the possible model predictions that are conditional upon an already observed response sequence, and one could certainly argue that these are not prior predictions in the conventional sense.

Finally, with respect to the inclusion of the stimulus value in the model specifications given the results thus far, it is still unclear whether the addition of stimulus information actually improves the fit of the dynamic criterion model. With the shifts confined to only errors it would seem that no benefit is gained. In light of the foregoing discussion, then, it would be prudent to consider a generalised class of dynamic criterion model where shifts can occur across all experimental trials; that is, the criterion can shift following correct decisions, too. In practice, for the PECM this requires minimal change to the model specifications detailed previously: just remove the shift matrix. Recall that it is the shift matrix that constrains the model to shift only following errors. For the ECM, however, allowing for shifts following correct decisions requires an adjustment to the shift matrix and the inclusion of an additional parameter. The details of these models and subsequent comparisons are considered next.

3.7. Generalising Criterion Shifts

Generalised models of the type described by Dorfman and Biderman (1971) and their inherent limitations were discussed in Chapter 1. However, it is worth reiterating that these models permit shifts on trials where correct decisions are made. It is assumed that the criterion, following correct decisions, shifts away from the sampled value (or internal stimulus effect) which serves to increase the area to the left or right of the criterion, depending upon which side the stimulus has fallen. The motivation behind this is to increase the probability of repeating the previously correct category/response, or what is referred to as assimilation (cf. Treisman &Williams, 1984). To model these shifts, however, requires two modifications to the ECM previously discussed. The first requires the shift matrix to be altered so that shifts can occur following correct decisions, and is done in the following way

$$\Delta = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}. \qquad \text{Eq. 3.3}$$

Note that the criterion is always shifted downward following signal (or high) trials, and always shifted upward following noise (or low) trials. The second is to include a shift

parameter for correct decisions, $\delta_{correct}$. The original shift parameter thus becomes δ_{error}. It is assumed that, like δ_{error}, values of $\delta_{correct}$ are drawn from a population-level distribution that is normally distributed, where the priors for the population mean are drawn from a uniform prior between -.5 and 0.5. That is, the model specifications are identical to the ECM; all we have done is to include an additional shift parameter. This model will be referred to as the General Deterministic Model (GDM; Dusoir, 1980).

Thomas (1975) and Dusoir (1980) have discussed dynamic criterion models where feedback is absent and reason that the criterion shifts as a function of the difference between the stimulus magnitude and the current criterion location. The criterion shifts in the same way as the PECM discussed in the previous section with the only modification being that the shift matrix is removed. This, then, allows the criterion to shift across all trials and is not dependent upon stimulus or feedback shift matrices. This is a particularly convenient feature of PECM, and one that allows the model to be generalized without including any extra parameters, and there is no reason why the model cannot be fit to the TTKR response data. We can now write the model equation as $c_{n+1} = c_n + \delta(x_n - c_n)$. This model will be referred to as the pure GDM (PGDM). The effects of TTKR should then be manifest in the relative weighting upon the difference between the criterion and the stimulus. Specifically, placing less weight upon the second right-hand term will result in smaller criterion shifts, and is something we might expect to see where $TTKR_i$ has been provided. On the other hand, if more weight is placed upon the term, then larger shifts in the criterion will be observed - something we might expect to see where $TTKR_e$ is provided.

The PGDM is a form of weighted average which is updated across trials in response to incoming stimulus information. However, a critical difference between a weighted average criterion and the deterministically assigned GDM criterion is that the weighted average will always shift *toward* the stimulus value. Consider for example a categorisation task, where on trial n the observer is presented with some amount of information, x_n, and responds signal. For a signal response to have been made it is necessarily the case that $x_n > c_n$. Now, if x_n happened to be a noise event, we assume that some proportion of the difference, $x_n - c_n$, is

used in updating the criterion where, because the difference will be positive for all $x_n > c_n$, the criterion will shift toward x_n. Note that this is precisely what occurs for error-correction. Conversely, if x_n was a signal event then the observer would still shift the criterion towardsx_n, and by doing so exhibit contrast behaviour (rather than the typical assimilation expected after a correct response; cf. Treisman &Williams, 1984). The key point to note is that if the criterion is modeled as a weighted average, then error-correction is an emergent property. It does not need explicit modeling. Moreover, weighted average models allow the criterion to shift across all trials without having to include additional parameters. That is, the generalisation can be achieved at no expense. The GDM, on the other hand, requires an additional parameter.

A further point to note is that a number of issues outlined in the previous section can be addressed if we model the criterion as a weighted average. Aside from the models making fuller use of the stimulus information across all trials, the weighted average models do not require the response sequence to be known. The only knowledge that is required is the stimulus sequence and the stimulus values on each trial, which are determined by the experimenter. What this means is that the model can be applied to cases where feedback is provided, or is not provided, because the shifts in the criterion are driven by the stimulus information and not a stimulus-by-response shift matrix.

An additional benefit that arises is that we could generate prior predictive distributions that are not conditional on data. When the criterion is modeled as a moving average the only thing we require in advance is the ordering of the stimulus and their respective values. In summarising, there is reason to suggest that the error-correction models considered up to this point have been too restrictive. We can generalise both the ECM and PECM by either including an additional shift parameter for correct trials (the GDM) or by removing the shift matrix altogether (the PGDM). In the next section, then, we will fit these two generalised dynamic criterion models. Apart from the aforementioned changes, all model specifications, number of chains, and burn-in are as previously discussed. Code for the generalised shift models can be found in Appendix A.

Table 3.7:

MAP Estimates for the GDM.

Cond	d'	2.5%	97.5%	δ_{error}	2.5%	97.5%	$\delta_{correct}$	2.5%	97.5%
1	.65	.57	.74	.007	-.021	.036	-.017	-.032	-.001
2	.95	.86	1.03	-.018	-.029	-.006	-.021	-.03	-.011
3	2.41	2.29	2.52	-.017	-.048	.013	-.022	-.044	-.001
4	2.43	2.31	2.55	-.02	-.04	.000	-.018	-.033	-.004

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

3.7.1. Conditions with TTKR

We first consider the GDM model. The parameter estimates are displayed in Table 3.7 and the posterior distributions for the shift parameter across correct and incorrect decisions are displayed in Figure 3.6. There is little need to consider in depth the results from the GDM fits. First, the sensitivity parameters are identical to the estimates obtained by the error-correction model fits and the static criterion model. That is, the GDM is making similar predictions to the SCM. The error shift parameters, with the exception of Group 1, are negative which again suggests that the criterion is shifting away from the stimulus value, which makes the probability of an error more likely. The negative shift parameters for the correct shifts are more readily interpretable. Rather than shifting the criterion away from the stimulus value these parameter suggest that the criterion is shifting toward the stimulus value, much like a weighted average, though generally these difference do not reliably depart from zero (see Figure 3.6). For the shifts following incorrect decisions, the mean posterior values across Groups 1-4 were 71, 3, 51, and 22 times more likely under the null. Following correct decisions the mean posterior values were 15, 14, and 7 times more likely under the null for Groups 1, 2, and 4, whereas the difference for Group 2 was around 11 times more likely under the alternative. Lastly, from the Bayes factors we see that the observed data are more likely under the static criterion model than under the GDM (Figure 3.7). The Bayes Factors ranged from 9 – 41, with a mean Bayes factor of 20 (SE = 1.5). All told, the results for the GDM are not that dissimilar from the results for the error-correction model fits.

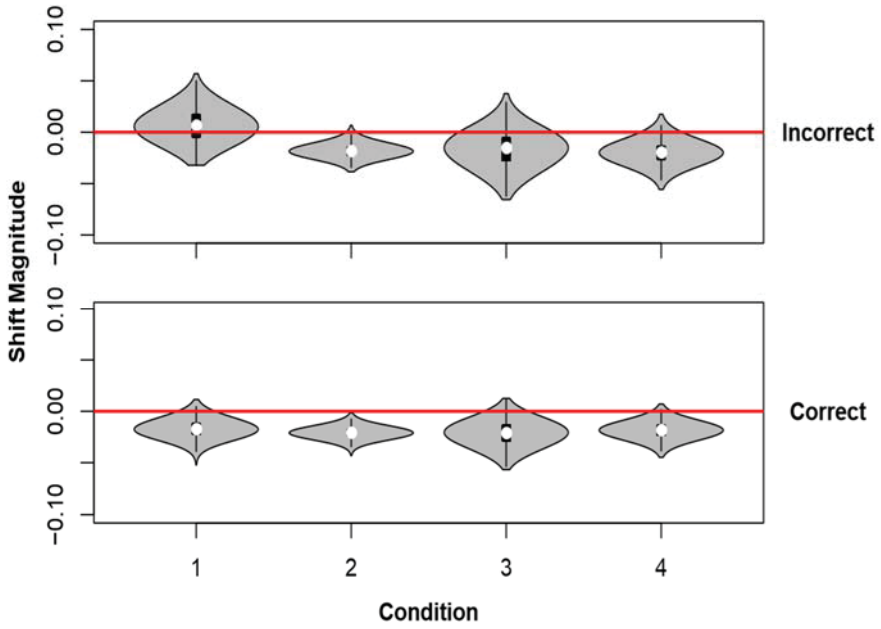

Figure 3.6: GDM population-level posterior distributions for the shift parameter following incorrect and correct decision across all experimental conditions.

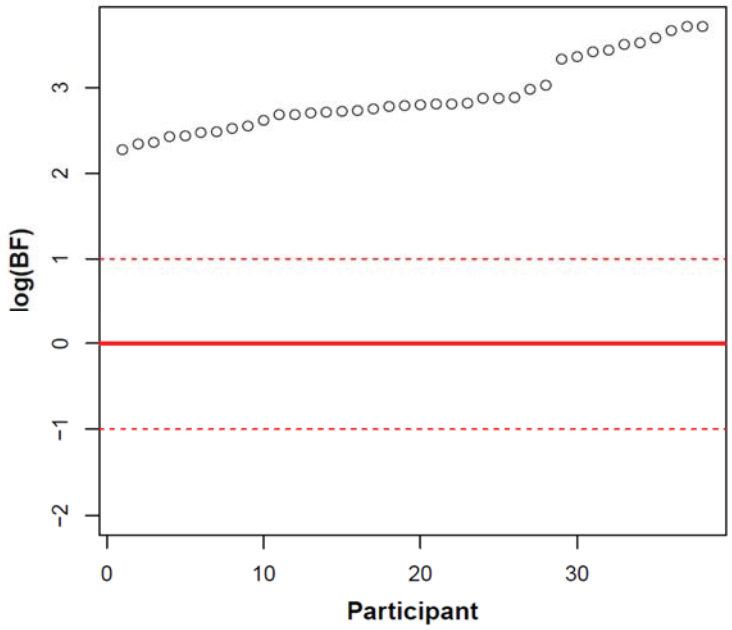

Figure 3.7: Individual level log Bayes Factors as evidence for the GDM across all experimental conditions. The values are ordered without regard for participant group because in all cases the evidence strongly favours the SCM. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

Table 3.8:

MAP Estimates for the PGDM.

Group	c_0	2.5%	97.5%	d'	2.5%	97.5%	δ	2.5%	97.5%
1	.14	.01	.40	.85	.78	.91	.239	.160	.322
2	.29	.02	.62	.94	.88	1.01	.177	.104	.253
3	1.27	1.09	1.42	2.70	2.56	2.81	.061	.006	.136
4	1.16	1.00	1.31	2.57	2.47	2.67	.042	.002	.105

Note: Group refers to TTKR/Difficulty combination: 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = $TTKR_e$/Easy; 4 = $TTKR_i$/Easy.

Turning now to the PGDM model fits, there are two encouraging points to note. The first is that the population-level means for the shift parameters appear to make some sense (Table 3.8; see also Figure 3.8). We see a larger proportion of the difference used by the Group 1 observers (δ = .24) compared to the Group 2 observers (δ = .18). This implies that the criterion is being shifted by a larger degree when observers receive $TTKR_e$ in the difficult condition. The difference between the posterior distributions, however, was 12 times more likely under the null.

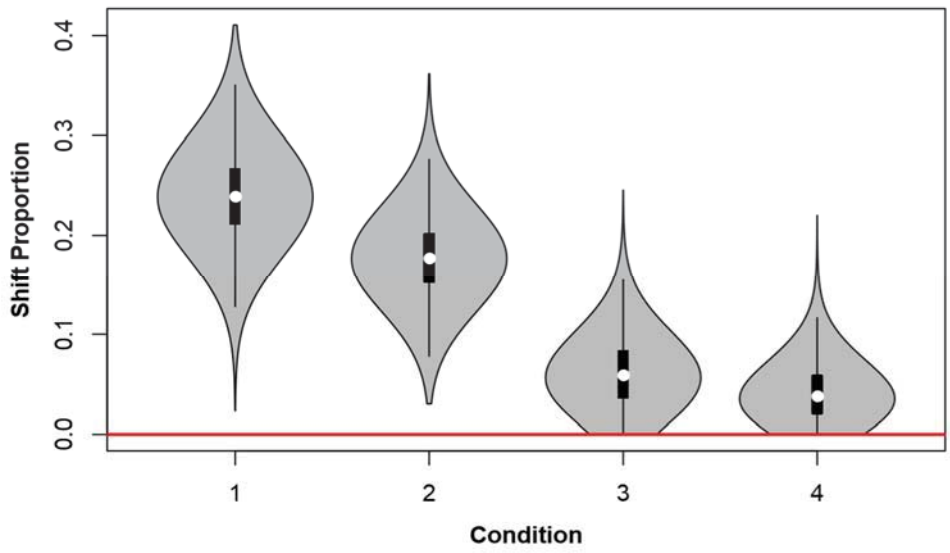

Figure 3.8: PGDM population-level posterior distributions for the shift proportion parameters across all groups.

In the easy condition much smaller proportion parameters occur, where they are fairly similar in magnitude across both the groups. The TTKR is not expected to have any differential effect upon the criterion across Groups 3 and 4, and this appears to the case. Additionally, the difference between the posterior distributions for Groups 3 and 4 were 26 times more likely under the null. The second point is that for the first time, with the exception of Group 2, we see an improvement in the estimates of sensitivity relative to the static criterion estimates. This implies that the model is accounting for some of the variance present within the data which are being captured by the shift parameters. Modeling the criterion as a weighted average appears to be more consistent with the empirical data; however, we now need to assess whether the PGDM is more consistent relative to the static criterion model. We return to the Bayes factors to answer this question (Figure 3.9).

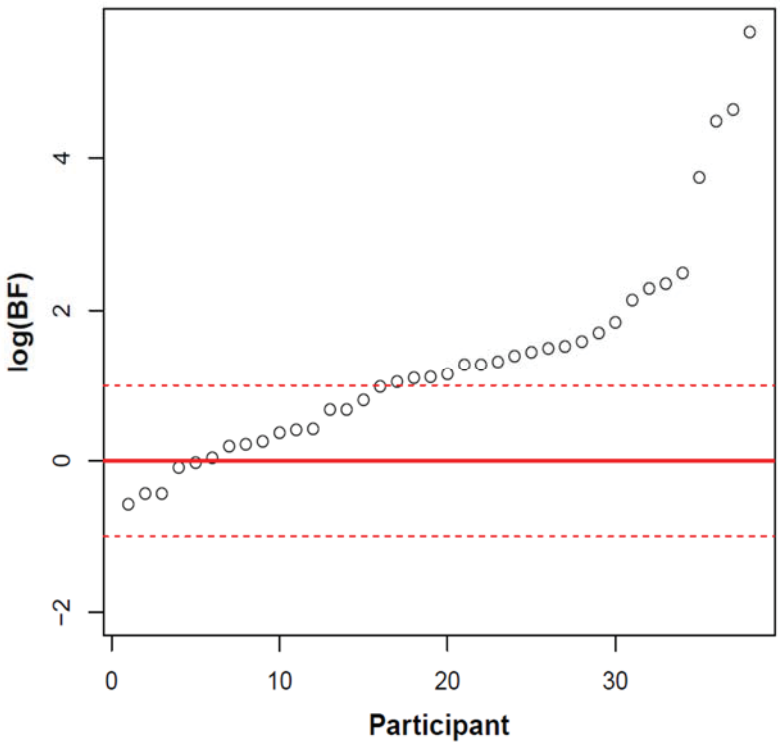

Figure 3.9: Individual level log Bayes Factors as evidence for the PGDM across all experimental conditions. The values are ordered without regard for participant group as for the majority of the cases the evidence favours the SCM. A handful of observers from Group 1 produced data that were more consistent with the PGDM, though the evidence is very weak. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

The range of the Bayes factors extended from .56 – 290, with a mean Bayes factor of 17 (SE = 8.2). On average, most observer data were still more likely under the SCM. There were a couple of observers whose Bayes factors were very large (BF = 290 and BF = 104; both observers were from Group 3) which probably upwardly biased the mean. If we consider the mean with these observers removed the Bayes factor is 7, and so ignoring these outliers does nothing to change the interpretability of the results. For the five observers whose data were more consistent with the PGDM (all of whom came from Group 1) we see the evidence is very weak. Though not apparent in Figure 3.8, approximately 40% of participants fell within the area where the evidence is indifferent toward either model. That is, for these observers their data are consistent with either model and there is no way to tell which model is better.

It should be noted that of this 40%, all but two observers were in either Group 1 or Group 2. That is, observers from the easy condition were generally more consistent with the SCM. So, despite the PGDM accounting for some of the variance in the data, we have very little evidence to suggest that the PGDM is more consistent with the observer data than the SCM. This conclusion will be further considered in Section 3.8.

3.7.2. Conditions with No TTKR

In addition to the TTKR sessions each observer participated in a full session of no TTKR trials (see Methods). First, signal detection indices were estimated by fitting the SCM model to the no TTKR data. The no TTKR observers in the difficult condition performed better (d' = .87; Bayesian 95% Credible Interval = .81 - .94) than the $TTKR_e$ observers (cf. Table 3.3). This finding agrees with the findings reported by Lee and Zentall (1966) and the predictions made by Schoeffler (1965). That is, observers tend to perform better when $TTKR_e$ is absent. However, observers appeared to perform more poorly without the TTKR in the easy condition ((d' = 2.25; Bayesian 95% Credible Interval = 2.17 – 2.35) compared to the TTKR observers (cf. Table 3.3). Such differences have been reported in the literature (e.g., Carterette et al., 1966) though were not found by Lee and Zentall (1966).

Table 3.9:

Psychometric Function MAP Estimates for Population Level Posterior Distributions.

Condition	α	2.5%	97.5%	β	2.5%	97.5%	μ	σ
Hard	-9.61	-10.13	-9.11	0.72	0.69	0.76	13.31	1.39
Easy	-8.49	-9.31	-7.77	0.48	0.44	0.54	17.53	2.07

The psychometric function parameters also corroborate this pattern (see Table 3.9). Considering first the hard condition, the slopes were steeper ($\beta = .72$) when compared to the $TTKR_e$ group ($\beta = .45$), though flatter when compared to the $TTKR_i$ group ($\beta = .83$). This translates to a reduced amount of estimated criterion variability when TTKR was not provided ($\sigma = 1.39$) compared to the $TTKR_e$ group ($\sigma = 2.23$), and greater when compared to the $TTKR_i$ group ($\sigma = 1.21$), though the difference is not overly large. With respect to the easy condition, there was a slight decrease in the slope estimates ($\beta = .48$) compared to the TTKR groups (both $\beta = .54$), which results in an increase in the estimated criterion variance ($\sigma = 2.07$) over the TTKR groups (both $\sigma = 1.86$).

Turing next to the model fits across the no TTKR conditions, from the model parameter estimates (Table 3.10) we see that larger shift proportions are found in the hard condition versus the easy condition. The posterior distributions for the shift proportion parameter (Figure 3.10) are quite sensible and tidy across both conditions. Furthermore, note the estimated sensitivity parameters for the PGDM are larger when compared to the SDT estimates displayed in Table 3.9. This suggests that the PGDM is accounting for some of the variability in the observer data.

Table 3.10:

MAP Estimates for the PGDM without TTKR.

Cond	c_0	2.5%	97.5%	d'	2.5%	97.5%	δ	2.5%	97.5%
Hard	.08	.00	.32	.97	.92	1.03	.24	.18	.30
Easy	.75	.10	1.44	2.46	2.37	2.54	.11	.05	.18

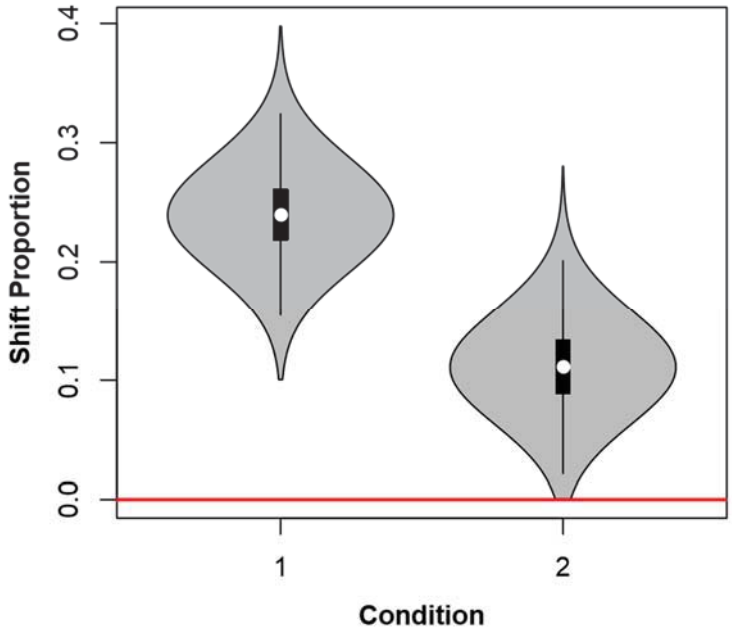

Figure 3.10: Posterior distributions for the population-level shift proportion parameters for the hard (Condition 1), easy (Condition 2), without TTKR.

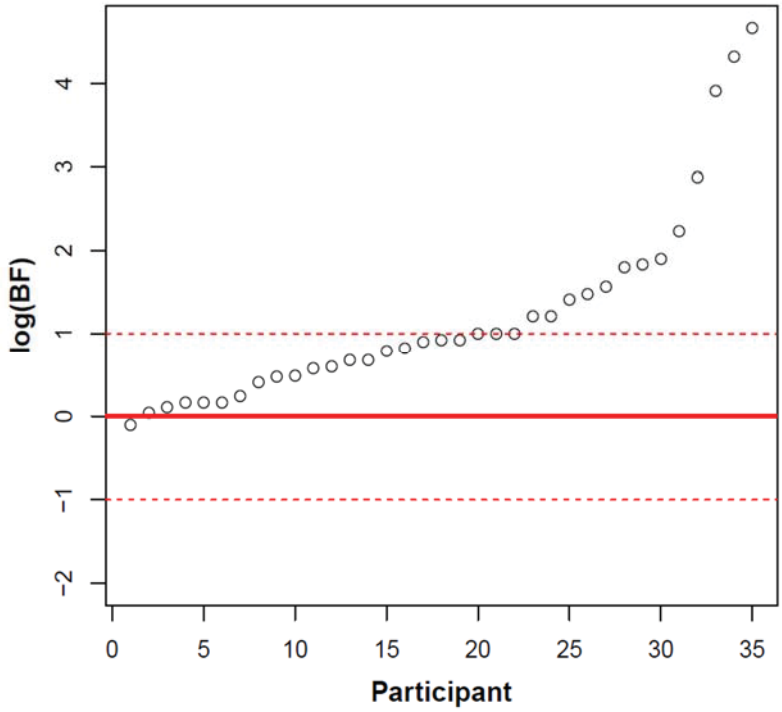

Figure 3.11: Individual level log Bayes Factors as evidence for the PGDM across all experimental conditions. The values are ordered without regard for participant group though for approximately 50% of the cases the evidence favours the SCM. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

However, when we consider the relative consistency between the PGDM and the empirical data, we again see very little evidence for the PGDM (Figure 3.11). The Bayes factors for the no TTKR data ranged from .09 – 107, with a mean of 9.7 (SE = 3.8). From Figure 3.11 we can see that approximately half the observers fall within the region where there is no conclusive evidence in favour of either model. In sum, we have very little evidence to suggest that the PGDM is more consistent with the empirical data than the SCM.

3.8. Summary

In this chapter a series of model comparisons were undertaken which sought to elucidate whether dynamic criterion models would be better able to predict observer data when stimulus magnitude was included within the model framework. Using the probabilistic categorisation task provides a unique opportunity to make inferences regarding how the criterion may shift. This is possible because we make facets of the experimental design explicit (specifically, the stimuli), thus making them amenable to measurement and modeling. We considered first a simple error-correcting version of the models and found that, whether the stimulus information was included or not, the model was not consistent with the empirical data. Through Bayesian model comparisons it was found that, universally, the data were more consistent with the conventional signal detection model. It was then considered that perhaps the error-correction model was too restrictive, and so generalised versions of both the ECM and PECM were fit to the data. The major finding from the generalised model fits was that the PGDM was more consistent with the empirical data than either of the alternative error-correction models. This can be inferred from the average Bayes factors across participants. Moreover, with respect to the general models themselves, it appears that the PGDM, when compared to the GDM, provided a more consistent account of the data. So, we may conclude that the addition of the stimulus magnitude is useful in modeling criterion shifts.

Additionally, there were encouraging findings that suggested that the PGDM has some predictive utility. These generalised dynamic criterion models attempt to capture a facet of

the data that the conventional signal detection model cannot; namely, decision noise. We assume that the estimates we obtain via the standard model are downwardly biased due to decision noise that is unaccounted for. We should, then, expect to see improvements in the sensitivity estimates if the model is capturing some of the decision noise. With this in mind, for all PGDM fits an evident increase in the sensitivity estimates was observed; which in turn suggests the model is accounting for some variability in the data. However, while this may be true, when the comparative fits between the SCM and the PGDM were considered, the evidence for the PGDM was inconclusive for some observers and weak for others. It is unreasonable to expect a model to completely account for all the variability in the data, though while the model as it stands does capture some amount of decision noise, the model comparisons suggest that the model is not specified correctly, or that the priors placed upon the model are still too uninformative.

One benefit of the PGDM is that it can be applied to both the TTKR and no TTKR case without any modification to the model specifications. The PGDM estimates across both the TTKR and no TTKR conditions were sensible and the effects of TTKR were apparent in the shift parameter posteriors. We did see that observers used a greater proportion of the difference between the current criterion and stimulus value in shifting the criterion where $TTKR_e$ was provided, and a smaller proportion where $TTKR_i$ was provided, in the hard condition. Qualitatively, then, the PGDM was able to capture the expected effect of TTKR. Moreover, there was little difference between the shift parameter posteriors between the TTKR groups in the easy condition. With respect to the actual performance across groups, we did see a difference in the sensitivity estimates between the $TTKR_e$ and $TTKR_i$ groups in the hard condition. Interestingly, where no TTKR was provided in the hard condition, performance was better when compared to the $TTKR_e$ group. This result does have empirical precedence and replicates previous findings (Lee & Zentall, 1966).

A non-trivial caveat in the present data is that there are dependencies in the data between the no TTKR and TTKR sessions which make interpreting the results problematic. Observers completed both TTKR and no TTKR sessions and in most instances there was little

time between sessions (a minimum of one hour had to pass between sessions and most observers chose to complete the second session as soon as possible). It is certainly conceivable that there were some effects of practice. Every effort was made to counterbalance the order in which observers completed each session; however, having previous experience with the stimuli may have influenced performance across sessions. This issue will also be considered further in Chapter 5.

Finally, the removal of the shift matrix that generalised the PECM meant that the effects of TTKR were not modeled *per se*; that is, the effects of TTKR were not explicitly modeled but rather inferred from the shift estimates. The only expectation was that the proportionality parameter should be lower in the $TTKR_i$ condition compared to the $TTKR_e$ condition, though this says less about the model specifications and more about the TTKR manipulation. Ideally it would be beneficial to model the effect of the TTKR in a more explicit fashion. That is to say, attempt to gain some insight into the process underlying the use and integration of TTKR across trials, and how this process in turn affects the decision criterion. The next chapter will explore this possibility. One immutable finding from the model comparisons performed presently is that any dynamic criterion model of probabilistic categorization should include information about the stimulus values. The model proposed in the next chapter incorporates this into the model framework.

Chapter 4

The Exponentially-Weighted Learner Model

In Chapter 1 a study by Kubovy and Healy (1977a) was briefly reviewed in which dynamic criterion models (or ALM) were rejected as an appropriate model for observer responding in probabilistic categorisation. Kubovy and Healy wanted to elucidate whether observer responding was better accounted for by a probabilistic process (more like Schoeffler's, 1965, model) or a shifting deterministic criterion (like Kac, 1962), using externally distributed numerical stimuli. To make observer criterion shifts more overt they employed a cut-off report method. The method required observers to first make a judgement based upon whether the numerical value exceeded their criterion value; and second, to explicitly nominate an updated criterion value once a decision had been made. Kubovy and Healy found that while observers were better approximated by an ALM-type model, the ALM was not considered an appropriate model because their observers shifted their criteria in ways inconsistent with the predictions of the ALM. The results from the model comparison in Chapter 3, however, suggest that the dynamic criterion models might better predict the data if the stimulus magnitudes are included in the model and the model allows for shifts following correct decisions. Kubovy and Healy's observers shifted their criterion in either direction following correct decisions, whereas the ALMs predict only unidirectional shifts after correct decisions have been made. None of the models considered in Chapter 3 can facilitate bi-directional shifts.

In an attempt to explain their cut-off report data, Kubovy and Healy (1977a) proposed an Ideal Learner Model (ILM). Their ILM attempted to extend the dynamic criterion process by prescribing a mechanism through which the criterion is able to shift in either direction following correct decisions. One way the ILM departs from the conventional dynamic criterion models is that the stimulus values themselves play a central role in the trial-by-trial updating of the criterion. Furthermore, the ILM begins to address the problem of how observers build and maintain stimulus representations over time (cf. Turner et al., 2011), and how these representations might be used in the establishment and updating of a criterion. Many dynamic criterion models must make assumptions about the way in which the criterion shifts (e.g., error correction), or the criterion sampling distributions (e.g., Muller

& Weidemann, 2008); all of which preclude a role for the stream of incoming sensory information. While it will be shown that the original ILM does not provide a good qualitative description of human performance, much of the ILM framework as it currently stands is immensely useful for the purposes of modeling categorisation behaviour, and so the first half of this chapter is devoted to elucidating the ILM's axioms. The evaluation will also consider various criticisms levelled at the ILM (e.g., Dorfman, 1977).

The second half of the chapter will introduce amendments that can be made to the ILM so that it more accurately reflects human categorisation performance. How can the ILM be modified so that it can accommodate the effects of TTKR described in Chapters 2 and 3? One of the central problems with the ILM is that it supposes that all information is veridically stored for all past events. Such a model lacks psychological validity in light of the well-known limitations in human memory. Therefore, ways in which the "memory" of the model can be altered will be explored. The modified ILM, however, does encounter some theoretical problems that limits is generalizability as a cognitive model, and these will be discussed at a later point.

4.1. Theoretical Evaluation of the ILM

To begin with, an assumption is made that the ILM possesses no specific knowledge about the stimulus distributions; only that each event occurs with equal probability (i.e., $\pi = .5$; pay-offs are also symmetric) and that the distributions have equal variances ($\sigma_0^2 = \sigma_1^2$). The ILM assumes observers attempt to maintain an estimate for each distributional mean, each of which is based upon the trial-by-trial integration of the most recent stimulus event. From a computational standpoint, the ILM is modeled by assigning each observational sample to one of two vectors – one for each stimulus event –and thus represents a store for all past stimulus occurrences. The mean values for each vector on each trial are thus assumed to correspond to estimates of the noise and signal distribution mean, and are continually updated across trials according to the most recently observed sample.

For the present discussion, let us first define the vector for the i^{th} stimulus class as $\mathbf{m}_i$, with the mean on each trial denoted $\bar{\mathbf{m}}_{i,n}$. Further, let k index the k^{th} trial, $k = 1, 2, 3 \ldots, n$, for each stimulus class; and let n denote the current trial. Finally, assume that for each stimulus class the total number of trials is N. Note that this means the maximum length for each $\mathbf{m}_i$ is also N, and the total number of trials is 2 x N. On each trial, the observer is presented with a stimulus that is sampled from one of the stimulus distributions, where $x_{i,n}$ is a scalar – representing evidential magnitude - sampled from the i^{th} stimulus distribution on trial n. On each trial $x_{i,n}$ is indexed according to which distribution it had been sampled from, where

$$\delta_n = \begin{cases} 1, & i = 0 \\ 0, & i = 1 \end{cases}. \qquad \text{Eq. 4.1}$$

That is, $\delta_n = 1 - i$. Let it also be noted that $x_{i,n}$ is always assigned to the vector that corresponds with the distribution it was sampled from: $x_{i,n} \rightarrow \mathbf{m}_i$. That is, all signal events are assigned to the signal vector, $x_{1,n} \rightarrow \mathbf{m}_1$, and all noise events are assigned to the noise vector, $x_{0,n} \rightarrow \mathbf{m}_0$. The estimate of the i^{th} distributional mean on trial n is thus based upon all previous x_i events, and is simply the arithmetic mean:

$$\bar{\mathbf{m}}_{0,n} = \frac{\sum_{k=1}^{n-1} x_{0,k}\, \delta_k}{\sum_{k=1}^{n-1} \delta_k}, \qquad \text{Eq. 4.2}$$

$$\bar{\mathbf{m}}_{1,n} = \frac{\sum_{k=1}^{n-1} x_{1,k}\, (1 - \delta_k)}{\sum_{k=1}^{n-1} (1 - \delta_{k)}}. \qquad \text{Eq. 4.3}$$

The expressions above more accurately refer to a weighted arithmetic mean, though it can be appreciated that all $x_{i,n}$ are assigned equal weight: $\delta_n = (1 - \delta_n) = 1$. Thus, the result is the arithmetic mean. Alternatively, we may specify the weight in the following way,

$$\omega_{i,n} = \begin{cases} \delta_n, & i = 0 \\ (1 - \delta_n), & i = 1 \end{cases}, \qquad \text{Eq. 4.4}$$

which then determines $\omega_{i,n}$ for the corresponding $x_{i,n}$ sample. The general formula for a weighted arithmetic mean may then be used (as above), where

$$\bar{\mathbf{m}}_{i,n} = \frac{\sum_{k=1}^{n-1} \omega_{i,k} x_{i,k}}{\sum_{k=1}^{n-1} \omega_{i,k}}. \qquad \text{Eq. 4.5}$$

Finally, the ILM assumes the observer attempts to locate the decision criterion at the optimal location. Accordingly, on each trial the criterion is defined as the midpoint between the estimated means: $c_n = 0.5 \cdot \left[\bar{\mathbf{m}}_{0,n} + \bar{\mathbf{m}}_{1,n}\right]$.

By a process similar to that assumed by Helson (1949), the magnitude of the sampled stimulus on each trial determines the characteristics of the criterion shift. Though, unlike the weighted standard, the direction and size of the shift in the criterion is determined by the magnitude of the sampled stimulus relative to both the observer's current criterion value and the estimated mean of the sampled distribution. For now, ignoring the relationship between $x_{i,n}$ and the criterion (that is, regardless of the decision the observer has made), it is necessary to describe how the criterion shifts in relation to the mean of the sampled distribution. Assume, then, that on trial n a sample is drawn from the signal distribution, $x_{1,n}$. If $x_{1,n} < \bar{\mathbf{m}}_{1,n}$, addition of a value smaller than the mean will decrease the mean value, thus $\bar{\mathbf{m}}_{1,n+1} < \bar{\mathbf{m}}_{1,n}$. The distance between the means has now decreased and so the midpoint (optimal decision criterion point) has also shifted. Consequently, this will result in the criterion shifting *away* from the estimated mean of the sampled distribution, $c_{n+1} < c_n$. Conversely, if $x_{1,n} > \bar{\mathbf{m}}_{1,n}$, the updated estimate will increase, $\bar{\mathbf{m}}_{1,n+1} > \bar{\mathbf{m}}_{1,n}$, resulting in a shift *toward* the estimated mean of the sampled distribution, $c_{n+1} > c_n$ (cf. Kubovy & Healy, 1977a). The same principle applies for $x_{0,n}$. To further clarify shifting toward or away, note that the following relations hold:

$$\Delta = \begin{cases} \left(c_{n+1} > c_n \middle| x_{0,n} > \bar{\mathbf{m}}_{0,n}\right), & Away \\ \left(c_{n+1} < c_n \middle| x_{0,n} < \bar{\mathbf{m}}_{0,n}\right), & Toward \\ \left(c_{n+1} < c_n \middle| x_{1,n} < \bar{\mathbf{m}}_{1,n}\right), & Away \\ \left(c_{n+1} > c_n \middle| x_{1,n} > \bar{\mathbf{m}}_{1,n}\right), & Toward \end{cases} \qquad \text{Eq. 4.6}$$

These relations have implications for criterion shifts, the first being that criterion shifts are not constant in magnitude. This is in contrast to the ALMs where all shifts are assumed to be of constant magnitude across all trials, depending on whether the shifts are following errors or correct decisions (see Section 1.4). Specifically, shift magnitude is determined by where $x_{i,n}$ falls in relation to $\bar{\mathbf{m}}_{i,n}$. The closer (or further away) the sample falls relative to the mean, the smaller (or greater) the shift magnitude. As the means becomes stable across trials the difference between the observational sample and the estimated mean has less of an influence upon the criterion shifts. In other words, the ongoing integration of information shifts the current mean less and less, meaning the criterion shifts less and less. Put another way, the ILM assumes that as n increases the shift size will approach zero; i.e., $\lim_{n\to\infty} \Delta = 0$. A second implication concerns the expected direction of shifts following specific decision outcomes. This concerns where $x_{i,n}$ falls in relation to the criterion, c_n, and Kubovy and Healy (1977a) demonstrated that there exists an ordered relationship in the expected frequency of shift direction following correct and incorrect decisions – though this relationship also depends on $\bar{\mathbf{m}}_{i,n}$.

To make the relationships clearer, consider Figure 4.1. Assuming again that we are dealing with a signal sample, $x_{1,n}$, the shaded areas in Figure 4.1 denote the regions where the sample may fall in relation to the both the criterion and the mean of the sampled distribution. These regions ultimately determine the direction of criterion shift following a specific decisional outcome. Considering for now only correct decisions, i.e., when $x_{1,n} \geq c_n$, the sample can fall within one of two regions (Panel A): either above $\bar{\mathbf{m}}_{1,n}$, or below it. The conditional probability that $x_{1,n} > \bar{\mathbf{m}}_{1,n}$ is (the light grey region)

$$P(x_{1,n} > \bar{\mathbf{m}}_{1,n} | x_{1,n} \geq c_n) = P(x_{1,n} > \bar{\mathbf{m}}_{1,n}) / P(x_{1,n} \geq c_n), \qquad \text{Eq. 4.7}$$

whereas the probability that $x_{1,n} < \bar{\mathbf{m}}_{1,n}$ is (the dark grey region)

$$P(x_{1,n} < \bar{\mathbf{m}}_{1,n} | x_{1,n} \geq c_n) = P(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n}) / P(x_{1,n} \geq c_n). \qquad \text{Eq. 4.8}$$

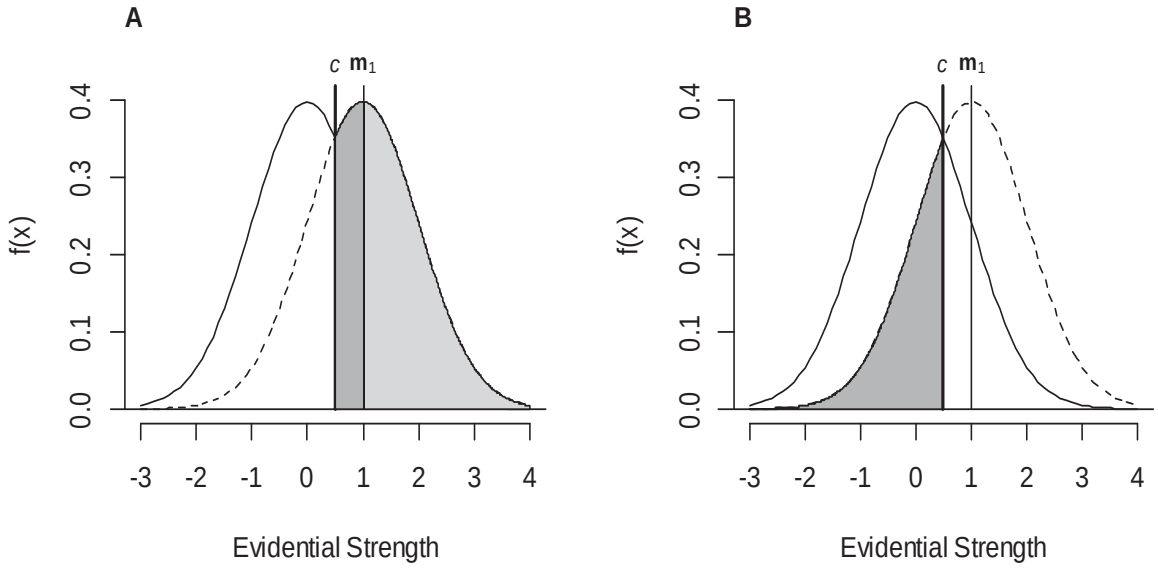

Figure 4.1: The region where a signal stimulus sample falls determines in which direction the criterion will shift. Following correct decisions (Panel A) the sample can fall in one of two regions, the smallest of which (dark grey) results in a shift away from the sample distribution mean. The larger (light grey) region will result in both a shift toward the sample mean and a shift toward the sampled stimulus value, and will occur with greater frequency. Following errors (Panel B), the sample can fall in only one region, resulting in a shift toward the sample itself while shifting away from the sampled mean.

If $P\left(x_{1,n} \geq \bar{\mathbf{m}}_{1,n}\right) > P\left(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n}\right)$ then the model predicts an increased frequency of criterion shifts toward the sampled mean: $c_{n+1} > c_n$ (cf. Equation 4.6). The same relationship holds for $x_{0,n}$, only in reverse. One can appreciate from Figure 4.1 that the area beneath the curve between the criterion and the mean, $P\left(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n}\right)$, is much smaller than the area to the right of the mean, $P\left(x_{1,n} \geq \bar{\mathbf{m}}_{1,n}\right)$, resulting in a much lower frequency of away shifts. The majority of criterion shifts following correct decisions, then, are toward the mean of the sampled distribution. Furthermore, because this can only occur in the event that $x_{1,n} \geq \bar{\mathbf{m}}_{1,n}$, the criterion is also shifting toward the stimulus sample - like a weighted standard. Considering now incorrect responses, when $x_{1,n} < c_n$, the stimulus sample can lie in only one region (Panel B). Consequently, because $\bar{\mathbf{m}}_{0,n} < c_n < \bar{\mathbf{m}}_{1,n}$, if $x_{1,n} < c_n$, it necessarily follows that $x_{1,n} < c_n < \bar{\mathbf{m}}_{1,n}$. It further follows that $x_{1,n} < \bar{\mathbf{m}}_{1,n}$, and therefore

$c_{n+1} < c_n$ (cf. Equation 4.6), meaning the criterion shifts away from the sampled mean following errors (i.e., error-correcting).

The ILM assumes that shift frequencies will occur according to the following ordered probabilities (still assuming a signal sample): $P(x_{1,n} \geq \bar{\mathbf{m}}_{1,n}) > P(x_{1,n} < c_n < \bar{\mathbf{m}}_{1,n}) > P(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n})$. In general, shifts toward the sample distribution mean following correct decisions are the most frequent, followed by away shifts after errors, and lastly by away shifts following correct decisions. Of course, toward shifts following errors are not expected (though can occur as will be discussed later). However, this order will change as the overlap in the distributions changes. Kubovy and Healy (1977a) only considered the case where $d'_{actual} = 1$; for present purposes further explication is required for easier tasks (where $d'_{actual} = 3$), though the changes are fairly straightforward. For the case where $d'_{actual} = 3$, the order will change to (signal sample) $P(x_{1,n} \geq \bar{\mathbf{m}}_{1,n}) > P(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n}) > P(x_{1,n} < c_n < \bar{\mathbf{m}}_{1,n})$; that is, the latter two terms have switched around. This switching reflects the increase in area between the criterion and the distribution mean by virtue of the rightward shift of the signal distribution: $P(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n} | d'_{actual} = 1) < P(c_n < x_{1,n} < \bar{\mathbf{m}}_{1,n} | d'_{actual} = 3)$. This rightward shift also means that a much smaller proportion of the signal distribution lies below the criterion, meaning away shifts following errors occur less frequently. In general, then, the order for easier tasks becomes: shifts toward the sampled distribution following correct decisions are the most frequent, followed by shifts away following correct decisions, and finally by shifts away following errors.

A natural consequence of the ILM is that the criterion acts very much like a weighted standard (cf. Helson, 1949; Thomas, 1973, 1975), though there is a non-zero probability that the criterion can shift away from the sampled value following correct decisions, much like the generalised error-correcting model described by Dorfman and colleagues (1971, 1975). For correct decisions there is a larger area beneath the stimulus distributions beyond the distributional means in which the stimulus may fall (cf. Figure 4.1). Naturally, this increases the probability of stimulus samples falling in these regions, which in turn pulls the criterion in the direction of the sampled value (toward the larger area where the stimulus has fallen).

The probability that the criterion will shift away from the sampled value, the area between the mean and the criterion, following correct decisions is much smaller. This type of shift is assumed by the generalised error-correction model, whereas the ILM treats the shift as a low probability event. If observers are relying on the stimulus information to update their criterion, then such a process may explain why the generalised error-correction model has had less success in fitting human data (cf. Dusoir, 1980). The ILM provides a mechanism allowing for bi-directional criterion shifts to occur following correct decisions, and so provides a further generalisation to the ALM framework. However, while these features of the ILM may serve it well, there are several caveats that limit its ability to account for human data. These issues will be evaluated in the next section by simulating the ILM.

4.2. Simulation of the ILM

To more fully evaluate the predictions made by the ILM, a simulation-based investigation was conducted. The simulated task was identical to that described above, with the inclusion of the $d'_{actual} = 3$ condition. The interest here is how the ILM performs the binary categorisation task described in Chapter 3. Prior to the simulation the stimuli were randomly drawn without replacement to derive a stimulus presentation order. On each simulation this sequence was shuffled for the new "observer". As described above, the expected performance values should approach the asymptotic values of d'_{actual} and a criterion value of approximately $d'_{actual}/2$. Each simulation consisted of a total of 400 trials (200 trials per stimulus type). The simulation was replicated 1000 times (that is, simulating 1000 hypothetical observers), the results of which are displayed in Table 4.1 and Figure 4.2.

It is clearly evident that the ILM performed the categorisation task optimally across both d'_{actual} levels, with both the criterion location and estimates of sensitivity close to their asymptotic values. Moreover, the estimated psychometric functions lack any appreciable slope (the psychometric function was fit using the `glm()` function in R). This suggests very little variability in the criterion, and each function almost completely approximates the step function specified by SDT (Figure 4.2). Noting that the total amount of criterion variance may

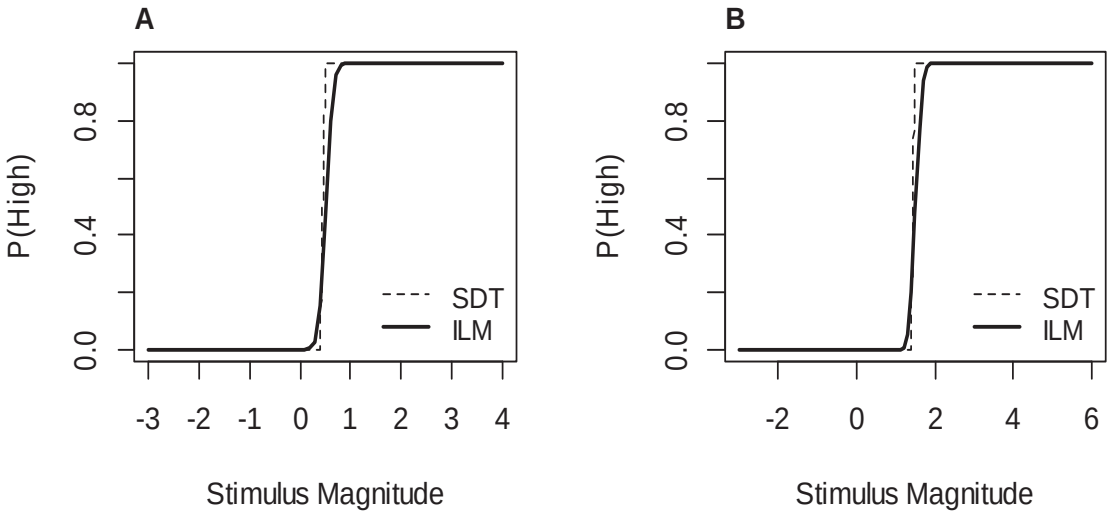

Figure 4.2: Psychometric functions based upon the averaged simulated data. The $d'_{actual} = 1$ (Panel A) and $d'_{actual} = 3$ (Panel B) functions both approximate the step function as specified by SDT. This implies that the variance associated with the decision criterion is negligible.

Table 4.1:

Average Results From Simulated ILM Across 1000 Replications.

					Shift Direction			
	k	d'	β_0	β_1	EA	ET	CA	CT
				$d'_{actual} = 1$				
M	.51	1.01	-4.74	9.30	122.51	.12	76.78	199.59
SD	.06	.04	4.69	9.05	2.45	.42	8.58	8.55
SE	.00	.00	.15	.29	.08	.01	.27	.27
Min	.72	.88	-24.96	2.01	115	0	46	168
Max	.31	1.12	-.95	49.91	132	4	110	226
				$d'_{actual} = 3$				
M	1.53	3.04	-12.01	7.97	25.29	.02	174.02	199.67
SD	.07	.06	16.80	11.18	1.46	.15	8.77	8.58
SE	.00	.00	.53	.35	.05	.01	.28	.27
Min	1.81	2.78	-71.43	1.41	21	0	140	172
Max	1.31	3.25	-2.03	47.62	32	2	202	232

Note: k = Criterion location; $d' = d'_{obtained}$; β_0 = Intercept of psychometric function; β_1 = Slope of psychometric function; EA = Away shift following error; ET = Toward shift following error; CA = Away shift following correct decision; CT = Toward shift following correct decision.

be estimated via the slope estimate, where $\sigma_c^2 = (1/\beta_1)^2$ (see Knoblauch & Maloney, 2012; Zak et al., 2012), reading the appropriate values off the table gives $\sigma_c^2 = .01$ and $\sigma_c^2 = .02$ for the $d'_{actual} = 1$ and $d'_{actual} = 3$ simulations, respectively.

In examining the direction of shifts following decision outcomes, the ILM expects an ordered relationship of shift frequency. From this point onward shifts will be referred to in the following manner: Correct-Toward (CT); Correct-Away (CA); Error-Toward (ET); and Error-Away (EA) – describing the decisional outcome and the associated direction of shift. The shift proportions are displayed in Table 4.1. For $d'_{actual} = 1$ and $d'_{actual} = 3$ simulations the expected order of shift frequencies was CT > EA > CA and CT > CA > EA, respectively. From the results in Table 4.1 these orders appear to be preserved, agreeing with the ordered relationship reported by Kubovy and Healy (1977a), and that outlined previously. Note, however, that there was a non-zero ET proportion across both simulations. While the ILM predicts that this should not happen there is a reasonable explanation for why the ET proportions are not zero. They are likely due to instability in the estimated distribution means across the initial trials, where because they are only based upon few observations they have temporarily interchanged positions. Consequently, while the criterion still remains bound between the means, the relationship has now reversed. These infrequent occurrences simply reflect a statistical property of the random shuffling, and in no way undermine the model predictions. Nevertheless, ET shifts do raise a conceptual question that was not addressed by Kubovy and Healy (1977a), and this is considered in Section 4.3.2.

4.2.1. Limitations of the ILM

One critical limitation is how the ILM integrates and stores information. The ILM assumes that all previous stimulus values contribute equally, and wholly, to the updated mean value. This implies there is no loss of information across time and thus assumes perfect memory for all past events. Accordingly, the ILM, *a priori*, describes an overly favourable interpretation of human decision-making. For each $\mathbf{m}_i$ vector, the expected mean and variance can be determined, where

$$\mathrm{E}[\mathbf{m_i}] = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}, \qquad \text{Eq. 4.9}$$

$$\mathrm{Var}[\mathbf{m}_i] = \mathrm{E}[(x_i - \mathbf{m}_i)^2]. \qquad \text{Eq. 4.10}$$

Assuming that each event is a normally distributed random variable, where $\mu_0 = 0, \mu_1 = 1, \sigma_0 = \sigma_1 = 1$, the expected mean and variance is equal to the mean and variance of the sampling distribution for the i^{th} stimulus event. This implies that where (cf. Equation 1.2)

$$d'_{obtained} = \frac{\mathrm{E}[\mathbf{m}_1] - \mathrm{E}[\mathbf{m}_0]}{\mathrm{Var}[\mathbf{m}_0]} = \frac{\mu_1 - \mu_0}{\sigma_1}, \qquad \text{Eq. 4.11}$$

performance will approach optimality; $d'_{obtained} \approx d'_{actual}$. Naturally, as $n \to \infty$ the variance associated with each $\mathbf{m}_i$ will approach asymptote, which, as previously mentioned, results in criterion variance approaching zero. In actuality, there remains evidence of non-zero shifts in the criterion, though they are of negligible magnitude and unlikely to have any great influence upon responding. That is, the criterion is now essentially fixed at the optimal location. It is therefore not surprising that Kubovy and Healy (1977a) remark that the performance of the ILM was "striking", noting the relative paucity of static criterion violations. However, such consistent performance is not typical of human data (see Chapter 1), meaning the ILM does not provide a qualitatively accurate description of human performance. It seems that the learning the ILM predicts is a little too "ideal". To further clarify this point, consider the plots in Figure 4.3. Both panels characterise criterion shifts given the ILM axioms. The plots were created by simulating an ILM observer for a binary categorisation task (for now only a single run was conducted). Specifically, the tonal stimuli for the second experiment in Chapter 3 were used by standardising the tone values according to their location along the decision axis. So, the stimuli were normally distributed random variables, where $\mu_0 = 0, \mu_1 = 1, \sigma_0 = \sigma_1 = 1$ ($d'_{actual} = 1$). The expected mean for each vector of estimates is thus $\mathbf{m}_0 = \mu_0$ and $\mathbf{m}_1 = \mu_1$, and the expected mean of the criterion is $\mathrm{E}[c] = .5$ (the ILM predicts probability

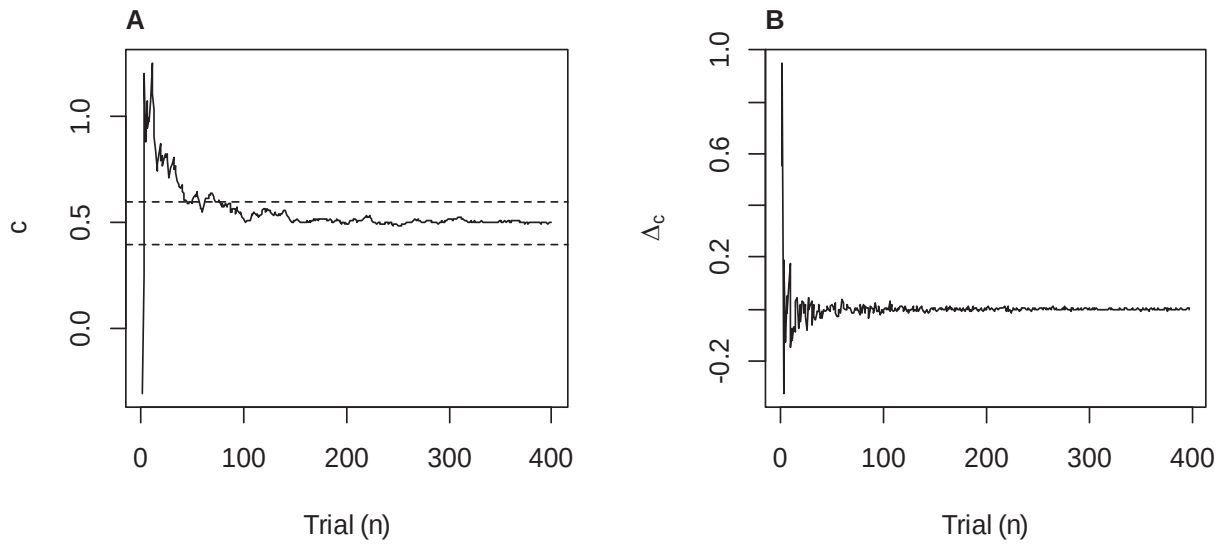

Figure 4.3: Simulated ILM data. Panel A maps the drift of the criterion across trials, while Panel B is the difference in criterion location between successive trials. It can be seen that the shifts in the criterion approach zero rapidly.

matching, at least for the $\pi = .5$ case; see Discussion). In total there were 400 trials (200 per stimulus event).

For this small simulation, the corresponding estimates were $\mathbf{m}_0 = 0$ and $\mathbf{m}_1 = .99$. Unsurprisingly, the simulated response data returned SDT estimates of $d'_{obtained} = .98$ and $c = .51$. From Panel A in Figure 4.3 one can see that the criterion tends toward the optimal location of 0.5, taking approximately 100 trials to do so. The dotted boundary lines denote the region $\mathrm{E}[c] \pm .1$ and are arbitrarily defined. They are included more as a descriptive aide in demonstrating the relative inertia of the criterion as n increases. Once the criterion enters this region the size of the shift across trials never exceeds the boundary; that is, $\Delta < .1$. This is clarified further by considering Panel B. This plot takes the difference between criterion locations on each trial, Δ_c, and so measures the shift magnitude on each trial. The apparent shifts in the earlier trials are larger than most, though can still be considered slight in context. Further, it is apparent that shifts tend toward zero quite quickly (around the 50 trial mark), where the average shift magnitude was approximately $\bar{\Delta}_{c,n} \approx .002$ across all trials. All in all,

the ILM observer performed the detection task much more like that of the ideal observer than a human observer.

4.3. Criticisms of the ILM

Kubovy and Healy's (1977a) ILM describes a dynamic criterion process that facilitates bidirectional criterion shifts and so qualitatively handles their empirical findings. However, upon further analysis Kubovy and Healy rejected the ILM, because it could not: a) account for a small number of ET shifts that were present in the cut-off report data; b) could not account for their observers making CA and CT shifts with almost equal frequency (the ILM predicts CA shifts to be the more frequent shift); and c) account for the fact that under equal prior probabilities the ILM predicts probability matching whereas the majority of Kubovy and Healy's observers did not exhibit such behaviour.

4.3.1. Probability Matching

A relatively robust feature of observer responding is that observers attempt to match the frequency of a specific response with the prior probability for that event (e.g., responding "signal" 50% of the time where $\pi = .5$ for signal events). ALMs have been shown to predict asymptotic response probabilities that match the prior probability of stimulus event (e.g., Dorfman, 1977; Dorfman & Biderman, 1971; Thomas, 1975) - at least for error-correcting models (see Thomas, 1973). The criterion is shifted toward the location that maximises the proportion of responses that match the more frequent event. The ILM also makes such predictions, though only for the $\pi = .5$ case considered here. Dorfman (1977) subsequently criticised the ILM for being less than ideal because it would fail to predict probability matching for all prior probabilities. More generally, the criterion would completely fail to shift toward the optimal location. He noted that, because the ILM attempts to always locate the criterion at the midpoint between the distributional means (and is thus postulating a sensory criterion and not a likelihood criterion, further limiting its utility), it ignores all prior information and as such should be rejected. In response, Kubovy and Healy (1977b) demonstrated that a general ideal learner model could be constructed that predicted

asymptotic probability matching across all prior probabilities. Therefore, the ILM as it was designed for the purpose of their study is a restricted version of a more general model that predicts probability matching for all possible priors. However, Kubovy and Healy (1977b) reason that, given that this simple ILM is unable to account for the observer data, a generalised ILM would also have to be rejected as it will, too, predict probability matching, which was not observed in the empirical data.

Despite offering a solution to Dorfman's probability matching criticism, Kubovy and Healy (1977b) note that it was precisely *because* the ILM predicted probability matching that it was rejected; their data indicated that some observers were not probability matching. Ultimately, Kubovy and Healy (1977b) agree with Dorfman (1977) that the ILM should be rejected, though each party cite antithetical reasons as to why. In fact, the emphasis placed upon the model's ability to predict probability matching may be wholly unnecessary for general dynamic criterion models. It has been well established that error-correction models predict asymptotic probability matching (see Theorem 2.1, Thomas, 1973); critically, there exists a unique stationary distribution for all prior probabilities. The criterion will consistently shift toward the position that facilitates probability matching and fluctuate about this point. However, Thomas (1973) established that this is not always the case for general error-correction models that allow for shifts following correct decisions (e.g., Dorfman & Biderman, 1971). A unique stationary distribution does not exist for all prior probabilities (though the general ALMs do predict probability matching for the $\pi = .5$ case) and so will not always predict probability matching. Dusoir (1980) also concluded that his observers did not universally probability match, and that observers could not be uniformly fit by a single ALM.

While the absence of probability matching may be overlooked, a requirement for any SDT-type model is a sensitivity to stimulus priors and so must necessarily shift the criterion in such a way that observers respond in accordance with the statistical properties of the stimuli. While many observers may fail to completely probability match, their conservativeness (or liberalism) will be relative to the optimal criterion location, which is determined by the prior probability of each stimulus event. The fact that the ILM is unable to

probability match is a serious limitation; this said, for the limited scope of enquiry herein (assuming only equal base rates, equal variances, and a monotonic decision axis), the model provides an adequate framework to which modifications may be applied. For the purpose of demonstrating the effect of TTKR, which for now have only been studied here assuming equal prior probabilities, the inability of the ILM to fully capture prior information may be put to one side.

4.3.2. The Direction of Criterion Shifts

Kubovy and Healy's (1977a) finding that observers were shifting their criterion in both directions with roughly equal probability following correct decisions violates the ordered relationship expected under the ILM predictions. The ILM was seen, then, not to provide a good qualitative description of the empirical response data. Furthermore, about 7% of criterion shifts were ET shifts that the ILM simply cannot account for; the origin of the shifts was left unexplained. While ET shifts were shown to have occurred during the simulations reported in Section 4.2, their presence reflects stochastic effects. It was noted, however, that ET shifts can be accounted for by a reversal of order in the estimates of the distributional means. Kubovy and Healy's (1977a) definition of away and toward shifts (cf. Equation 4.6) was predicated on the assumption that observers locate their criterion between the means of the distributions, such that $\bar{\mathbf{m}}_0 < c < \bar{\mathbf{m}}_1$, and that the order of the means remains preserved across all trials, where the noise mean is always lower than the signal mean. While the ILM axioms ensure that this is always the case, it is unlikely that human observers will always be able to maintain this order. To demonstrate how this scenario permits ET shifts, consider a criterion that is located above the signal mean, and assume that we are dealing with a signal sample. In this situation, the following relation must hold: $\bar{\mathbf{m}}_1 < c < \bar{\mathbf{m}}_0$. This is essentially a reversal of the scenario depicted in Figure 4.1. There now exists two regions *below* the criterion (instead of one) where the signal sample could fall. Further, there is now only a single region *above* the criterion (instead of two). While the relationship between the sample location and distribution mean remains

unchanged ($\bar{\mathbf{m}}_{1,n+1} > \bar{\mathbf{m}}_{1,n}$ if $x_{1,n} > \bar{\mathbf{m}}_{1,n}$), the way the criterion shifts in relation to the decisional outcome is now slightly different. Following correct decisions the criterion can now only shift in one direction, and that is upward - a toward shift (CT). However, though the criterion is now above the mean it is actually an away shift.

Following errors, the criterion can now shift in two directions. The higher probability will see the criterion shift in the usual direction following errors on signal trials (CA shifts; again in relation to the assumed ordering of the means), because $P(x_{1,n} < \bar{\mathbf{m}}_{1,n}) > P(\bar{\mathbf{m}}_{1,n} < x_{1,n} < c_n)$. However, there is a small probability that the sample falls within the region between the signal mean and the criterion, which would lead to both a) an error, and b) a *toward shift* (ET; according to the assumed ordering). That is, Kubovy and Healy's (1977a) analysis would note that $c_{n+1} > c_n$ following a $x_{1,n}$ sample and so the shift would be registered as a toward shift. However, because they assume the signal mean is always located above the criterion, the relationship between the criterion shift and signal mean is reversed and so it appears that a toward shift has occurred following an error on a signal trial. The increase in criterion ($c_{n+1} > c_n$) occurs because the sample has fallen above the estimated mean ($x_{1,n} > \bar{\mathbf{m}}_{1,n}$), though does so following an error only because the criterion itself is located above the estimated mean ($c_n > \bar{\mathbf{m}}_{1,n}$), between which $x_{1,n}$ has fallen. Of course, these relations hold for a criterion located below the noise mean. Furthermore, the relations considered currently will also change in response to an increase in area between the distributional means. The probability of a switch in ordering of the distributional means decreases considerably as the signal mean increases, and so these considerations become less of an issue. Indeed, the present considerations pertain to the $d'_{actual} = 1$ case exclusively. In sum, Kubovy and Healy's (1977a) analysis of the away and toward shift frequencies appears to be confounded by the assumption that the means preserve their ordering throughout all trials.

An alternative interpretation of the ET shifts may see them as nothing more than random effects. As such they do not reflect anything strategic or meaningful. For example, observers may have shifted their criterion differently just to see what happens, or to alleviate

boredom, or simply, just for something different. Or, they may reflect lapses of concentration and are nothing more than outliers. The empirical shift percentages may also be called into question in light of the method used to infer them. The cut-off report method requires observers to explicitly report an integer that reflects their current criterion value. As Dorfman (1977) notes, "A cut-off decision rule induced by instruction and requiring numerical report of the cut-off may obey very different laws from the cut-off process on the internal observation axis" (p. 448). In other words, the ET shifts may simply be an artefact of the report method. If this is the case then the rest of the shift percentages would be subject to the same criticism; however, there is certainly evidence to suggest that the shift percentages are not an artefact of the response format. For example, the similarities in the number of static criterion violations between the two conditions suggest there is merit to the report method. The ET shifts also appeared consistently across all sessions (the percentages were relatively stable across sessions). This suggests that the shifts perhaps reflect more than a random re-positioning of the criterion due to boredom or concentration lapses. Also, the fact that the ALM cannot accommodate bi-directional shifts means that it offers very little insight into what *is* an appropriate shift process. Conversely, cut-off reports do offer some insight into observer shift processes, and do so without imposing directional constraints upon the shifts. The method, then, may provide a valid proxy for evaluating criterion shifts in a more observable way. A patent limitation of the ALM is that it normatively prescribes shifts; that is, "specifying that the direction of shift be entirely determined by the identity of the distribution from which the observation on the preceding trial had been drawn" (Kubovy & Healy, 1977a, p. 444). Further, the ALMs all assume criterion shifts are of a constant magnitude. While the ALM provides one version of a dynamic decision criterion, a better model might do well to assume that shifts can occur in both directions.

4.4. Summary of the ILM

While the ILM provides a generalisation to the current dynamic criterion accounts, from the present evaluation and the assessments made by Kubovy and Healy (1977a), the

ILM performance can only be described as optimal. Evaluation of the asymptotic performance suggests negligible criterion shifts across trials and psychometric slopes that approximate a step function. While the criterion is shifting, the magnitudes of the shifts is negligible and unlikely to perturb the criterion significantly beyond the optimal location. In effect the criterion appears to be almost fixed. While this facet of the ILM renders it inconsistent with empirical data, Kubovy and Healy rejected the model on the ground that the empirical shift frequencies following correct decisions were not entirely consistent with the model predictions. Observers appeared to be shifting their criterion in both directions with equal frequency following correct decisions, whereas the ILM predicts that CT shifts should occur more frequently. Additionally, the small percentage of ET shifts cannot be accounted for by the ILM. It seems that the disparity between the empirical and model shift frequencies is superfluous given the criterion exhibits no appreciable variance, and should be rejected on this fact alone. However, the observed ET frequencies can be explained by the ILM, provided the assumption that the noise mean remains below the signal mean all of the time is dropped. Though the axioms of the ILM do not permit this to happen, one can envision circumstances that would perhaps precipitate such a scenario. For example, consider a run of noise trials that all engender relatively high evidential magnitudes. It would generally be the case that such samples exceed the noise mean, which, if the observer is faithfully using the TTKR, would in turn increase the noise mean estimate (because the observer accepts the samples as truly noise samples). If the estimated signal mean happened to be lower than its asymptotic value then it is possible for the mean estimates to cross over. That is, given a sequence of stimuli where the TTKR appears to be violating observers' expectations of what the stimuli should be, observers may switch their responding to match the reinforcing information, thus reversing their responding. While this seems counterintuitive, such practices are not without precedent.

It is well established that TTKR can influence observer responding and the nature of sequential dependencies (e.g., Atkinson et al., 1964; Friedman et al., 1968; Lee & Zentall, 1966; Massaro, 1969; Tanner et al., 1970). In particular, Friedman et al. (1968) suggest that

the feedback-reinforced response is strengthened when the evidential strength for stimuli is ambiguous. Furthermore, the bias toward a response is a monotonically increasing function with the frequency of the TTKR event, so if a signal response is more frequently reinforced, then the observer will respond signal more often. Effectively, observers are trying to match their response with the TTKR in the absence of certainty. However, Massaro (1969) demonstrated that this tendency was not exclusively confined to states of uncertainty. In fact, observers tracked the TTKR and attempted to match their responses with the reinforcing events even where the evidential strength for a stimulus was unambiguous. Massaro, manipulating the veracity of the TTKR, found that as the reliability of the TTKR approached zero observers were actively switching their responses (e.g., calling a loud tone soft and vice versa).

While such extreme instances are not possible under the present conditions, it certainly seems that TTKR could have similar, if temporary, effects that could periodically displace the estimates of the means. Indeed, Massaro (1969) notes that these effects would be difficult to explain solely in terms of a shifting criterion account. In effect, the criterion would remain located somewhere between the means, though what has occurred is a temporary reversing of the psychophysical environment, as it were. However, while providing a possible explanation for ET shifts, a crossing over of the distributional means cannot account for the CT and CA shift percentages, which suggests other factors are unaccounted for, such as random criterion shifts.

A strength of the ILM is that it considers not only where the stimulus had been sampled from, but also how information is integrated across trials, and how this process might affect the positioning of the decision criterion. The evidential magnitude of the observational sample is also accounted for, which in turn proportionally influences the criterion location. Intimately tied to this is how observers use TTKR in determining the true state of affairs. It is through the tracking and collation of information from past stimulus events that the shifts in the criterion are determined. This process is reminiscent of Treisman and Williams' (1984) criterion-setting theory (CST). The CST suggests that observers track

events within their psychophysical environment, which ultimately determines where the criterion is located. Each response induces a shift in the criterion, which is determined by a shift parameter, and is of a constant size. These shifts are recorded as individual traces (all past shifts are recorded) which decay at a constant rate, pulling the criterion back toward a reference value. The current criterion location, then, is the sum of all latent traces (those that have not decayed to zero), the number of which determines the depth of the sequential dependency. How quickly the traces decay determines the availability of past shift records in contributing to the current criterion location (e.g., if three traces are active then a sequential effect with a lag of $k = 3$ is assumed). While the ILM prescribes a similar tracking of past observational samples, an issue with the ILM as it stands is that *all* past events are stored. There is no loss of information across time. It is this facet of the model that results in performance that is very close to optimal. However, this tracking behaviour makes the ILM amenable to modeling TTKR effects. In the default case, the ILM receives $TTKR_e$ and simply updates the estimated mean that matches the TTKR label. The distribution that the sample was drawn from contributes toward the estimated mean of that sampled distribution; however, this need not be the case. As will be demonstrated in the next section, the distributional label (i.e., TTKR) attached to the observed sample may affect how, or if, the information is stored and integrated. In light of this it appears that the ILM has more to offer in terms of describing categorisation performance.

In summarising, while the ILM to some extent addresses the problem of how a criterion is established and maintained in terms of stimulus representations, which is a vast improvement upon the ALM framework, the ILM fails to accurately account for observer data (Kubovy & Healy, 1977a; see also Turner et al., 2011). While this does not completely rule out the ILM as a potentially useful cognitive model, what the foregoing discussion largely boils down to is that inherent cognitive limitations are not accounted for in the ILM. For the most part, many of the incompatibilities between the ILM and the data come about because an idealised model is being utilised to cope with imperfect data. In the section to follow we will explore ways in which the ILM may be modified to better model human decision making.

Additionally, we will explore ways in which the modified ILM can be used to model the effects of TTKR.

4.5. The Exponentially Weighted Learner Model (EWLM)

The major reason why the ILM provided a poor qualitative description of human performance was that there was no information loss; the ILM had a perfect memory for all past events. The EWLM is a modified ILM whereby the stimulus information available in the updating of the distribution means is exponentially weighted so that only the most recent stimulus occurrences are able to contribute. This, then, addresses the issue of information loss. Interestingly, the need to limit the amount of information stored by the ILM has been investigated before. Ward and colleagues (1988) presented an alternate model of probabilistic categorisation where they used Markov chains as the underlying shift mechanism (more aligned with the ALM approach). Their "ignorant" Markovian observer was required to estimate the transition probabilities, the probability of shifting the criterion following certain event outcomes, based upon previous observational samples. The number of available past occurrences was manipulated which influenced the observer's "memory". Ward et al. established, via simulation, that as the availability of past observational samples decreased, the model began to better approximate human performance. Unfortunately, like the ILM, all past informational values were assumed to have been stored veridically, and so the model presumed no loss of information *per se*; just a limiting on the number of available items. While the present discussion endorses such a limitation, it is argued that the decay function described below provides a more principled, or psychologically congruent, approach to limiting information. An additional modification that has not previously been considered is the way in which stimulus values are assigned to each distributional estimate. It is possible to manipulate the way items are integrated, allowing one to explicitly model the effects of TTKR. In this way, the EWLM offers a novel approach to modeling the supposed TTKR effects reported in Chapter 2.

4.5.1. Modeling Information Loss

The load on memory may be manipulated in a number of ways, one of which is to change the number of items that can be used to contribute to the estimation of the mean (see Turner et al., 2011; Ward et al., 1988). Another way might be to store items on only some proportion of the trials (Turner et al., 2011). Finally, we may limit the contribution of each stored item to the updating of the mean. This is achieved by assuming that the more recent items contribute more strongly, whereas older items contribute less. All of these approaches have psychological interpretations. Manipulating the total number reflects limitations in the capacity of our memory; storing only some proportion of the stimuli may reflect attentional lapses; while differences in the value of the stored stimuli reflect the fact that information decays over time.

Ultimately, what these three possible parameters serve to do is affect the way the distributional information is accumulated across trials, which is summarized by the trial-by-trial updates in the distributional means. Having a decay, an attention, and a limit parameter, all varying, provides very little in terms of tractability as they all can trade off against one another. This can result in various parameter combinations that are not unique, limiting what can be inferred from the model. Instead, and in the interests of keeping the model as simple as possible, only a single "memory" parameter will be used: decay. Not only does item decay address the problem of having all past events contribute equally to the updated distributional means, but for all practical purposes item decay inherently places a limit upon capacity. This is so because items that were presented further back contribute negligibly. The items that provide a significant contribution are only going to be the most recent items. Additionally, for convenience it will be assumed that all items are stored across trials. Though lapses in attention are likely to occur, the inclusion of an attention parameter will not be of substantive value. Whether one chooses to limit the number of items through either a decay process or through only allowing a proportion of the items to be stored, the overall effects are identical;

the distributional means begin to vary. As mentioned previously, having a number of parameters that are all influencing a process in the same way seems redundant.

With this in mind, given that decay can freely vary does imply that "capacity" may be quite large, given certain decay values. It should be pointed out that capacity in this sense does not refer to what we would usually consider capacity to be, as, for example, in working memory research. In fact, where limits have been placed on the number of past items (see both Turner et al., 2011, and Ward et al., 1988), the item limit far exceeds what we would consider reasonable in terms of short term memory. Rather, in a vein similar to that taken by Treisman and Williams (1984), the model allows for a perceptual prototype (in this case a distributional mean) to be maintained and updated through the tracking of past events. This does not necessarily imply that *all* past events are stored in their entirety, but a weighted representation that has been continually updated on each trial. The storing of items reflect this effect of stimulus history, or tracking, and not a memory store *per se*. I do not believe it is necessary to place too much emphasis upon this idea of capacity, and it has not overly concerned others who have modeled human performance in similar ways (Kubovy & Healy, 1977a; Turner et al., 2011; Ward et al., 1988). Insofar as placing a psychological interpretation upon this storing and tracking process, one would surmise that the process is somewhat passive and is not an active maintenance of past representations. That is to say, observers are not explicitly aware that they are maintaining a distributional representation of each stimulus type. Instead, it is more that the observers have a general sense of what a high or low tone is given their past experiences, and that this general perception can change over time given the type of information that is available (e.g., TTKR). So, the critical feature, and fundamental point of enquiry, rests with how information about the stimulus contributes toward the establishment and maintenance of the decision criterion. In this sense we might refer to it as an effect of past history, where larger decay values correspond to a greater reliance upon past stimulus occurrences.

In now making the decay process more explicit, we note that a core assumption of the EWLM (as it is for the ILM) is the maintenance of a mean for each stimulus distribution,

between which it is assumed the criterion is located. The weighting upon past events essentially defines the "memory" of the model, which in turn determines the extent to which previous values contribute toward the current estimate. The difference between the ILM and the EWLM is the weighting scheme placed upon the means. While for the ILM the weights are uniform for all past stimulus values, the EWLM instead applies an exponential weighting scheme. Accordingly, we may define the weight on trial n as

$$\omega_n = \lambda^{n-k}; \quad 0 \leq \lambda \leq 1\,, \qquad \text{Eq. 4.12}$$

where λ is the decay rate at lag (or time) $n-k$, k indexes the n^{th} trial, $k = 1, 2, 3 \ldots, n$, and n denotes the current trial. Larger values of λ mean that the decay occurs slowly, and so the memory for past events is very good. Smaller values imply faster decay meaning that items in memory are relatively short lived. It requires little effort, then, to modify the existing ILM model so the estimated distributional means are subject to the effects of imperfect memory. Assuming again the mean for the i^{th} stimulus class is $\bar{\mathbf{m}}_i$, and recalling that each event is indexed by the delta parameter, δ_n (cf. Equation 4.1), we can define the weight according to all past event occurrences as

$$\omega_{i,n} = \begin{cases} \lambda^{\delta_{n-k}}, & i = 0 \\ \lambda^{(1-\delta_{n-k})}, & i = 1 \end{cases}. \qquad \text{Eq. 4.13}$$

Applying the weighting to all previous observational samples, the corresponding mean is then calculated by way of Equation 4.5, and so we have

$$\bar{\mathbf{m}}_{i,n} = \frac{\sum_{k=1}^{n-1} \omega_{i,k} x_{i,k}}{\sum_{k=1}^{n-1} \omega_{i,k}}\,, \qquad \text{Eq. 4.14}$$

$$= \frac{\omega_{i,1}x_{i,1} + \omega_{i,2}x_{i,2} + \cdots + \omega_{i,n-1}x_{i,n-1} + \omega_{i,n}x_{i,n}}{\omega_{i,1} + \omega_{i,2} + \cdots \omega_{i,n-1} + \omega_{i,n}}, \qquad \text{Eq. 4.15}$$

$$= \frac{\lambda^{199}x_{i,1} + \lambda^{198}x_{i,2} + \cdots \lambda^{1}x_{i,199} + \lambda^{0}x_{i,200}}{\lambda^{199} + \lambda^{198} + \cdots \lambda^{1} + \lambda^{0}}. \qquad \text{Eq. 4.16}$$

So, the estimated means are now exponentially weighted arithmetic means, where Kubovy and Healy's (1977a) ILM becomes a special case (where $\lambda = 1$). Note that, for both the ILM

and the EWLM, the mean is only updated on trials where $\delta \neq i$. This means that on trials where $\delta = i$, $\bar{\mathbf{m}}_{i,n}$ remains unchanged. It would also seem plausible that there is some change in the value of the mean even on trials where it is not being updated; i.e., having an item added to $\mathbf{m}_i$. Instead, on the trials where a mean is not updated the previously added stimulus sample, $\mathbf{m}_{i,n-1}$, is simply carried over. Specifically, the trial-by-trial accumulation can be expressed as,

$$\mathbf{m}_{0,n} = \begin{cases} x_{i,n}, & i = 0 \\ \mathbf{m}_{0,n-1}, & i = 1 \end{cases}, \qquad \text{Eq. 4.17}$$

$$\mathbf{m}_{1,n} = \begin{cases} \mathbf{m}_{1,n-1}, & i = 0 \\ x_{i,n}, & i = 1 \end{cases}. \qquad \text{Eq. 4.18}$$

As an additional matter of convenience, instantiating this assumption means that the δ parameter can be removed altogether, leading to a simpler weighting where$\omega_n = \lambda^{n-k}$. This value can then be plugged into Equation 4.13 above. With the new weighting scheme in hand, we may define the criterion as before, where $c_n = 0.5 \cdot \left[\bar{\mathbf{m}}_{0,n} + \bar{\mathbf{m}}_{1,n}\right]$. It should be noted that the EWLM predicts that performance will continue to improve as more items become available for averaging. This might seem somewhat counterintuitive as one could argue that the maintenance of more items in memory would compromise performance, owing to increased cognitive load which may introduce greater efforts in remembering. This is a fair consideration, though the idea can be made more palatable when we consider that the total number of past stimuli that can be held is limited (cf. Ward et al., 1988). As previously discussed, this conjecture is necessary given the exponential loss.

The limitations imposed on how much information can be retained mean that observers may need to place varying reliance on alternative sources of information (e.g., TTKR). For example, observers may choose to track, or remain increasingly mindful of, past events (which will consume capacity, though should facilitate performance by increasing the number of active stimuli) when the nature of stimuli is increasingly ambiguous. This tendency, if viable, should be reflected in the decay parameter. While performance is expected to become poorer for lower decay parameter values, it is assumed that some level of

retention is present; in other words, there will be at least one item in memory. This must be so in order to maintain a mean distributional value, and so decay values approaching zero should not be expected. Thus, smaller decay values (which result in fewer active items) do not necessarily entail poorer performance, and by the same token larger decay values (which result in more active items) do not necessarily entail better performance. The way in which the items have been stored, and their associated values, will ultimately affect how performance is affected. In conditions of uncertainty we might expect to see larger decay values (observers are placing greater weight on past events and they are decaying more slowly), though the overall quality of the items contributing toward each distributional mean is inconsistent. Conversely, if the observers are integrating information that is more consistent then we might expect to see lower decay rates and less reliance on items further back. The next section explores how TTKR may be modeled and how the various types of TTKR can affect the trial-by-trial updating of the criterion.

4.5.2. Modeling the Effects of TTKR

In this section an approach to modeling the effect of TTKR is discussed, where the effects of TTKR reported in Chapter 2 and 3 should be reflected in the predictions made by the EWLM. To model the effect of TTKR the importance of the category label is stressed as it is this that is assumed to determine which mean (i.e., which vector) the stimulus is assigned to. The default case for the EWLM to this point (and for the ILM, too) is that the stimulus is assigned to the vector that corresponds to the distribution sampled from. This, then, effectively describes $TTKR_e$ and is tantamount to providing event-based information. To begin with, let us define the TTKR event on each trial as φ_n and so determine which $\mathbf{m}_{i,n}$ the stimulus value is assigned to. For the default case ($TTKR_e$) we define

$$\varphi_n = \begin{cases} 0, & i = 0 \\ 1, & i = 1 \end{cases}, \qquad \text{Eq. 4.19}$$

and so the TTKR value corresponds to the i^{th} distributional value. As a general consideration, let it be noted that the element-wise update to each distributional vector occurs with the following probabilities,

$$\mathbf{m}_{0,n} = \begin{cases} x_n, & (1-\varphi_n) \\ \mathbf{m}_{0,n-1}, & \varphi_n \end{cases} \quad \text{Eq. 4.20}$$

$$\mathbf{m}_{1,n} = \begin{cases} \mathbf{m}_{1,n-1}, & (1-\varphi_n) \\ x_n, & \varphi_n \end{cases}. \quad \text{Eq. 4.21}$$

These equations are essentially rewrites of Equations 4.17 and 4.18, and define, by case, how each vector is updated according to the TTKR that has been presented on each trial. Thus, the trial-by-trial updated mean estimates can be expressed in the following way:

$$\bar{\mathbf{m}}_{0,n} = \frac{\sum_{k=1}^{n-1} \omega_k \left[(1-\varphi_k)x_k + \varphi_k \mathbf{m}_{0,n-1}\right]}{\sum_{k=1}^{n-1} \omega_k}, \quad \text{Eq. 4.22}$$

$$\bar{\mathbf{m}}_{1,n} = \frac{\sum_{k=1}^{n-1} \omega_k \left[\varphi_k x_k + (1-\varphi_k)\mathbf{m}_{1,n-1}\right]}{\sum_{k=1}^{n-1} \omega_k}. \quad \text{Eq. 4.23}$$

Note that the assignment is deterministic. That is, some value is either added or carried over on each and every trial. This, of course, need not be the case, and one could conceive of instances where the stimulus may not be assigned to any vector, or assigned to the wrong vector, but such cases will not be explored here.

With these considerations in mind we can now explore ways in which the TTKR can affect how the stimulus information is accumulated over trials. Turning to the $TTKR_i$ case, the distributional origin of the stimulus is irrelevant. Instead, φ_n is determined according to which side of the optimal decision criterion the stimulus is sampled from. Thus, we rewrite Equation 4.19 above in the following way:

$$\varphi_n = \begin{cases} 0, & x_n < c_o \\ 1, & x_n \geq c_o \end{cases}, \quad \text{Eq. 4.24}$$

noting that c_o is the optimal decision criterion and proportional to $d'_{actual}/2$. Having determined the TTKR label in this way the values need only be plugged into Equations 4.22

and 4.23 above. Similarly, in the no TTKR case we assume that observers use their response as a proxy for feedback, and so we again rewrite Equation 4.19 in the following way:

$$\varphi_n = \begin{cases} 0, & R_n = 0 \\ 1, & R_n = 1 \end{cases}. \qquad \text{Eq. 4.25}$$

In demonstrating the TTKR effects, consider Figure 4.4. To generate these plots the EWLM was simulated across five decay values, $\lambda = \{.1, .3, .5, .7, .9\}$, while keeping difficulty constant ($d'_{actual} = 1$). It is evident that as the decay parameter approaches zero the volatility in the criterion (and the mean estimates) increases substantially. In fact, there is significant overlap and crossing over between the mean estimates, though the degree to which this occurs does depend on the type of TTKR provided. As more items are allowed to contribute to the mean estimates, the separation between each estimate increases, and so the criterion shifts less and less. The crucial point is the $TTKR_i$ column (right column) reveals consistent separation between the distributional means across all decay values; this is not apparent for the $TTKR_e$ and no TTKR cases. So, while there may be great volatility in the means, performance can be improved by providing more consistent TTKR. Further, in order to improve performance where the quality of the information is not as good, improvements in performance require the observer to place greater reliance upon past stimulus events in order to stabilise the concomitant trial-by-trial mean and criterion updates. However, relying on an increased number of stimuli does not necessarily entail, as discussed previously, improved performance. This need not be the case with $TTKR_i$, and so performance can be improved while also relying upon fewer items.

This trade-off between TTKR type and reliance on past events (i.e., the decay parameter) is displayed in Figure 4.5 where the changes in d' across both levels of difficulty are plotted as a function of both TTKR type and decay. In the difficult condition, performance in uniformly expected to be better across the whole range of decay values where $TTKR_i$ is provided. In the easy condition performance is expected to fairly similar between the types of TTKR across the whole range of the decay values. In the $TTKR_e$ and no TTKR cases,

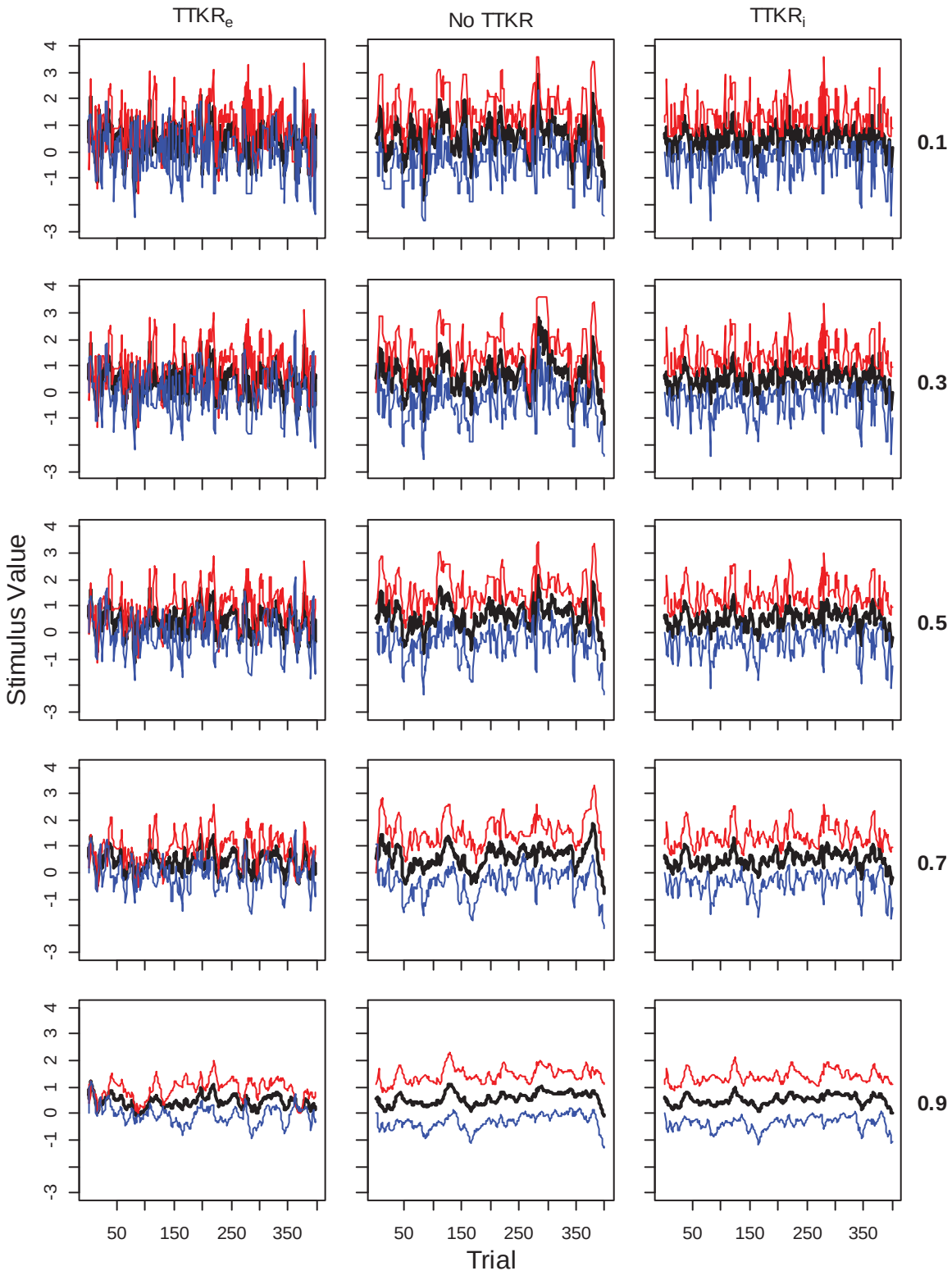

Figure 4.4: The simulated effects of TTKR manipulations for the EWLM. The dark thick line reflects the criterion shifts across trials, where the line above (or below) is the signal (or noise) mean estimates. As the speed with which items decay increases (rows), both the distributional estimates and the criterion begin to exhibit increased volatility, and the distributional estimates increasingly overlap. The extent to which this occurs can be influenced by the type of TTKR provided (columns). Decay parameter values are expressed on the right-hand ordinate.

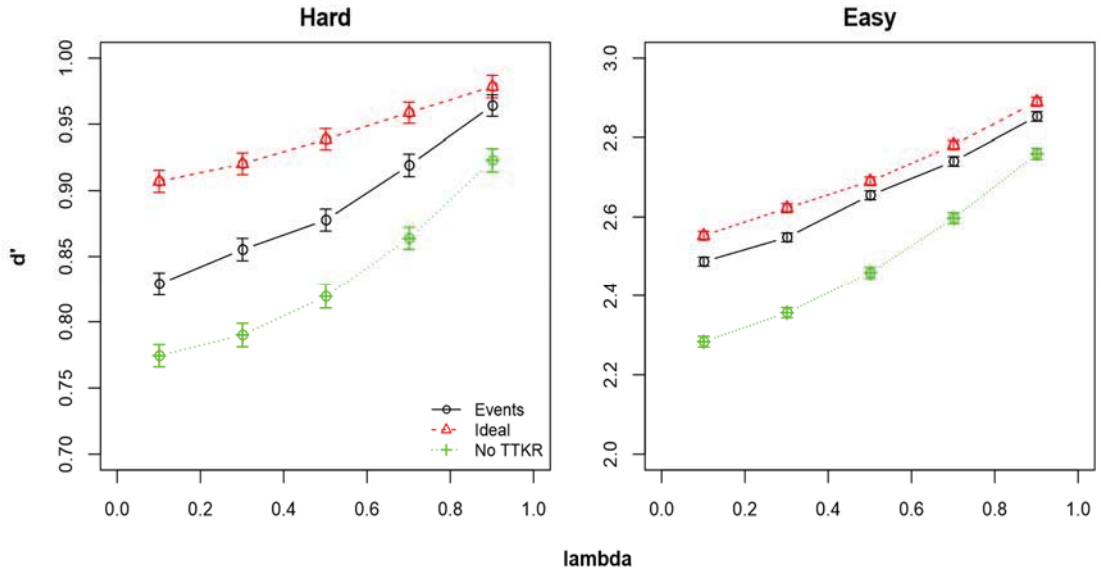

Figure 4.5: Asymptotic d' values based upon 1000 simulations per decay value while manipulating the type of TTKR the EWLM received across both the hard (left panel) and easy (right panel) conditions. The patterns displayed in this figure resemble the findings described by Chapters 2 and 3. In difficult conditions, providing $TTKR_i$ increases performance quite considerably, even when only a few items are available to update the distributional mean. Error bars are SE of mean. Legend details: Events = Events-based TTKR; Ideal = Ideal-based TTKR.

performance can only improve by placing more reliance on past events; however, in both cases performance will always be inferior to the $TTKR_i$ case.

One evident discrepancy in the model predictions for d' and the empirical data presented in Chapter 3 is that performance is expected to be worse for observers receiving no TTKR compared to $TTKR_e$ in the hard condition (Figure 4.5, left panel). However, the model predictions are correct for the easy condition. It was mentioned in Chapter 3 that there might be some dependencies between the observed TTKR and no TTKR data (observers completed both TTKR and no TTKR sessions) and so the data observed for the no TTKR group in the hard condition may be influenced by this. That is, it is unclear whether the model is making a poor prediction or whether the observed data are wrong. Before drawing any further conclusions regarding the EWLM predictions, it will be best to wait until more data have been collated.

One issue that has been ignored up to this point is the degree of asymptotic criterion variance that the EWLM can predict. Under the assumption that the criterion must remain constrained between the distributional means, shift magnitudes are also necessarily constrained. There is a limit on the absolute amount the criterion may shift from trial-to-trial. The consequences of this are threefold: first, and advantageously, the criterion distribution is stationary. This implies that the criterion will sensibly shift around a mean point and not wander off into unreasonable areas of the decision axis. The second, and not so useful, consequence is that having a constrained criterion poses limits upon the degree to which the model can recover observer response data, particularly in conditions where criterion variance is expected to be quite large (where $TTKR_e$ is provided). To get a sense of the magnitude of criterion shifts, the model was simulated 1000 times (assuming that $TTKR_e$ was provided) with the decay parameter set to zero, and where the standard deviation of the trial-by-trial criterion values were used as a proxy for a general criterion shift magnitude, $\Delta_c = SD(c_n)$. The asymptotic shift magnitude was $\Delta_c \approx .7$, which corresponded with a $d'_{obtained} \approx .83$. This is a much higher sensitivity estimate than those observed empirically. The third and final consequence is that the model cannot be generalized to cases where the stimulus base rates are not equal. This severely limits the generalisability of the model beyond the equal base rate case. This will be discussed in the section that follows.

4.6. Discussion and Summary of the EWLM

The second half of this chapter was devoted to describing modifications that can be applied to Kubovy and Healy's (1977a) ILM so that it may more appropriately model human categorisation data. The resultant EWLM model assumes that exponentially decaying information is used in updating mean values for each stimulus class across trials. The criterion, which is assumed to be located at the midpoint between these means, thus shifts with every update made to the distributional means. The model allows for the trial-by-trial stimulus information to be used across trials, where the type of feedback that is provided influences the way the stimulus information is stored. In this way, the model is able to predict

the effect of TTKR described in Chapters 2 and 3. The EWLM appears, then, to offer some solutions to problems associated with dynamic criterion models, though still having some limitations.

Many dynamic criterion models must make assumptions about the criterion *a priori*; that is, how the criterion is initially established, the nature of the distribution that criterion values are sampled from, and how stimulus information is represented across trials. The modified ILM attempts to address some of these issues by proposing that the criterion is subject to a continual, though gradual, accumulation of stimulus information. In this sense the EWLM is similar to the Dynamic SDT model posited by Turner et al. (2011). In contrast to simply summarizing the past stimulus as distributional means, Turner and colleagues reasoned that the observer uses the past information to establish distributional representations. In other words, a limited number of past events are used to build a distribution for each stimulus type. The model is non-parametric, though requires Kernal Density Estimation[13] (KDE) to obtain the PDF for each stimulus distribution. Each PDF is then updated on each trial according to a weighted combination of the current stimulus value and the previous PDF. Having an estimate of the PDF also means that a likelihood ratio criterion can be implemented; so Turner et al.'s dynamic signal detection model possess properties that a highly desirable.

A critical shortcoming of the EWLM is the constraint placed upon the location of the criterion, meaning that a) it still suffers from *a priori* assumptions having to be made about the criterion; and b) as discussed above, it significantly constrains the model, limits its generalizability, and does not completely behave as SDT models should. In the limiting case where the priors for each stimulus are equal, the EWLM can be used to describe human performance. This said, any model that is rooted in signal detection theory must be sensitive

[13] Kernal Density Estimation (KDE) is a non-parametric technique used to estimate the probability density of a random variable. Given a finite sample of random variables with unknown density, KDE is used to estimate the shape of the underlying density function, where the sample's kernel density estimator is usually defined as $f_h(x) = \frac{1}{nh}\sum_{1=1}^{n} K\left(\frac{x-x_i}{h}\right)$, where $K(\cdot)$ is the kernel, and h is the smoothing parameter which must take on values greater than 1.

to statistical information. The criterion needs to shift in response to changes in the statistical properties of the distributions. The only way this can effectively be achieved is through treating the criterion as a likelihood ratio, as was originally proposed by Green and Swets (1966). Only then will the criterion be able to shift appropriately. Turner et al.'s (2011) Dynamic SDT model possesses such abilities and can appropriately locate the criterion in cases where the base rates are not equal because the criterion is a likelihood ratio criterion. This circumvents the manifest problems with the EWLM which see it only applicable to the equal base rate case. Even so, given that the research to date has not extended beyond the equal variance case the EWLM may serve as a useful analogue and is designed with this equal base rate case in mind. Furthermore, the model can also make qualitative predictions regarding the effects of TTKR that agree with the empirical data.

The EWLM provides a further generalisation on the way the criterion shifts across trials. The EWLM predicts that error-correction will occur though it also allows the criterion to shift in both directions following correct decisions. It remains to be seen whether the EWLM can predict categorisation performance any better than the generalised models discussed in Chapter 3. Both the PGDM and EWLM consider the stimulus magnitudes on each trial and work it into the shift mechanism and so are similar in many respects. In the chapter to follow we will compare the EWLM to the generalised PECM.

Chapter 5

Testing the Exponentially Weighted Learner Model

The previous chapter first outlined an alternative dynamic criterion model known as the Ideal Learner Model (ILM; Kubovy & Healy, 1977a). Kubovy and Healy (1977a) proposed the ILM as an alternative to the conventional ALM, having reasoned that the ALMs do not provide an adequate fit to categorization data. Their conclusion was based upon a cut-off report method, where the direction of the reported shifts were deemed to be inconsistent with the axioms of the ALM. Kubovy and Healy did not perform model fits and sought a more descriptive account of the data. However, the model fitting and comparison in the present study (Chapter 3) established that criterion models that allowed for the stimulus information to be included, and the criterion to be allowed to shift on all trials, might be better able to predict the data than simpler ALM models that only used the stimulus and response sequences. The ILM, however, is overly constrained and cannot describe human data very well.

Chapter 4 proposed as an alternative to the ILM, the Exponentially Weighted Learner Model (EWLM), which sought to remedy problems with the ILM that prohibited it being considered as a viable model of human categorisation. The critical shortcoming was that the ILM assumed there was no information loss over time, which meant that shifts in the decision criterion rapidly asymptote toward zero. The EWLM instead assumes that information exponentially decays over time and so limits the number of past occurrences that can be used in updating the distributional means. An additional feature of the EWLM is that it explains shifts in the criterion through a process of information accumulation, and not simply through a deterministic shift matrix. The EWLM also provides an explanation as to how this process may be influenced through providing various types of feedback. The EWLM was shown to qualitatively capture the pattern of results discussed in Chapters 2 and 3, though there is some disagreement with the model predictions where no TTKR is provided.

In this chapter the EWLM will be put to test. A central question is whether the EWLM is an improvement over the general dynamic criterion models discussed in Chapter 3. Accordingly, the sections that follow will describe a small study which will provide some data for the purpose of model comparison. The study is essentially a replication of the second

experiment discussed in Chapter 2, with a few modifications. The model comparisons in this chapter will be very straightforward and proceed via Bayes Factors only. This is because we only need to address the predictive ability of the EWLM. The analysis side of this chapter, then, is necessarily brief. However, before moving on to the model comparisons, how the models can be constrained so that they might make better predictions will be discussed. The final section presents a general discussion and will address the findings, implications, and theoretical issues.

5.1. Some Further Consideration on Priors

In choosing what priors to use it is often beneficial to generate prior predictive distributions. By doing so some idea regarding the sensibility of the model predictions may be gleaned. If the model is making a large number of varied predictions perhaps the model is not constrained enough and requires some modification to the priors. Conversely, the model may be making predictions that are far too constrained and so the priors need expanding. In this way one could assess which priors give the model a fair chance at making sensible predictions while still being reasonably uninformative. The danger is that some may abuse this practice and specify priors that are biased toward an (erroneously) expected data pattern, serving only to maximise the consistency between the data and model predictions. The simple solution to this conundrum is absolute transparency in the reporting and access to model specifications (which includes the priors used) and code. With this in mind some further considerations regarding the priors for the dynamic criterion models are discussed here.

Firstly, it is difficult to visualise what the prior predictive distributions might look like for the trial-by-trial predictions made by the dynamic criterion models. The predictions may be plotted in the ROC space though the resulting clustering of points provide no real sense of the varying predictions made by each model. So, there is little to be gained from assessing the prior predictive distributions, and the modifications that have been made thus far have largely drawn upon assumptions that are considered reasonable given the demands of the

categorisation task (e.g., limiting the sensitivity parameter so that it does exceed the asymptotic values). In the present chapter, then, an attempt is made to model the process the observer goes through, making as few assumptions as possible. Moreover, it is desirable to try and understand how the criterion is actually established and maintained across trials. The latter is an issue that was raised by Turner et al. (2011) and is one that the EWLM, and the modeling in this chapter, might shed some light on. While the weighting schemes and mechanisms involved in the placement of the criterion may differ for each model, there is an underlying premise, outlined next, that should remain unchanged no matter which model is used.

The modeling conducted thus far has assumed that for all criterion models the initial criterion value (i.e., the starting point for the dynamic process) is a free parameter. This assumption allows for additional flexibility in the model, but it may be unnecessary and in fact psychologically incongruent. Consider that on the very first trial the observer, in most instances, will simply make a guess as to the nature (high tone or low tone) of the stimulus event. In some cases, however, the initial tone will reveal some information regarding the distribution of origin given how high, or low, it is, though for the tones that fall close to the mid-point of the tonal range a guess will be required. For all practical purposes the assumption for the initial criterion value is reflected in the prior specifications by always setting the value equal to $d'/2$. That is, it is a fixed, deterministic, parameter.

On the second trial, the observer now has some additional information that can be used to update, or indeed establish, a decision criterion. So, when the next stimulus is presented, the placement of the decision criterion is slightly more informed. In the most naïve version of this process the observer simply uses the previous stimulus value as the criterion on the current trial. What is more likely is that the new information is weighted and then added to the extant information. This process continues over all trials, where the criterion is updated in response to the newest information available. These assumptions can be applied to all the models considered in this chapter, including the EWLM. The details regarding the

models will be discussed in Section 5.3.4; the next section will detail the small empirical study undertaken.

5.2. Tone Discrimination Experiment

To ensure that all the stimuli were discriminable, a discrimination analysis was conducted which assessed the pair-wise discriminability of adjacent tones. A 2AFC task was used in which all observers were required to identify whether the tone in the second interval was higher or lower than the tone in the first interval. The stimuli and procedure used here are identical to those used in the study described in Chapter 3.

5.2.1. Observers

A total of 10 observers (nine males and one female) completed the tonal discrimination task. All observers were recruited from Massey University and had normal hearing.

5.2.2. Results

All observers were able to discriminate between tone pairs with perfect accuracy. Critically, where the numbers of JNDs were diminishing toward the higher end of the range (see Appendix C), observers were still able to correctly decide whether the tone was higher or lower than the reference. These results demonstrate that the tones used in the following study were highly discriminable.

5.3. Main Study

The experimental method used is almost identical to that used in Chapter 3. Accordingly, only the unique aspects of the method will be reported here.

5.3.1. Observers

A total of 36 observers with normal hearing participated in the experiment (six observers per experimental condition). The sample consisted of 19 males (52.8%) and 17 females with an age range of 18 to 34 years (M = 24.40; SD = 4.98).

5.3.2. Procedure

Unlike the previous experiment, the current experiment was built and presented using PsychoPy (Peirce, 2007), an open source experiment builder that uses Python code. The procedural differences largely lie in the visual layout for the response screen and feedback. All other factors (number of trials, instructions, etc.) are identical to the procedure outlined in Chapter 3. The first change was that the feedback interval was self-paced and required the observer to click the mouse to continue. The self-paced feedback interval was expected to allow observers to maximally benefit from the feedback. However, once observers were comfortable with the task, the time spent on the feedback screen shortened considerably. The second change was that observers were presented with a binary scale that had "Low" and "High" as the anchor points, and registered their responses by clicking on the blue arrow above the desired response selection. Finally, the third change was that feedback displayed the correct response in the centre of the screen, and was coloured green if the observer response was correct or red if the response was incorrect. If the observer failed to make a response within the response interval, the correct response was displayed in black. In the no KR condition a "+" was displayed in place of the feedback. Session lengths typically lasted around 45 minutes and observers were reimbursed for their time.

5.3.3. Design and Analysis

In terms of design, the only major difference for the current experiment was that all groups were independent. That is, across both levels of difficult there were independent $TTKR_e$, $TTKR_i$, and no TTKR groups. As before, all SDT parameters were estimated using hierarchical Bayesian models and the slopes of the psychometric function were also evaluated. The model specifications and prior distributions remain unchanged for the Bayesian psychometric function model, though there was a slight change in the prior distributions for the signal detection model. These differences are detailed in the next section. All code can be found in Appendix A.

5.3.4. Model Specifications

In this chapter two dynamic criterion models will be considered. With respect to the PECM, much of the discussion has already been had in regards to its specification (see Chapter 3). The only change that is made to this model is that we are no longer assuming the initial criterion is a free parameter. Instead it is fixed to be equal to $d'/2$. This means that the PGDM is now a two parameter model and requiring estimation of only d' and δ. The priors placed upon the population-level means for these parameters remained unchanged. Turning next to the EWLM, this model also requires estimation of only two parameters: d' and λ. In specifying the priors for the EWLM the same assumptions that were applied to the PGDM are also used here. Specifically, for the population-level d' parameter, in the hard condition a $Uniform(.5, 1)$ prior is used, whereas in the easy condition a $Uniform(2, 3)$ prior is used. This reflects the assumption that performance should not exceed the upper limit (i.e., the limit imposed by the overlap between the stimulus distributions) across either level of difficulty. The EWLM assumes that the information decays over trials and so requires a prior placed upon the population-level λ parameter. Seeing as we are agnostic about the range of values λ might assume a $Uniform(0, 1)$ prior was placed on the decay parameter.

With the priors established, the Bayesian implementation for the EWLM will be briefly discussed. We have assumed for all the dynamic criterion models up to this point that an observer's responses are governed by a binomial process which is driven by the hit rate and false alarm rate. So, for the j^{th} observer on the i^{th} trial we denote hits, h_{ij}, and false alarms, f_{ij}, resulting in

$$h_{ij} \sim Binomial(H_{ij}, n_{ij}^1) \quad \text{Eq. 5.1}$$

$$f_{ij} \sim Binomial\left(F_{ij}, n_{ij}^2\right), \quad \text{Eq. 5.2}$$

where n_{ij}^1 and n_{ij}^2 are the number of hits and false alarms, respectively, and

$$H_{ij} = 1 - \mathbf{\Phi}(c_{ij} - d_j') \quad \text{Eq. 5.3}$$

$$F_{ij} = 1 - \mathbf{\Phi}\left(c_{ij}\right). \quad \text{Eq. 5.4}$$

The critical difference between the PECM and the EWLM is how the EWLM assumes the criterion is shifted across trials. The EWLM assumes the criterion is located at the midpoint of the distributional means on the i^{th} trial, M_{ij}; that is,

$$c_{i+1j} = 0.5 \cdot \left[M_{ij}^{0} + M_{ij}^{1}\right], \qquad \text{Eq. 5.5}$$

where M_{ij}^{0} and M_{ij}^{1} are the estimated low and high distributional means, respectively. The criterion then shifts in response to the trial-by-trial updates in the estimated means.

5.4. Results

Turning next to the data obtained from the observers, the signal detection estimates for all experimental groups are displayed in Table 5.1. Considering first the performance in the hard condition, the effect of TTKR in the hard condition is similar to that seen in Chapter 3. The $TTKR_e$ group performance was poorer ($d' = .74$) when compared to the $TTKR_i$ group ($d' = .96$). The no TTKR observers also performed better ($d' = .85$) than the $TTKR_e$ group. This tends to confirm the pattern observed in Chapter 3, though it poses a problem for the EWLM. The EWLM predicts that performance should be worse for the no TTKR groups across both levels of difficulty. A similar result emerges in the easy data, also.

For the $TTKR_i$ group performance was better ($d' = 2.31$) than both the $TTKR_e$ ($d' = 2.14$) and the no TTKR group ($d' = 2.18$), with the no TTKR group performing marginally

Table 5.1:

MAP Signal Detection Estimates with 95% Bayesian Credible Intervals

Group	c	2.5%	97.5%	d'	2.5%	97.5%
1	.38	.24	.51	.74	.62	.85
2	.47	.33	.61	.96	.84	1.07
3	.40	.30	.53	.85	.74	.97
4	1.02	.87	1.17	2.14	2.01	2.27
5	1.08	.93	1.22	2.31	2.17	2.45
6	1.31	1.16	1.46	2.18	2.05	2.32

Note: Groups correspond to the KR Type/Difficulty combination; 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = No TTKR/Hard; 4 = $TTKR_e$/Easy; 5 = $TTKR_i$/Easy; 6 = No TTKR/Easy.

Table 5.2:

MAP Psychometric Function Estimates.

Group	α	2.5%	97.5%	β	2.5%	97.5%	μ	σ
1	-7.88	-10.01	-5.80	.59	.43	.76	13.40	1.70
2	-12.93	-14.18	-11.81	.96	.88	1.06	13.44	1.04
3	-9.38	-11.38	-7.38	.71	.58	.85	13.25	1.41
4	-7.83	-9.02	-6.78	.46	.40	.52	17.16	2.19
5	-8.95	-9.89	-8.08	.53	.48	.58	17.05	1.90
6	-8.65	-9.73	-7.62	.46	.42	.51	18.67	2.16

Note: Groups correspond to the KR Type/Difficulty combination; 1 = $TTKR_e$/Hard; 2 = $TTKR_i$/Hard; 3 = No TTKR/Hard; 4 = $TTKR_e$/Easy; 5 = $TTKR_i$/Easy; 6 = No TTKR/Easy.

better than the $TTKR_e$ group. Interestingly, it seems that providing no TTKR results in better, or at least no worse, performance than providing $TTKR_e$. The same pattern also emerges in the psychometric functions. The parameter estimates are displayed in Table 5.2 as are the transformed slope posterior distributions (Figure 5.1). There are evident differences in the slope parameters between the TTKR groups in the hard condition.

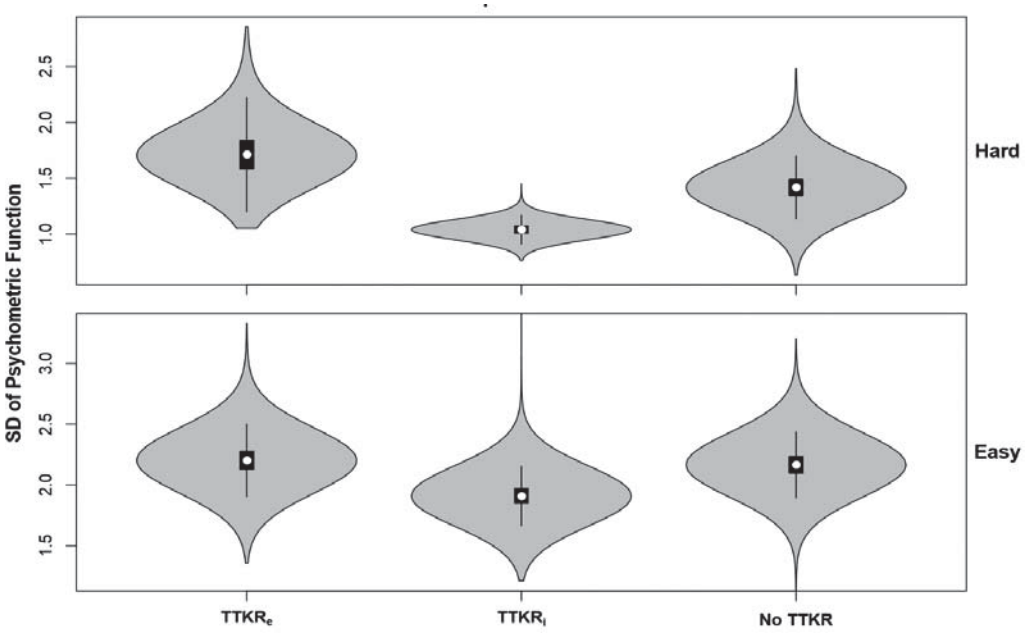

Figure 5.1: Transformed population-level posterior distributions for the slope parameter, expressed on the stimulus scale to correspond with the standard deviation of the psychometric function.

We see that the estimated criterion variance for the Condition 1 group was larger (σ = 1.7) than both the Condition 2 (σ = 1.04) and Condition 3 groups (σ = 1.41). Criterion variance was thus greater when $TTKR_e$ was provided compared to both $TTKR_i$ and no TTKR. Conversely, the slope parameters were not too dissimilar across the TTKR groups in the easy condition, though there is an effect for TTKR in the Condition 5 observers. The estimated criterion variance is smaller (σ = 1.9) than for both the Condition 4 (σ = 2.19) and Condition 6 observers (σ = 2.16). This is unexpected and suggests that TTKR is in fact having some effect in the easier tasks. However, the difference in the degree of criterion variance between the groups is huge, which suggests the difference in sensitivity may simply be stochastic, an artefact of the small number of participants per group[14].

5.4.1. Model Comparisons & Discussion

This section draws only upon the comparative evidence between the PGDM and the EWLM at the individual level. There are three comparisons of interest: whether the evidence favours the PGDM over the SCM; whether the evidence favours the EWLM over the SCM; and where the balance of evidence lies between the PGDM and the EWLM. Turning our attention toward the first comparison, Figure 5.2 displays each participant's Bayes factor across conditions. The individual Bayes factors ranged from .59 to 41, with a mean of 4.28 (SE = 1.12). In all but two cases the evidence was more likely under the SCM; however, for nearly half the observers the evidence is not overly compelling for either the SCM or PGDM.

Focussing now on the EWLM, the immediate point to note is that we have a larger proportion of observers whose data are more consistent with the EWLM than was observed for the same comparison with the PGDM (Figure 5.3); however, much like the PGDM the comparative evidence is not very strong. While 38% (13 out of 36) of observers' data were more likely under the EWLM, there was no obvious pattern underlying the observers that

[14] There were significant difficulties in recruiting participants for this experiment which meant that the sample size per condition was much smaller than anticipated, N = 6.

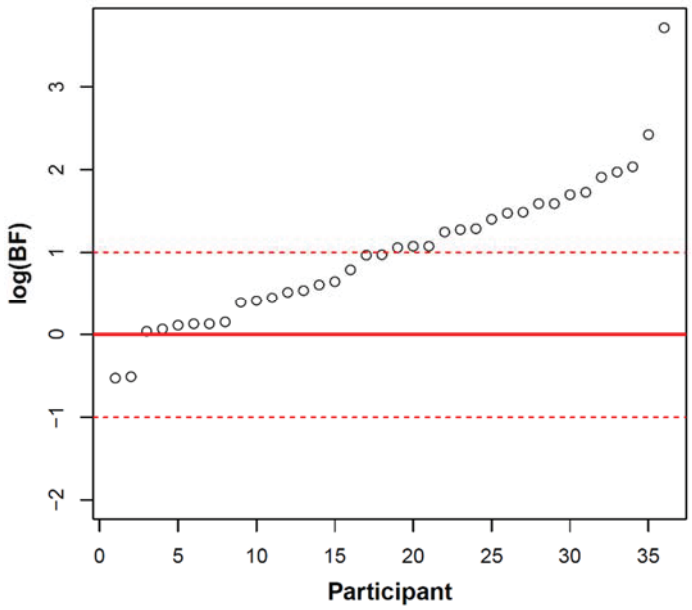

Figure 5.2: Individual level log Bayes Factors as evidence for the PGDM across all experimental conditions. The values are ordered without regard for participant group because in nearly all cases the evidence favours the SCM. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

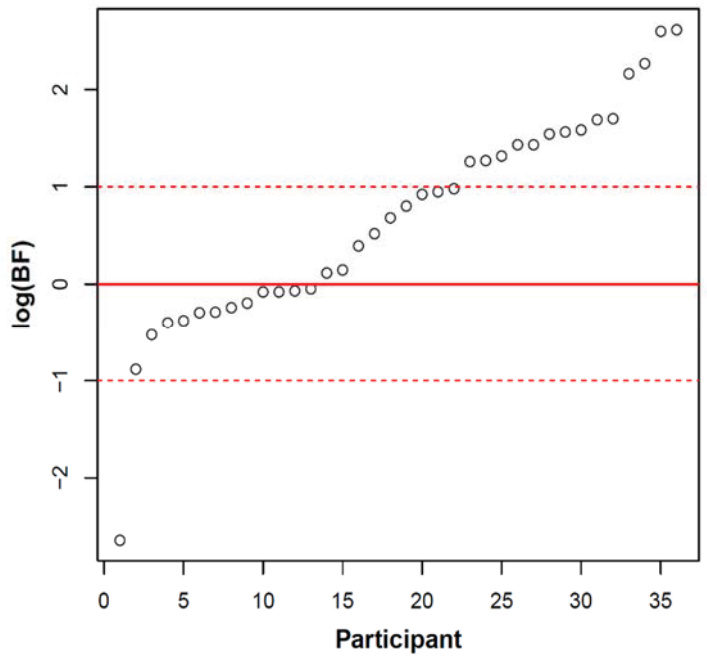

Figure 5.3: Individual level log Bayes Factors as evidence for the EWLM across all experimental conditions. The values are ordered without regard for participant group because in nearly all cases the evidence favours the SCM. The area between the dotted lines reflects values that provide no real evidence in favour of either model.

were more consistent with the EWLM. The individual Bayes factors ranged from .07 to 13, with a mean of 3.24 (SE = .57).

Finally, the PGDM was compared with the EWLM (Figure 5.4). The range of the individual Bayes factors extends from .01 to 12, with a mean observer Bayes factor of 2 (SE = .42). It can be seen that approximately 50% of observers are more consistent with the EWLM. When this is broken down by group, no observers from Group 1 were consistent with the EWLM. Conversely, the data from all but one observer from Groups 2 and 3 were more consistent with the EWLM than the PGDM. In the easy condition, only one observer had data that were consistent with the EWLM in Group 4, whereas 58% (7 out of 12) observers from Groups 5 and 6 produced data that were consistent with the EWLM. The evidence for the superiority of the EWLM over the PGDM is fairly mixed. It can be seen that the EWLM does a poor job of predicting the data where $TTKR_e$ is provided to observers.

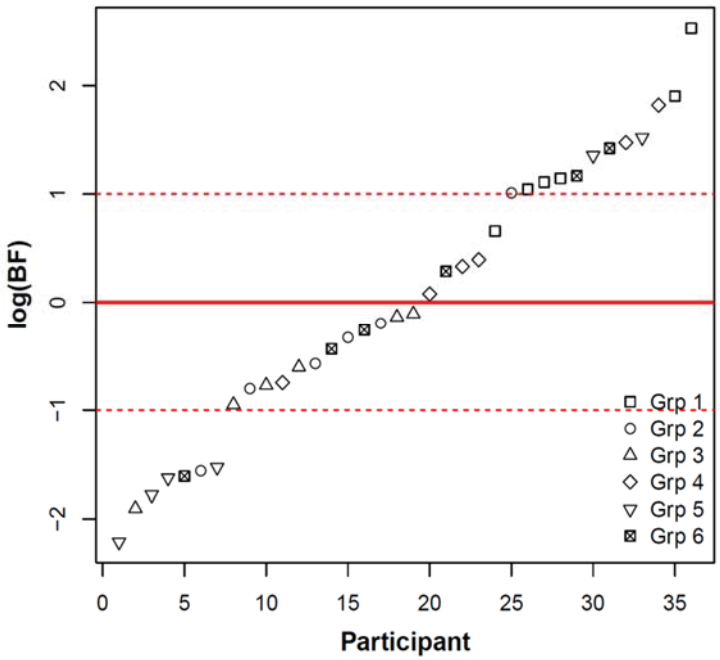

Figure 5.4: Individual level log Bayes Factors across all experimental conditions. The area above zero is the evidential strength for the PGDM whereas the area below is the evidential strength for the EWLM. Observer groups have been denoted in the plot. The area between the dotted lines reflect values that provide no real evidence in favour of either model. There is almost a 50/50 split in the observers falling above or below the zero point.

This is perhaps not a surprising result given that the EWLM assumes that the criterion must remain bounded by the distributional means, limiting the degree to which the criterion can shift. The PGDM, however, is not limited and it appears to better predict the Group 1 and 4 data. However, there is a clear preference for the EWLM in conditions where criterion variability has been shown to be reduced by either providing $TTKR_i$ or no TTKR. Here the EWLM is better able to predict the data than the PGDM, for the most part. This finding may be due to one of two factors; the first is that shifts in both directions following correct decisions are occurring and the EWLM gains merit in its ability to predict these shifts; the second is that the bounded nature of the criterion does very little for the model and it is effectively behaving like the original ILM; the criterion is, for all intents and purposes, fixed. Stated another way, while there may be a benefit for being able to shift the criterion in either direction following a correct decision, it is unclear the extent to which the EWLM is predicting these shifts. Alternatively, it may be that the best fitting decay parameters virtually render the criterion fixed.

Some insight may be gleaned when the actual decay parameter estimates are considered. The estimated decay values across all groups ranged fairly uniformly between .88 and .95. These values are all very high which suggests the information loss is occurring very slowly; past items are being retained over long periods of time. Bearing in mind that a decay value of one mimics the ILM, these decay values are suggestive of a criterion that exhibits relatively small shifts. If so, the EWLM would not make predictions that differed greatly from the static criterion model. Thus, it would not be surprising to see a preference for the EWLM, given all previous model comparisons have tended to favour the SCM. However, this conclusion makes the inconsistencies between the EWLM and the $TTKR_e$ observers unclear. Based on the reasoning above, it should be the case that the observer data in Groups 1 and 4 should also be better predicted by the EWLM, but they are not. It could instead be implied that the criterion *was* shifting enough to make varying trial-by-trial predictions, though the predictions were inconsistent with the data because a) the criterion could not shift widely enough, given that it is constrained between the distributional means, or b) that the bi-

directional shifts add no predictive utility and actually result in poorer predictions. Unfortunately, these results do not lend themselves to a straightforward, interpretation. So, while the EWLM was able to qualitatively model the effects of TTKR (though not the effects of no TTKR), it appears that it does not provide a substantive model for human categorisation.

In summarising these analyses, the foregoing has detailed a replication study that hoped to better elucidate the effects of TTKR. Specifically, all experimental cells were completely independent and the experimental stimuli were further evaluated to ensure adequate discriminability, remediating some earlier issues with experimental design and implementation. While the study suffered from very small participant numbers, the overall results tended to substantiate the earlier findings reported in Chapters 2 and 3. The main purpose for collecting these experimental data was to enable comparisons to be made in regard to the predictions of the PGDM and the EWLM. Both models share a common feature that relies on a type of averaging process that determines the direction and magnitude of criterion shifts, though the specific mechanism through which this is achieved differs slightly between the models. It is precisely these differences in mechanism that allow for a more generalised shift process in the EWLM. While the EWLM provides a very neat qualitative description of the effects of TTKR (Chapter 4), the ability of the model to capture the full range of data is clearly limited. This suggests that the extra shift the EWLM introduces may not be of any great use. Moreover, the variability in the relative consistency between the EWLM predictions and observer data could not be interpreted in any clear way.

5.5. General Discussion

Signal detection theory may be used as both a measurement and a process model (Turner et al., 2011). In the former, we may make inferences regarding one's ability to categorise or detect stimuli according to some perceptual characteristics that are completely independent from any biases one has in responding. The benefits that this approach engenders cannot be understated, an approach that fundamentally reshaped the field of psychophysics and being applicable to many areas of theoretical and applied psychology. As a

process model, the idea that an observer weighs the likelihood, or perceptual magnitude, of the stimulus against a criterion is remarkably intuitive. However, the process account of signal detection breaks down when we assume that the criterion is fixed across trials. The chief concern is an increased level of uncertainty in the SDT parameter estimates. This is so because a shifting criterion introduces noise that cannot be accounted for by the SDT model, meaning the parameter estimates one obtains may underestimate the true categorisation abilities of the observer. However, in the absence of any acceptable way to effectively deal with, or remove, the variance associated with the criterion, we are left to assume that the parameter values obtained accurately reflect observer performance. Furthermore, in most cases it is not possible to define an ideal observer against which human observers can be compared. The presence of unmodeled variability is pernicious and significantly affects the inferential validity of performance measures. Moreover, the work detailed here also raises the concern that various factors thought to be largely immaterial to the overall performance of observers, may actually influence performance in ways that are also deleterious. Feedback is a prime example.

This thesis has focussed on two aspects. The first concerns the effect that providing trial-by-trial feedback to observers has upon decision noise. A typical obstacle in determining the relative magnitude of decision noise in detection tasks is that decision noise is confounded with sensory noise; this is to say, the two usually cannot be separated. In an attempt to circumvent this problem, a methodological design was implemented which involved distributing stimuli externally. The idea was to reduce the detection task, in which performance is contingent upon internally distributed stimulus effects, to a categorisation task where the stimulus effect is observable to both the experimenter and observer. The motivation behind the design was to make the decision stage as explicit, and observable, as possible and remove confounding factors associated with the perceptual stage. This approach provides the best possible method for studying human decision behaviour under conditions analogous to those experienced in detection tasks. As was discussed in Chapter 1, this method has been widely used as a means of observing decision behaviour. To this end, analyses were

undertaken through which the effects of TTKR might become more apparent. A key aspect of these analyses was to examine the observer psychometric, or choice, functions. Of most interest were the comparative slopes across TTKR groups, and when these were considered it was apparent that the type of TTKR that is provided influenced the way that observers distribute their responses. Providing information related to the statistical properties of the distributions (i.e., $TTKR_e$), in conditions where the true state of nature is very uncertain (the "hard" condition), results in poor categorisation performance. Performance was compared with a group where, instead of providing $TTKR_e$, feedback was deterministically related to the optimal decision criterion. So, rather than providing individuals with the actual distribution the stimulus had been sampled from, the observer is instead informed of what the optimal response for that stimulus should have been. Providing feedback in this way improved performance, where observers were performing almost at asymptote in the hard categorisation condition.

Conversely, the type of TTKR was expected to have no differential effect in conditions where the true event is less ambiguous; that is, in conditions where the stimulus distributions overlap less (the "easy" condition). The underlying reason that drives the effect of TTKR is that, where there is greater distributional overlap, stimuli that present the same amount of evidential information may have been drawn from different distributions on different occasions. When this information is relayed to the observer the feedback may seem inconsistent, which in turn precipitates often unnecessary changes in response, and this is precisely what $TTKR_i$ circumvents because it is determined relative to a fixed value, and thereby inherently more consistent. In the case where the stimulus distributions overlap less, confusion that arises from stimuli being drawn from the overlapping regions is mitigated, thus markedly reducing the benefit of TTKR. The effects reported here were inconsistent with respect to the effects of TTKR in easy tasks. On the whole, it appeared that $TTKR_e$ and $TTKR_i$ had no differential effect, as expected. The outcome where no TTKR was provided is less clear. There is evidence to suggest that providing TTKR or not may influence performance in psychophysical tasks (e.g., Carterette et al., 1966). In addition, there has been some interest

within the recognition memory field concerning how TTKR affects performance, with some evidence suggesting that the effects described here are present in recognition memory. For example, Katner and Lindsay (2010) reported drops in performance where TTKR is provided compared to where it is not. In the absence of any systematic psychophysical investigation this suggests that the effects may hold under orthodox detection conditions. Schoeffler (1965) made no predictions regarding the role that task difficulty played in the effect of TTKR, citing the Cartertte et al. (1966) study as evidence for a putative effect. Lee and Zentall (1966) established that the effect was present, but only for hard tasks. The evidence reported in the present research tends to confirm this position, though does not shed conclusive light upon the effects in easier tasks. The results from the present chapter's analysis are equivocal given the small sample size; however, the effects reported in Chapters 2 and 3 appear to be a little more robust, and suggest that there is little difference in performance when either type of TTKR is provided (Chapters 2 and 3), and that providing no TTKR negatively affects performances (Chapter 3). However, more definitive studies and modeling will have to be conducted in the future to further elucidate the interplay, if any, between task difficulty and TTKR type.

A fairly consistent finding in the present study was that performance in the easy task was sub-optimal. Additionally, the estimates of criterion variance suggested that the criterion was certainly fluctuating, where in some cases the amount of variation was greater than that observed in the hard task. It is not immediately obvious why performance was so sub-optimal given the relatively easy task demands, though performance was equivalently sub-optimal across the TTKR groups. There is one feature of the methodological design that presents as a possible reason for the poorer performance in the easy tasks. The number of stimuli used in the easy task was greater than the number used for the hard task. This happens due to the rightward shift of the "high" distribution. Ultimately, more stimuli are required to facilitate this shift, which in turn means that observers in the easy condition are presented with more stimuli. While on an intuitive level this should not pose an issue, there may be an argument that an increased cognitive load is pushing performance down. In this sense, the tasks were

not strictly equivalent. Any future extension of this work needs to consider this possibility. Fortunately, there is a straight forward solution as it is possible to keep the same number of stimuli (and thereby range) of tones constant across levels of difficulty by simply reducing the standard deviation of the distributions in the easy task. This approach could lead to a clearer interpretation of performance in easier detection tasks.

What is clear is that the effect of decision noise in categorisations tasks is not trivial. Furthermore, if we accept the distribution of external stimuli as a valid analogue of the detection task, then the present work demonstrates just how much decision noise can exist in binary categorisation tasks. The implications for SDT have already been stated in Chapter 1, though it is important to understand that the existence of decision noise needs acknowledgement and that decision noise may be affected by the type of information available within the observer's environment. The present results, however, do not imply that feedback should not be provided. Feedback is essential in learning and skill acquisition. For example, in training situations providing feedback to radiographers is essential in establishing an ability to discriminate between cancerous growths and broken bones. Rather, alternatives in the way TTKR is provided may yield faster, and better, results than the typical trial-by-trial methods. In cases where ideal observer performance may be calculated, such as in categorisation tasks, providing deterministic feedback can improve overall performance. Providing TTKR after blocks of trials rather than trial-by-trial may also improve performance (Podd, 1975). It may be the case that observers receiving feedback on every trial are simply being overwhelmed with information and so are constantly trying to adjust their criterion to suit the incoming information. Providing performance updates across blocks may mitigate this tendency and might improve performance. These are certainly worthwhile avenues for future investigation. What we also need are ways of describing the mechanism underlying response inconsistencies. That is, establishing that a difference exists between two groups should not be the overarching goal for psychology, but rather to understand the cognitive processes that drive these differences. This, then, promotes the idea of a process-driven approach to psychology. The natural consequence of this approach is to fit, modify, or

contrive, models that can go some way to explaining the data. This was the second focus of the present thesis.

Theories as to why criterion fluctuation occurs largely centre on explanations that evoke strategic motivations (e.g., error-correction) or memory failures. Prevailing theories suggest that the criterion is shifted in ways that reduce future errors, and there is good empirical evidence that such an interpretation has some explanatory value (e.g., Dorfman & Biderman, 1971; Kac, 1962). Alternative theories suggest instead that the criterion on each trial is a random sample from a criterion distribution (e.g., Dorfman et al., 1975; Muller & Weidemann, 2009), though the underlying assumption of such models, namely, that the criterion is a random rather than a fixed variable, lacks support. While many of the models can describe trial-by-trial choice data reasonably well, many fail to assess how stimulus effects influence the criterion in terms of how observers build up representations of the criterion location across trials (cf. Turner et al., 2011), or how this criterion is established or maintained. A further issue in measuring criterion variance, addressed in the present work, was the inherent confounding between decision and perceptual noise. Through using externally distributed stimuli the present work attempted to gauge the appropriateness of various dynamic criterion models.

The approach taken here in comparing models relied upon the relative evidence of each model. This is encapsulated by the marginal likelihood of the model and offers a single numerical value that determines the model's predictive ability. In terms of model comparisons, the ratio of marginal likelihoods, or Bayes factor, provides an unambiguous metric in determining the evidence in favour of a particular model, circumventing many of the inherent problems in conventional model fitting (likelihood ratio tests, information criteria) that can often be biased. Good models should make a large number of good predictions that are both sensible and constrained. Often this is achieved by constraining the priors placed upon the model, usually reflecting some *a priori* sense of where the parameter values should fall. However, models must not be overly constrained such that they make very few predictions; there needs to be good trade-off between constraint and flexibility. The

model comparisons undertaken in the present study attempted to incorporate and make use of the marginal likelihood to better inform the model selection process. The comparisons undertaken in this chapter, and Chapter 3, made use of constrained priors so that the predictions made by each model were reasonable. Parameter estimation was not undertaken as interest focussed solely upon whether the EWLM could better account for the observed data than the PGDM. For this simple task all that was required were the marginal likelihoods for each model.

The general findings from the model comparisons were that the dynamic criterion models can fit the data; this was true of both the restricted and general models. However, the overall fit and predictive ability of the models was greatly enhanced when stimulus magnitude was included within the model architecture. Specifically, the data were better predicted when criterion shifts were modeled as a proportional shift of the distance between the current stimulus and criterion value. One of the implications of modeling the criterion in this way is that when the shift matrix is removed, allowing the criterion to shift on all trials, the shifts following correct decisions are in the direction opposite to those often cited in the literature (e.g., Dorfman & Biderman, 1971; Dorfman et al., 1975). The criterion tended toward the most current stimulus value in a fashion more consistent with a weighted average (e.g., Dyjas et al., 2012; Laming, 2014; Thomas, 1975). This behaviour naturally emerges when the difference between the criterion and the stimulus value is considered (e.g., Thomas, 1975) and so offers a natural psychological interpretation. After all, the difference between the stimulus and the criterion is precisely what guides our decision, so why cannot it affect the shift in the criterion? The theoretical consequence that emerges from this conjecture is that it is at odds with other psychologically established phenomena; namely, the tendency for observers to repeat the previously correct category (assimilation). A criterion shifting toward the stimulus value following correct decisions, however, would promote the idea of a contrast effect. It had been reasonably assumed that the criterion should shift away from the current stimulus value following correct decisions, because it increases the area within which the stimulus lies providing a process through which assimilation may come about. A criterion

that shifts toward the current stimulus may pose a theoretical issue, though for now it remains uncertain how much of an issue it may be. Given the fact the PGDM was able to make reasonable predictions would suggest that there is empirical validity to such shifts. This said, different tasks may affect the criterion in different ways. The inferences drawn about the criterion given the tasks described in the present study may be very different if other tasks had have been used. Further investigation is required in further elucidating the implications of contrastive and assimilative criterion shifts.

A persistent finding across all the model comparisons was that, while the PGDM was able to make predictions that were consistent with the observer data, when the predictions made by the PGDM were compared to the static criterion, or conventional, signal detection model, the evidence mostly favoured the fixed criterion account. This was a counterintuitive finding in light of the clear differences in the slopes of the psychometric functions across TTKR groups. However, the Bayes factors did not suggest that the evidence for the SCM was overly strong. That is, while the evidence leaned toward the SCM, would there be enough evidence to rule out the PGDM? In many cases the evidence was indifferent to both models and so no clear preference was possible; both models were equally as probable. However, one might argue that, given how agnostic the Bayes factors were, there would not be enough evidence to persuade one to consider the PGDM as a viable model; and this would be fair a conclusion. Does this mean the criterion is not shifting, though? Absolutely not; the working hypothesis was that the model was perhaps not specified correctly and that a better consistency between data and model might be gained by explicitly modeling the effects of TTKR.

A further goal in examining the cognitive processes underlying criterion shifts was to try and model both the effects of TTKR and the way that information might be integrated across trials. This effort resulted in the EWLM, the only model known that has explicitly modeled TTKR effects. The EWLM shares in common with weighted average models (e.g., the PECM) the idea that stimulus information influences the maintenance of the decision criterion, and that shift magnitudes are not constant across trials. This is unlike the

assumptions governing the conventional dynamic criterion models approach whereby the criterion simply shifts by a fixed amount according to deterministic rules. A major criticism levelled at the ALM framework models is that they are divorced from informational content and the value of the stimulus itself. However, this is not an entirely fair criticism as the information typically available in conventional detection tasks does not include the stimulus effect. Therefore, the models had to make reasonable assumptions about the ways the criterion might shift, and to this end they were fairly adequate. The advantage in constructing the models used in the present investigation is that such information is available and so might be used to explore ways in which the basic mechanism can be improved. Accordingly, the EWLM essentially builds up a representation of the distributions (via the mean values), using a limited number of items per category determined by the inherent limits upon memory. Observers are then assumed to locate their criterion at the midpoint of these two distributions. A further difference is that the EWLM generalizes shifts further by allowing shifts in all directions following correct decisions. Another advantage was that the influence of TTKR could be explicitly modeled such that the EWLM could perform very differently when different types of TTKR were used. The only other model that had suggested such effects was Schoeffler's (1965) learning model. The EWLM, from a qualitative perspective, was able to capture the pattern of results between the $TTKR_e$ and $TTKR_i$ observers, though the predictions it made for the no TTKR observers were at odds with the data. This was so for the hard condition at least. However, the predictions made by the EWLM are not so different from those made by Schoeffler, and nor should they be. In essence, with the EWLM decay rate set to zero, Schoeffler's model emerges and is thus a special case of the EWLM. Additionally, the EWLM, much like Schoeffler's model, predicts an amount of decision noise that is ostensibly much less than is present within the observed data. Even with the EWLM fully degraded ($\lambda = 0$) the EWLM predicts asymptotic categorisation performance that far exceeds the empirical data. Simulations revealed that the degraded EWLM predicts a criterion distribution with a standard deviation of less than unity which implies a fairly truncated distribution of values. In order to attain performance estimates that are similar to the

empirical data one would require a distribution of criterion shifts with a standard deviation of at least unity, on the average.

Some of the failings in the predictive ability of the EWLM can be sourced in the constraint placed upon the criterion, which sees it bound between the estimated means of the stimulus distributions. It conditions where the criterion is assumed to shift less, the EWLM appeared to be able to predict the data moderately well, though it often did no better than the PGDM. This implies that the existence of the additional criterion shift offers little benefit. An additional issue with the EWLM is that it cannot be extended beyond the equal base rate case. This is a critical failing in any model that is to be applied to detection-type tasks because the criterion must be sensitive to changes in the frequency with which events occur; an issue that has largely been circumvented by Turner et al. (2011). The EWLM shares many similarities with Turner et al's. non-parametric dynamic SDT model and both are predicated upon the trial-by-trial integration of stimulus information. The dynamic SDT model, however, uses a likelihood ratio criterion which allows it to be generalised to cases other than the equal base rate case. Given that the basic architecture of the models is the same it would seem likely that the effects of TTKR could be modeled using the dynamic SDT model also. In future modeling efforts it would be interesting to see whether a parametric version of the model could be instantiated and fit to categorisation data. This said, based upon the findings reported here it would seem that the criterion may be best described by a weighted average of the incoming stimulus information (i.e., the PGDM), though this requires further investigation. Additionally, it seems that only a single shift after a correct decision may be necessary, though this also needs further investigation. While in some cases observer data were more likely under the EWLM, it is quite apparent that the PGDM has greater generalisability, and thereby psychological utility. If, however, the EWLM was the true data generating model, then in order for it to account for the empirical data one may speculate that there are additional, possibly idiosyncratic, processes that the EWLM has simply missed. This is not an unreasonable assumption, though it is unlikely. Moreover, given the aforementioned constraints placed upon the magnitude of criterion shifts, one would have to conclude that

the EWLM either requires some major modifications, or is not sufficiently generalizable. My own sentiments rest with the latter.

Future research will also have to address whether the reported TTKR effects are applicable beyond the equal base rate case. Evidence from the criterion learning literature (e.g., Maddox, 2002) suggests that an effect of TTKR probably would. Maddox and colleagues (2002, 2004, 2005) have established that providing optimal classifier feedback (a type of idealised feedback) can eliminate competition between decision criteria. For example, in cases where both the base rate and pay off for events are asymmetric it is possible for two decision criteria to exist along the decision axis: one that maximizes accuracy and one that maximizes reward. A robust finding under such conditions is that observers tend not to favour either criterion, but rather place their decision criterion between the two (cf. Healy and Kubovy, 1977a); that is, there is competition between maximizing reward and accuracy. Maddox and colleagues demonstrated that when feedback is provided relative to one of the criteria (e.g., the reward criterion), observers shift their criterion toward the reinforced criterion, thus eliminating the competition between the criteria. Proposed extensions to the current work would investigate whether the differential effects of feedback would hold where one event was more common than the other.

The present work dealt only with choice data and so was limited in the number of models that could be evaluated. Reaction time data would have been useful so that modeling via the diffusion model (e.g., Ratcliff, 1978, 2002), or the Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2005, 2008), might have been used to provide alternative means through which we could assess the effects of TTKR. It would have been interesting to evaluate how decision thresholds differed across TTKR groups. Drift rates generally reflect some inherent quality of the stimulus and information available, though given that the stimulus information the same across all conditions one would expect to see differences in reaction times due to changes in thresholds across TTKR groups. For example, it might be reasonable to assume that faster reaction times would be evident for the $TTKR_i$ group because there decision boundaries are set much lower whereas $TTKR_e$ observers may set theirs higher. This might

reflect a tendency for $TTKR_i$ observers to be more confident in their decision owing to the consistency of the trial-by-trial feedback. $TTKR_e$ observers, on the other hand, may be less sure of their decisions and take longer. Of course, this is mere speculation; an empirical investigation to pursue this possibility in the future would be of interest.

The current findings may also be generalised to many situations where classification or categorisation of stimuli are required. It is invariably the case that models of perception find an appropriate place within the fields of cognitive and experimental psychology (e.g., Malmberg & Annis, 2012). Relatively early on, signal detection was applied in the analysis of memory, providing a tool with which people's ability to recognize previously seen items could be analysed in terms of hits and false alarms (e.g., Wickelgren, 1968; MacMillan & Creelman, 2005). More recently, the ideas of criterion variance (e.g., Benjamin et al., 2009) and sequential dependencies (e.g., Jones et al., 2006) are receiving extensive attention within the field. Probabilistic and perceptual categorisation play a fundamental role in our daily cognitive experiences, and accordingly have both received a lot of research attention (e.g., Ashby & Gott, 1988; Maddox, 2002; Nosofsky, 1986; Rouder & Ratcliff, 2004, 2006; Stewart et al., 2002, 2005). In what is an analogous state of affairs between probabilistic (e.g., Schoeffler, 1965) and deterministic response processes in detection (e.g., Green & Swets, 1966) much of the modeling work in cognition has been spent investigating differences between exemplar (e.g., the GCM; Nosofsky, 1986) and threshold -based accounts (e.g., GRT; Ashby & Gott, 1988) of categorisation. The debate between exemplar and decision-bound theories still is not resolved; a complication is the fact that probabilistic processes can be mimicked by a threshold-process if the decision bound is allowed to vary. The early probabilistic response models are models that assume a type of stimulus generalisation (cf. Shepard, 1957, 1987) whereby the similarity of the current stimulus is weighted against category exemplars. Responses are then determined by gauging how similar the current stimulus is to either category exemplar. The more dissimilar the current stimulus is to a specific exemplar, the higher the probability that the stimulus will be assigned to the opposing category. Nosofsky's (1986) GCM is perhaps the most well-known of these models. Another of similar ilk is Stewart

et al.'s (2005) memory and contrast model (MAC). Like the GCM, the MAC compares the psychological distance between the current stimulus and some exemplar. In restricted cases the exemplar may be the immediately preceding trial. In such instances the MAC can perform much like a threshold model where the immediately preceding stimulus value *is* the decision criterion. What this suggests is that many decisions are made in relation to other stimuli and so are in no way absolute.

With respect to threshold-based models, the general recognition theory model (GRT) is the most favoured. Though is it generally applied to multi-dimensional categorisation tasks, it can also be applied to uni-dimensional tasks like those in the current study. In this case the model fits a psychometric function to the response contingencies for each stimulus value. Interestingly, in doing so one estimates a mean threshold value and a variance parameter, though the variance parameter is routinely viewed as being solely comprised of stimulus noise. It is, however, likely that decision noise makes up a large part of the variance in responding, and in cases where the stimuli are highly discriminable decision noise probably constitutes a very large proportion of the variance. Such considerations are usually not made.

Given the nature of the task used presently, it may also be prudent to question whether the observer used an exemplar-type approach in arriving at decisions on each trial. While exemplar models have been shown to be of value in probabilistic categorisation tasks where the number of items is relatively small, it seems unlikely that such models would work well for the present tasks, given the large number of items per category. Rouder and Ratcliff (2004, 2006) suggest that as the number of simple stimuli increase, behaviour is better accounted for by a decision rule process. While this may be so, it does not completely preclude relativistic strategies like those suggested by Stewart et al. (2005; i.e., the MAC). The PECM incorporates such a process as a special case where $\delta = 1$, where the total proportion between the current criterion and the current stimulus is used. In effect, the criterion becomes the current stimulus value, and so is a type of MAC model. From the present data it appears that MAC-type models may not work well given that none of the parameters' values approached zero; however, further model fitting would have to be undertaken to firmly

establish this. It may be the case that some observers use such a strategy. It would be of interest to find out if an exemplar model could provide a better fit, and under what circumstances.

In drawing this present thesis to a close, trial-by-trial shifts in the observer decision criterion places considerable limits upon the interpretability of SDT parameter estimates. The SDT model lacks provision for such extraneous noise; instead, human observers are considered to be consistent decision makers holding their decision criterion fixed. Factors within the observer's environment can also influence the way in which the criterion shifts across time. Paradoxically, while feedback is usually provided to help the observer perform detection or categorisation tasks, it may also negatively influence performance. This appears to be the case in environments where the uncertainty surrounding the stimulus events is relatively high. The models presented herein specify a mechanism through which the TTKR drives the choice behaviour according to the trial-by-trial flow of stimuli. Despite a number of limitations, the results suggest that the stimulus magnitude play a critical role in the way the observer established and maintains the decision criterion. Rather than shifting the criterion according to rules determined by the outcomes of each trial, the criterion instead shifts in response to the accumulation of information across trials, which is used to establish a long-term representation that shifts much like a weighted moving average. The need to account for and model such processes is paramount if the signal detection model is to be used in any way other than as a descriptive tool. If signal detection is being used as a process model, then variability in the criterion must be included into the model.

References

Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical framework. *Journal of General Psychology, 54,* 279-299.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33-53.

Ashby, F. G., & Lee, W. W. (1993). Perceptual variability as a fundamental axiom of perceptual science. In S. C. Masin (Eds.), *Foundation of perceptual theory* (pp. 369-399). Elsevier Science Publishers B. V.

Atkinson, R. C., Carterette, E. C., & Kinchla, R. A. (1964). The effects of information feedback upon psychophysical judgements. *Psychonomic Science, 1,* 83-84.

Balcazar, E., Hopkins, B. L., & Suarez, Y. (1985). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management*, 7, 65-89.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection and criterion noise: Applications to recognition memory. *Psychological Review, 116*, 84-115.

Blackwell, J. R., & Newell. K. M. (1996). The informational role of knowledge of results in motor learning. *Acta Psychologica, 92,* 119-129.

Bohil, C. J., & Maddox, W. T. (2003a). A test of the optimal classifier's independence assumption in perceptual classification. *Perception & Psychophysics, 65,* 478-493.

Bohil, C. J., & Maddox, W. T. (2003b). On the generality of optimal versus objective classifier feedback effects on decision criterion learning in perceptual categorisation. *Memory & Cognition, 31,* 181-198.

Bonnel, A. M., & Miller, J. (1994). Attentional effects on concurrent psychophysical discriminations: Investigations of a sample-size model. *Perception & Psychophysics, 55,* 162-179.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112,* 117–128.

Brown, S., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology, 57,* 153–178.

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115,* 396 – 425. doi:10.1037/0033-295X.115.2.396

Bush, R. R., Luce, R. D., & Rose, R. M. (1964). Learning models for psychophysics. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 201-217). Stanford: Stanford University Press.

Carterette, E. C., & Wyman, M. J. (1962). Application of a Markov learning model to simple detection situation involving social pressure. In J Crisell, H . Soloman, and P. Suppes (Eds.), *Mathematical methods in small group processes* (pp. 74-100). Stanford: Stanford University Press.

Carterette, E. C., Friedman, M. P., & Wyman, M. J. (1966). Feedback and psychophysical variables in signal detection. *Journal of the Acoustical Society of America, 39*, 1051-1055.

Chinn, R. M. C., & Alluisi, E. A. (1964). Effects of three kind of knowledge of results information on three measures of vigilance performance. *Perceptual & Motor Skills, 18,* 901-912.

Chun, M. M., & Wolfe, J. M. (1996). Just say no: How are visual searches terminated when there is no target present? *Cognitive Psychology, 30,* 39–78. doi:10.1006/cogp.1996.0002

Collier, G., & Verplanck, W. S. (1958). Nonindependence of successive responses at threshold as a function of interpolated stimuli. *Journal of Experimental Psychology, 55*, 429-437. doi:10.1037/h0047574

DeCarlo, L. T. (1998). Signal detection and generalised linear models. *Psychological Methods, 3,* 186-205.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54,* 304-313.

DeCarlo, L. T., & Cross, D. V. (1990). Sequence effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General, 119,* 375-396.

Dorfman, D. D. (1973). The likelihood function of additive learning models: Sufficient conditions for strict log-concavity and uniqueness of maximum. *Journal of Mathematical Psychology, 10,* 73-85.

Dorfman, D. D. (1977). Comments on "The decision rule in probabilistic categorisation: What it is and how it is learned," by Kubovy and Healy. *Journal of Experimental Psychology: General, 106,* 447-449.

Dorfman, D. D., & Biderman, M. (1971). A learning model for a continuum of sensory states. *Journal of Mathematical Psychology, 8,* 264-284.

Dorfman, D. D., Saslow, C. F., & Simpson, J. C. (1975). Learning models for a continuum of sensory states re-examined. *Journal of Mathematical Psychology, 12,* 178-211.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America, 46,* 372-383.

Dusoir, A. E. (1980). Some evidence on additive learning models. *Perception & Psychophysics, 27*, 163-175. doi: 0031-5117/80/020163·13

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics, 74,* 1819-1841.

Eijkman, E., & Vendrik, A. J. H. (1964). Detection theory applied to the absolute sensitivity of sensory systems. In J. A. Swets, *Signal detection and recognition by human observers.* New York: Wiley.

Ell. S., Ing. A. D., & Maddox, W. T. (2009). Criterial noise effects on rule-based category learning: The impact of delayed feedback. *Attention, Perception, & Psychophysics, 71,* 1263-1275. doi: 10.3758/APP.71.6.1263.

Embrey, D. E. (1975). Training the inspector's sensitivity and response bias. In C.G. Drury and J.G. Fox (Eds.), *Human reliability in quality control* (pp. 123-131). Taylor and Francis: London.

Erev, I. (1998). Signal detection by human observers: A cut-off reinforcement learning model of categorisation decision under uncertainty. *Psychological Review, 105,* 280-298.

Estes, W. K., & Johns, M. D. (1958). Probability learning with ambiguity in the reinforcing stimulus. *The American Journal of Psychology, 71,* 219-228.

Fechner, G. T. (1860). *Elemente der psychophysik*. Breitkopf and Hartel: Leipzig.

Friedman, M. P., Carterette, E. C., Nakatani, L., & Ahumada, A. (1968). Comparisons of some learning models for response bias in signal detection. *Perception & Psychophysics, 3*, 5-11.

Glass, B. D., Maddox, W. T., & Markman, A. B. (2011). Regulatory fit effects on stimulus identification. *Attention, Perception, & Psychophysics, 73,* 927-937.

Green, D. M., & Luce, R. D. (1974). Counting and timing mechanisms in auditory discrimination and reaction time. In D. H. Krantz, R. O. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary development in mathematical psychology* (pp. 372-415). San Francisco: Freeman.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Green, D. M., McKey, M. J., & Licklider, J. C. R. (1959). Detection of a pulsed sinusoid in noise as a function of frequency. *Journal of the Acoustical Society of America, 31,* 1446-1452.

Gundy, R. F. (1961). Auditory detection of an unspecified signal. *Journal of the Acoustical Society of America, 33,* 1008-1012.

Hammerton, M. (1970). An investigation into changes in decision criteria and other details of a decision-making task. *Psychonomic Science, 21,* 203-204.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog to signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 344-354.

Helson, H. (1947). Adaption-level as a frame of reference for prediction of psychophysical data. *The American Journal of Psychology, 60,* 1-29.

Helson, H. (1964). *Adaption-level theory.* New York: Harper & Row.

Holland, M. K., & Lockhead, G. R. (1968). Sequential effects in absolute judgments of loudness. *Perception & Psychophysics, 3,* 409–414. doi:10.3758/BF03205747

Howarth, C. I., & Bulmer, M. G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology, 8*, 163-171. doi:10.1080/17470215608416816

Jeffreys, H. (1961). *Theory of probability* (3rd. ed). Oxford, UK: Oxford University Press.

Jones, M., Love, B. C., & Maddox, W. T. (2006). Recency effects as a window to generalisation: Separating decisional and perceptual sequential effects in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 316-332. doi:10.1037/0278-7393.32.3.316

Jones, M., & Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 626 – 640. doi:10.1037/0278-7393.29.4.626

Kac, M. (1962). A note on learning signal detection. *IRE Transactions on Information Theory, 8*, 126-128.

Kac, M. (1969). Some mathematical models in science. *Science, 166,* 695-699.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389-406.

Kary, A., Taylor, R., & Donkin, C. (in press). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284. doi: 0033-2909/96

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York: Springer.

Kubovy, M., & Healy, A. F. (1977a). The decision rule in probabilistic categorisation: What it is and how it is learned. *Journal of Experimental Psychology: General, 106,* 427-446.

Kubovy, M., & Healy, A. F. (1977b). Numerical decision and the ideal learner: A reply to Dorfman. *Journal of Experimental Psychology: General, 106,* 450-452.

Kubovy, M., Rapoport, A., & Tversky, A. (1971). Deterministic vs. probabilistic strategies in detection. *Perception & Psychophysics, 9,* 427-429.

Laming, D. (2014). Signal detection with $d' \equiv 0$: A dynamic model for binary prediction. *Journal of Mathematical Psychology, 60,* 35-46.

Larkin, W. (1971). Response mechanism in detection experiments. *Journal of Experimental Psychology, 91*, 140-153.

Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal setting. *Organizational Behavior and Human Decision Processes*, *50,* 212-247.

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian cognitive modeling.* Cambridge, UK: Cambridge University Press.

Lee, W. (1963). Choosing among confusably distributed stimuli with specified likelihood ratios. *Perceptual and Motor Skills, 16,* 445-467.

Lee, W. (1969). Relationships between Thurstone category scaling and signal detection theory. *Psychological Bulletin, 71*, 101-107.

Lee, W., & Janke, M. (1964). Categorizing externally distributed stimulus samples for three continua. *Journal of Experimental Psychology, 68,* 376-382.

Lee, W., & Janke, M. (1965). Categorizing externally distributed stimulus samples for unequal molar probabilities. *Psychological Reports, 17,* 79-90.

Lee. W., & Zentall, T. R. (1966). Factorial effects in the categorisation of externally distributed stimulus samples. *Perception & Psychophysics, 1,* 120-124.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia, PA: SIAM.

Luce, R. D., Nosofsky, R. M., Green, D. M., & Smith, A. F. (1982). The bow and sequential effects in absolute identification. *Perception & Psychophysics, 32,* 397–408. doi:10.3758/BF03202769

MacMillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* Cambridge, UK: Cambridge University Press.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Analysis of Behavior*, *78,* 567–595.

Maddox, W. T., Baldwin, G. C., & Markman, A. B. (2006). A test of the regulatory fit hypothesis in perceptual classification learning. *Memory & Cognition, 34,* 1377–1397.

Maddox, W. T., & Bohil, C. J. (1998). Base-rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1459–1482.

Maddox, W. T., & Bohil, C. J. (2001). Feedback effects on cost-benefit learning in perceptual categorisation. *Memory & Cognition, 29,* 598-615.

Maddox, W. T., & Bohil, C. J. (2004). Probability matching, accuracy maximisation, and a test of the optimal classifier's independence assumption in perceptual categorisation. *Perception & Psychophysics, 66,* 104-118.

Maddox, W. T., & Bohil, C. J. (2005). Optimal classifier feedback improves cost-benefit but not base-rate criterion learning in perceptual categorisation. *Memory & Cognition, 33,* 303-319.

Maddox, W. T., & Dodd, J. L. (2001). On the relation between base-rate and cost–benefit learning in simulated medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1367–1384.

Maddox, W. T., Markman, A. B., & Baldwin, G. C. (2007). Using classification to understand the motivation–learning interface. In A. B. Markman & B. H. Ross (Eds.), *The psychology of learning and motivation* (Vol. 47, pp. 213–249). San Diego, CA: Academic Press.

Magill, R. A. (1994). The influence of augmented feedback on skill learning depends on characteristics of the skill and the learner. *QUEST, 46,* 314-327.

Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141,* 233-259.

Markman, A. B., Baldwin, G. C., & Maddox, W. T. (2005). The interaction of payoff structure and regulatory focus in classification. *Psychological Science, 16,* 852–855.

Markman, A. B., Maddox, W. T., Worthy, D. A., & Baldwin, G. C. (2007). Using regulatory focus to explore implicit and explicit processing. *Journal of Consciousness Studies, 14,* 132–155.

Massaro, D. W. (1969). The effect of feedback in psychophysical tasks. *Perception & Psychophysics, 6,* 89-91.

McNicol, D. (1972). *A primer of signal detection theory*. Sydney: Allen & Unwin.

McNicol, D. (1975). Feedback as a source of information and as a source of noise in absolute judgements of loudness. *Journal of Experimental Psychology: Human Perception and Performance, 104*, 175-182.

Micalizzi, J., & Goldberg, J. H. (1989). Knowledge of results in visual inspection decisions: Sensitivity or criterion effect? *International Journal of Industrial Ergonomics, 4,* 225–235.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15*, 465-494.

Munson, W. A., & Karlin, J. E. (1956). The measurement of the human channel transmission characteristics. *Journal of the Acoustical Society of America, 26,* 542-553.

Nosofsky, R. M. (1983). Information integration and the identification of stimulus noise and criterial noise in absolute judgement. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 299-309.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorisation relationship. *Journal of Experimental Psychology: General, 115, 39-57.*

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" *A*′ and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10,* 556-569.

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8 -13.

Petersen, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *IRE Transactions on Information Theory, PGIT-4,* 171-212.

Petrov, A. A., & Anderson, J. R. (2005). The dynamics of scaling: A memory-based anchor model of category rating and absolute identification. *Psychological Review, 112,* 383–416. doi:10.1037/0033-295X.112.2.383

Podd, J. V. (1975). *Type I and type II ROC analysis of change in human decision axis.* Unpublished Masters Thesis, Victoria University of Wellington, Wellington, New Zealand.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278–291.

Rosner, B. S., & Kochanski, G. (2009). The law of categorical judgement (corrected) and the interpretation of changes in psychophysical performance. *Psychological Review, 116*, 116-128.

Rouder, J. N., & Ratcliff, R. (2004). Comparing categorisation models. *Journal of Experimental Psychology: General, 133,* 63-82.

Rouder, J. N., & Ratcliff, R. (2006). Comparing exemplar- and rule-based theories of categorisation. *Current Directions in Psychological Science, 15,* 9-13.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical appraisal. *Psychological Bulletin, 95,* 355-386.

Schmidt, R. A., Young, D. E., Swinnen, S. P., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 352-359. doi: 0278-7393/89

Schoeffler, M. S. (1965). Theory for psychophysical learning. *Journal of the Acoustical Society of America, 37,* 1124-1133.

Shepard, R. N. (1957). Stimulus and response generation: A stochastic model relating generalisation to distance in a psychological space. *Psychometrika, 22,* 325-345.

Shepard, R. N. (1987). Toward a universal law of generalisation for psychological science. *Science, 237,* 1317-1323.

Shower, E. G., & Biddulph, R. (1931). Differential pitch sensitivity of the ear. *Journal of the Acoustical Society of America, 3,* 275-287.

Stewart, N., & Brown, G. D. A. (2004). Sequence effects in categorizing tones varying in frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 416 – 430. doi:10.1037/0278-7393.30.2.416

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 3–11. doi:10.1037/0278-7393.28.1.3

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112,* 881–911. doi:10.1037/0033-295X.112.4.881

Swets, J. A., (1961). Detection theory and psychophysics: A review. *Psychometrika, 26,* 49-63.

Swinnen, S. P., Schmidt, R. A., Nicholson, D. E., & Shapiro, D. C. (1990). Information feedback for skill acquisition: Instantaneous knowledge of results degrades learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 706-716. doi: O278-7393/9O

Szalma. J. L., Hancock, P. A., Dember, W. N., & Warm, J. S. (2006). Training for vigilance: The effect of knowledge of results format and dispositional optimism and pessimism on performance and stress. *British Journal of Psychology, 97*, 115-135. doi:10.1348/000712605X62768

Tanner, T. A. Jr., Haller, R. W., & Atkinson, R. C. (1967). Signal recognition as influenced by presentation schedules. *Perception & Psychophysics, 2*, 349–358. doi:10.3758/BF03210070

Tanner, T. A. Jr., Rauk, J. A., & Atkinson, R. C. (1970). Signal recognition as influenced by information feedback. *Journal of Mathematical Psychology, 7,* 259-274.

Tanner, W. P. (1961). Physiological implications of psychophysical data. *Science, 89,* 752-765.

Tanner, W. P., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *Journal of the Acoustical Society of America, 30*, 922-928.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401-409.

Taylor, R. T. (2010). *Criterion variance in signal detection theory: the interactive effect of knowledge of results and task difficulty on binary decision tasks.* Unpublished Masters Thesis, Massey University, Palmerston North, New Zealand.

Thomas, E. A. C. (1973). On a class of additive learning models: Error-correcting and probability matching. *Journal of Mathematical Psychology, 10,* 241-264.

Thomas, E. A. C. (1975). Criterion adjustment and probability matching. *Perception & Psychophysics, 18,* 158-162.

Thomas, E. A. C., & Myers, J. L. (1972). Implication of latency data for threshold and non-threshold models of signal detection. *Journal of Mathematical Psychology, 9,* 253-285.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34,* 273–286. doi:10.1037/h0070288

Triesman, M. (1985). The magical number seven and some other features of category scaling: Properties for a model of absolute judgment. *Journal of Mathematical Psychology, 29,* 175–230. doi:10.1016/0022-2496(85)90015-X

Triesman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*, 68-111. doi:10.1037/0033-295X.91.1.68

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118,* 583-613.

Verde, M. F., MacMillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics, 68,* 643-654.

Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Nonindependence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology, 44*, 273-282. doi:10.1037/h0054948

Wagenaar, W. A. (1973). The effect of fluctuations of response criterion and sensitivity in a signal detection experiment. *Psychol. Forcsh. 36*, 27-37.

Wald, A. (1950). *Statistical decision functions.* New York: Wiley.

Ward, L. M. (1973). Use of Markov-encoded sequential information in numerical signal detection. *Perception & Psychophysics, 14,* 337-342.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics, 9,* 73–78. doi:10.3758/BF03213031

Ward, L. M., Livingston, J. W. Jr., & Li, J. (1988). On probabilistic categorisation: The Markovian observer. *Perception & Psychophysics, 43,* 125-136.

Weidenfeller, E. W., Baker, R. A., & Ware, J. R. (1962). Effects of knowledge of results (true and false) on vigilance performance. *Perceptual & Motor Skills, 14,* 211-215.

Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgements. *Journal of Mathematical Psychology, 5*, 102-122.

Zak, I., Katkov, M., & Sagi, D. (2012). Decision criteria in dual discrimination tasks estimated using external-noise methods. *Attention, Perception, & Psychophysics, 74,* 1042-1055.

Appendix A

R and JAGS Code

A.1. Chapter 2: Psychometric Function Model Code

```
model{
   for (j in 1:nsubj){
      for (i in 1:ntone){
         r[i,j] ~ dbin(thetalim[i,j], n[i,j])
         logit(thetalim[i,j]) <- lthetalim[i,j]
         lthetalim[i,j] <- min(999, max(-999, ltheta[i,j]))
         ltheta[i,j] <- alpha[j] + beta[j]*tones[i]
      }
      beta[j] ~ dnorm(mu.beta[cond[j]], tau.beta[cond[j]])
      alpha[j] ~ dnorm(mu.alpha[cond[j]], tau.alpha[cond[j]])
   }

   for (k in 1:ncond){

      mu.beta[k] ~ dnorm(0,.001)
      mu.alpha[k] ~ dnorm(0,.001)
      sigma.beta[k] ~ dunif(0,10)
      sigma.alpha[k] ~ dunif(0,10)
      tau.beta[k] <- pow(sigma.beta[k], -2)
      tau.alpha[k] <- pow(sigma.alpha[k], -2)
   }
}
```

A.2. Chapter 2: JAGS instantiation for Psychometric Function

```
data = c("r", "n", "nsubj", "ntone", "tones", "ncond", "cond")
parameters = c("mu.alpha","mu.beta")
myinits = list(
  list(mu.beta = runif(ncond,0,.5), mu.alpha = runif(ncond,-2,2)),
  list(mu.beta = runif(ncond,0,.5), mu.alpha = runif(ncond,-2,2)),
  list(mu.beta = runif(ncond,0,.5), mu.alpha = runif(ncond,-2,2)),
  list(mu.beta = runif(ncond,0,.5), mu.alpha = runif(ncond,-2,2))
)

samples = jags(data, inits = myinits, parameters, model.file =
"model.hier.txt", n.chains = 3, n.iter = 10000, n.burnin = 500,
n.thin = 1, DIC = T)

alpha = samples$BUGSoutput$sims.list$mu.alpha
beta = samples$BUGSoutput$sims.list$mu.beta
# Transform posterior so values expressed in psychometric SD units
postE = 1/beta2[,1]
postI = 1/beta2[,2]
```

A.3. Chapter 2: Error-Correction Model Code

```
model{
   for(j in 1:Subs){
      for(i in 1:N){
         h[i,j] ~ dbin(hr[i,j],n1[i,j])
         f[i,j] ~ dbin(far[i,j],n2[i,j])
         hr[i,j] <- 1-phi(c[i,j]-d[j])
         far[i,j] <- 1-phi(c[i,j])
```

```
      }
      d[j] ~ dnorm(D[cond[j]],precD)
      c[1,j] <- c1[j]
      for(i in 1:(N-1)){
         c[i+1,j] <- c[i,j]+Shift[s[i,j]+1,r[i,j]+1]*delta[j]
      }
      delta[j] ~ dnorm(Delta[cond[j]],precDelta)
      c1[j] ~ dnorm(C1[cond[j]],precC1)
   }

   for(k in 1:Cond){
      D[k] ~ dunif(0,4)
      Delta[k] ~ dunif(-1,1)
      C1[k] ~ dunif(0,4)
   }

   sdD ~ dunif(0,10)
   sdDelta ~ dunif(0,10)
   sdC1 ~ dunif(0,10)

   precD <- pow(sdD,-2)
   precDelta <- pow(sdDelta,-2)
   precC1 <- pow(sdC1,-2)
}
```

A.4. Chapter 2: JAGS instantiation for Error-Correction Model

```
data = list("s", "r", "h", "f",
"n1","n2","cond","N","Subs","Cond","Shift")
parameters = c("D", "C1", "Delta")

samples = jags.parallel(data,inits=NULL,parameters,
model.file="model.ecm.txt",n.chains = 4, n.iter = 10000, n.burnin
= 1000, n.thin=2, DIC=T)

D=samples$BUGSoutput$sims.list$D
C1=samples$BUGSoutput$sims.list$C1
Delta=samples$BUGSoutput$sims.list$Delta
```

A.5. Chapter 2: Marginal Likelihoods for ECM and SCM

ECM

```
nsim = 1e4
D=runif(nsim,0,4)
C=runif(nsim,0,4)
G=runif(nsim,-1,1)

H=1-pnorm(C-D)
F=1-pnorm(C)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
```

```
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C = C+Shift[ss[i]+1,rs[i]+1]*G
      H=1-pnorm(C-D)
      F=1-pnorm(C)
   }
}

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C = C+Shift[ss[i]+1,rs[i]+1]*G
      H=1-pnorm(C-D)
      F=1-pnorm(C)
   }
}

# Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C = C+Shift[ss[i]+1,rs[i]+1]*G
      H=1-pnorm(C-D)
      F=1-pnorm(C)
   }
}

# Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C = C+Shift[ss[i]+1,rs[i]+1]*G
```

```
      H=1-pnorm(C-D)
      F=1-pnorm(C)
   }
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-ecm.RData")
```

SCM

```
nsim = 1e5
D=runif(nsim,0,4)
C=runif(nsim,0,4)
G=runif(nsim,0,1)

H=1-pnorm(C-D)
F=1-pnorm(C)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j]
   fs=f[,j]
   n1s=n1[,j]
   n2s=n2[,j]

   for(i in 1:ntrials) ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
   for(i in 1:ntrials) ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
}

#Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j]
   fs=f[,j]
   n1s=n1[,j]
   n2s=n2[,j]

   for(i in 1:ntrials) ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
   for(i in 1:ntrials) ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
}

#Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j]
   fs=f[,j]
   n1s=n1[,j]
```

```
   n2s=n2[,j]

   for(i in 1:ntrials) ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
   for(i in 1:ntrials) ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
}

#Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j]
   fs=f[,j]
   n1s=n1[,j]
   n2s=n2[,j]

   for(i in 1:ntrials) ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
   for(i in 1:ntrials) ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-scm.RData")
```

A.6. Chapter 2: Simulating d' based upon psychometric parameters

```
# Number of trial per event and number of simulations
nHigh = nLow = 200
nsim=1e4

# Means and SDs for criterion distribution: from psychometric
function
tmp1 = c(7.40,7.45)
tmp2 = c(1.28,.62)

# Standardise the criterion values
c = (tmp1-6.5)/2
sd = tmp2/2

# Set up array to store d prime values
ds = array(dim=c(2,nsim))

# Run simulation loop
for(i in 1:nsim){
   C1 = c[1]
   C2 = c[2]

   # criterion variance degraded d prime value
   D1 = 1/sqrt(1+sd[1]^2)
   D2 = 1/sqrt(1+sd[2]^2)

   # calculate hits and false alarms
   H1 = 1-pnorm(C1-D1); F1 = 1-pnorm(C1)
   H2 = 1-pnorm(C2-D2); F2 = 1-pnorm(C2)
```

```
   h1 = sum(rbinom(nHigh,1,H1))/nHigh
   f1 = sum(rbinom(nLow,1,F1))/nLow
   h2 = sum(rbinom(nHigh,1,H2))/nHigh
   f2 = sum(rbinom(nLow,1,F2))/nLow

   # estimate d prime from "observed" values
   ds[1,i] = qnorm(h1)-qnorm(f1)
   ds[2,i] = qnorm(h2)-qnorm(f2)
}
```

Note: The code for the psychometric function and the error-correction model is identical to that in Chapter 2. Additionally, code for obtaining the marginal likelihoods for the SCM is unchanged. For this reason the code has not been reproduced in the following sections.

A.7. Chapter 3: Pure Error-Correction Model Code

```
model{
   for(j in 1:Subs){
      for(i in 1:N){
         h[i,j] ~ dbin(hr[i,j],n1[i,j])
         f[i,j] ~ dbin(far[i,j],n2[i,j])
         hr[i,j] <- 1-phi(c[i,j]-d[j])
         far[i,j] <- 1-phi(c[i,j])
      }

      d[j] ~ dnorm(D[cond[j]],precD)
      c[1,j] <- c1[j]
      for(i in 1:(N-1)){
           c[i+1,j] <- c[i,j]+Shift[s[i,j]+1,r[i,j]+1]*(x[i,j]-
           c[i,j])*delta[j]
      }

      delta[j] ~ dnorm(Delta[cond[j]],precDelta)
      c1[j] ~ dnorm(C1[cond[j]],precC1)
   }

   D[1] ~ dunif(0.5,1)
   D[2] ~ dunif(0.5,1)
   D[3] ~ dunif(2,3)
   D[4] ~ dunif(2,3)
   C1[1] ~ dunif(0,1)
   C1[2] ~ dunif(0,1)
   C1[3] ~ dunif(1,2)
   C1[4] ~ dunif(1,2)

   for(k in 1:Cond){
      Delta[k] ~ dunif(0,1)
   }

   sdD ~ dunif(0,10)
```

```
   sdDelta ~ dunif(0,10)
   sdC1 ~ dunif(0,10)

   precD <- pow(sdD,-2)
   precDelta <- pow(sdDelta,-2)
   precC1 <- pow(sdC1,-2)
}
```

A.8. Chapter 3: JAGS instantiation for PECM

```
# Shift matrix contains only positive values to indicate on which
#type of trial shifts can occur. Here it is only on errors.
# The direction is determined by x - c.
Shift = matrix(c(0,1,1,0),2,2,byrow=T)

data = list("x", "s", "r", "h", "f", "n1",
"n2","cond","N","Subs","Cond","Shift")
parameters = c("D", "C1", "Delta")

samples = jags.parallel(data,inits=NULL,parameters,
model.file="model.pecm.txt",n.chains = 4, n.iter = 10000, n.burnin
= 500, n.thin=1, DIC=T)

D=samples$BUGSoutput$sims.list$D
C1=samples$BUGSoutput$sims.list$C1
Delta=samples$BUGSoutput$sims.list$Delta
```

A.9. Chapter 3: Marginal Likelihood for PECM

```
nsim = 1e5
D1=runif(nsim,0.5,1); D2=runif(nsim,2,3)
C1=runif(nsim,0,1); C2=runif(nsim,1,2)
G=runif(nsim,0,1)

H=1-pnorm(C1-D1)
F=1-pnorm(C1)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+Shift[ss[i]+1,rs[i]+1]*G*(xs[i]-C1)
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
```

```
}

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+Shift[ss[i]+1,rs[i]+1]*G*(xs[i]-C1)
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

H=1-pnorm(C2-D2)
F=1-pnorm(C2)

# Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+Shift[ss[i]+1,rs[i]+1]*G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

# Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+Shift[ss[i]+1,rs[i]+1]*G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
```

```
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-pecm.RData")
```

A.10. Chapter 3: General Deterministic Model Code

```
model{
   for(j in 1:Subs){
      for(i in 1:N){
         h[i,j] ~ dbin(hr[i,j],n1[i,j])
         f[i,j] ~ dbin(far[i,j],n2[i,j])
         hr[i,j] <- 1-phi(c[i,j]-d[j])
         far[i,j] <- 1-phi(c[i,j])
      }
      d[j] ~ dnorm(D[cond[j]],precD)
      c[1,j] <- c1[j]
      for(i in 1:(N-1)){
         c[i+1,j] <- c[i,j]+(Shift1[s[i,j]+1,r[i,j]+1]*delta1[j])+
           (Shift2[s[i,j]+1,r[i,j]+1]*delta2[j])
      }
      c1[j] ~ dnorm(C1[cond[j]],precC1)
      delta1[j] ~ dnorm(Delta[cond[j],1],precDelta)
      delta2[j] ~ dnorm(Delta[cond[j],2],precDelta)
   }

   D[1] ~ dunif(.5,1)
   D[2] ~ dunif(.5,1)
   D[3] ~ dunif(2,3)
   D[4] ~ dunif(2,3)
   C1[1] ~ dunif(0,1)
   C1[2] ~ dunif(0,1)
   C1[3] ~ dunif(1,2)
   C1[4] ~ dunif(1,2)

   for(k in 1:Cond){
      for(n in 1:2){
         Delta[k,n] ~ dunif(-.5,.5)
      }
   }

   sdD ~ dunif(0,10)
   sdDelta ~ dunif(0,10)
   sdC1 ~ dunif(0,10)
   precD <- pow(sdD,-2)
   precDelta <- pow(sdDelta,-2)
   precC1 <- pow(sdC1,-2)
}
```

A.11. Chapter 3: JAGS instantiation for GDM

```
data = list("s", "r", "h", "f",
"n1","n2","cond","N","Subs","Cond","Shift1","Shift2")
```

```
parameters = c("D", "C1", "Delta")

samples = jags.parallel(data,inits=NULL,parameters,
model.file="model.gdm.txt", n.chains = 4, n.iter = 10000, n.burnin
= 500, n.thin=1, DIC=T)

D=samples$BUGSoutput$sims.list$D
C1=samples$BUGSoutput$sims.list$C1
Delta=samples$BUGSoutput$sims.list$Delta
```

A.12. Chapter 3: Marginal Likelihood for GDM

```
nsim = 1e5
D1=runif(nsim,.5,1); D2=runif(nsim,2,3)
C1=runif(nsim,0,1); C2=runif(nsim,1,2)
G1=G2=runif(nsim,-.5,.5)

H=1-pnorm(C1-D1)
F=1-pnorm(C1)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+Shift1[ss[i]+1,rs[i]+1]*G1+
           Shift2[ss[i]+1,rs[i]+1]*G2
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+Shift1[ss[i]+1,rs[i]+1]*G1+
           Shift2[ss[i]+1,rs[i]+1]*G2
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}
```

```
H=1-pnorm(C2-D2)
F=1-pnorm(C2)

# Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+Shift1[ss[i]+1,rs[i]+1]*G1+
           Shift2[ss[i]+1,rs[i]+1]*G2
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

# Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+Shift1[ss[i]+1,rs[i]+1]*G1+
           Shift2[ss[i]+1,rs[i]+1]*G2
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-gdm.RData")
```

A.13. Chapter 3: Pure General Deterministic Model Code

Note that model code and implementation for the pure error-correction model is the same with and without TTKR.

```
model{
   for(j in 1:Subs){
      for(i in 1:N){
         h[i,j] ~ dbin(hr[i,j],n1[i,j])
```

```
         f[i,j] ~ dbin(far[i,j],n2[i,j])
         hr[i,j] <- 1-phi(c[i,j]-d[j])
         far[i,j] <- 1-phi(c[i,j])
      }
      d[j] ~ dnorm(D[cond[j]],precD)
      c[1,j] <- c1[j]
      for(i in 1:(N-1)){
         c[i+1,j] <- c[i,j]+(x[i,j]-c[i,j])*delta[j]
      }
      delta[j] ~ dnorm(Delta[cond[j]],precDelta)
      c1[j] ~ dnorm(C1[cond[j]],precC1)
   }

   D[1] ~ dunif(.5,1)
   D[2] ~ dunif(.5,1)
   D[3] ~ dunif(2,3)
   D[4] ~ dunif(2,3)
   C1[1] ~ dunif(0,1)
   C1[2] ~ dunif(0,1)
   C1[3] ~ dunif(1,2)
   C1[4] ~ dunif(1,2)

   for(k in 1:Cond){
      Delta[k] ~ dunif(0,1)
   }

   sdD ~ dunif(0,10)
   sdDelta ~ dunif(0,10)
   sdC1 ~ dunif(0,10)

   precD <- pow(sdD,-2)
   precDelta <- pow(sdDelta,-2)
   precC1 <- pow(sdC1,-2)
}
```

A.14. Chapter 3: JAGS instantiation for PGDM

```
data = list("x", "s", "r", "h", "f", "n1",
"n2","cond","N","Subs","Cond")
parameters = c("D", "C1", "Delta")

samples = jags.parallel(data,inits=NULL,parameters,
model.file="model.pgdm.txt",n.chains = 4, n.iter = 10000, n.burnin
= 500, n.thin=1, DIC=T)

D=samples$BUGSoutput$sims.list$D
C1=samples$BUGSoutput$sims.list$C1
Delta=samples$BUGSoutput$sims.list$Delta
```

A.15. Chapter 3: Marginal Likelihood for PGDM

```
nsim = 1e5
D1=runif(nsim,0.5,1); D2=runif(nsim,2,3)
C1=runif(nsim,0,1); C2=runif(nsim,1,2)
G=runif(nsim,0,1)
```

```
H=1-pnorm(C1-D1)
F=1-pnorm(C1)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+G*(xs[i]-C1)
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+G*(xs[i]-C1)
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

H=1-pnorm(C2-D2)
F=1-pnorm(C2)

# Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}
```

```
# Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-pgdm.RData")
```

A.16. Chapter 5: Marginal Likelihood for EWLM (w/ TTKR)

```
# Assign tones to low and high vectors based upon TTKR values
z = as.vector(tone)
low = matrix(ifelse(tmp$KR==0,z,0),400,nsubs)
high = matrix(ifelse(tmp$KR==1,z,0),400,nsubs)

for(i in 2:400) for(j in 1:nsubs) low[i,j] =
ifelse(low[i,j]==0,low[i-1,j],low[i,j])
for(i in 2:400) for(j in 1:nsubs) high[i,j] =
ifelse(high[i,j]==0,high[i-1,j],high[i,j])

nsim = 1e5
D1=runif(nsim,.5,1); D3=runif(nsim,2,3)
G=matrix(runif(nsim,0,1),nsim,ntrials)
lag = matrix(seq(0,399),nsim,ntrials, byrow=T)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   lows=matrix(low[,j],nrow=nsim,ncol=ntrials,byrow=T)
   highs=matrix(high[,j],nrow=nsim,ncol=ntrials,byrow=T)
   Mn = Ms = array(dim=c(nsim,ntrials))

   C=0.5
   H=1-pnorm(C-D1)
   F=1-pnorm(C)
```

```
   ll[1,,1,j]=dbinom(hs[1],n1s[1],H)
   ll[2,,1,j]=dbinom(fs[1],n2s[1],F)
   Mn[,1] = lows[,1]
   Ms[,1] = highs[,1]

   for(i in 1:(ntrials-1)){
      Mn[,i+1]=apply(lows[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Ms[,i+1]=apply(highs[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Mn[,i+1]=ifelse(is.na(Mn[,i+1]),0,Mn[,i+1])
      Ms[,i+1]=ifelse(is.na(Ms[,i+1]),0,Ms[,i+1])
      C = 0.5*(Mn[,i+1]+Ms[,i+1])
      H=1-pnorm(C-D1)
      F=1-pnorm(C)
      ll[1,,i+1,j]=dbinom(hs[i+1],n1s[i+1],H,log=T)
      ll[2,,i+1,j]=dbinom(fs[i+1],n2s[i+1],F,log=T)
   }
}

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   lows=matrix(low[,j],nrow=nsim,ncol=ntrials,byrow=T)
   highs=matrix(high[,j],nrow=nsim,ncol=ntrials,byrow=T)
   Mn = Ms = array(dim=c(nsim,ntrials))

   C=0.5
   H=1-pnorm(C-D1)
   F=1-pnorm(C)
   ll[1,,1,j]=dbinom(hs[1],n1s[1],H)
   ll[2,,1,j]=dbinom(fs[1],n2s[1],F)
   Mn[,1] = lows[,1]
   Ms[,1] = highs[,1]

   for(i in 1:(ntrials-1)){
      Mn[,i+1]=apply(lows[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Ms[,i+1]=apply(highs[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Mn[,i+1]=ifelse(is.na(Mn[,i+1]),0,Mn[,i+1])
      Ms[,i+1]=ifelse(is.na(Ms[,i+1]),0,Ms[,i+1])
      C = 0.5*(Mn[,i+1]+Ms[,i+1])
      H=1-pnorm(C-D1)
      F=1-pnorm(C)
      ll[1,,i+1,j]=dbinom(hs[i+1],n1s[i+1],H,log=T)
      ll[2,,i+1,j]=dbinom(fs[i+1],n2s[i+1],F,log=T)
   }
}

# Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   lows=matrix(low[,j],nrow=nsim,ncol=ntrials,byrow=T)
```

```
   highs=matrix(high[,j],nrow=nsim,ncol=ntrials,byrow=T)
   Mn = Ms = array(dim=c(nsim,ntrials))

   C=1.5
   H=1-pnorm(C-D3)
   F=1-pnorm(C)
   ll[1,,1,j]=dbinom(hs[1],n1s[1],H)
   ll[2,,1,j]=dbinom(fs[1],n2s[1],F)
   Mn[,1] = lows[,1]
   Ms[,1] = highs[,1]

   for(i in 1:(ntrials-1)){
      Mn[,i+1]=apply(lows[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Ms[,i+1]=apply(highs[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Mn[,i+1]=ifelse(is.na(Mn[,i+1]),0,Mn[,i+1])
      Ms[,i+1]=ifelse(is.na(Ms[,i+1]),0,Ms[,i+1])
      C = 0.5*(Mn[,i+1]+Ms[,i+1])
      H=1-pnorm(C-D3)
      F=1-pnorm(C)
      ll[1,,i+1,j]=dbinom(hs[i+1],n1s[i+1],H,log=T)
      ll[2,,i+1,j]=dbinom(fs[i+1],n2s[i+1],F,log=T)
   }
}

# Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   lows=matrix(low[,j],nrow=nsim,ncol=ntrials,byrow=T)
   highs=matrix(high[,j],nrow=nsim,ncol=ntrials,byrow=T)
   Mn = Ms = array(dim=c(nsim,ntrials))

   C=1.5
   H=1-pnorm(C-D3)
   F=1-pnorm(C)
   ll[1,,1,j]=dbinom(hs[1],n1s[1],H)
   ll[2,,1,j]=dbinom(fs[1],n2s[1],F)
   Mn[,1] = lows[,1]
   Ms[,1] = highs[,1]

   for(i in 1:(ntrials-1)){
      Mn[,i+1]=apply(lows[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Ms[,i+1]=apply(highs[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Mn[,i+1]=ifelse(is.na(Mn[,i+1]),0,Mn[,i+1])
      Ms[,i+1]=ifelse(is.na(Ms[,i+1]),0,Ms[,i+1])
      C = 0.5*(Mn[,i+1]+Ms[,i+1])
      H=1-pnorm(C-D3)
      F=1-pnorm(C)
      ll[1,,i+1,j]=dbinom(hs[i+1],n1s[i+1],H,log=T)
      ll[2,,i+1,j]=dbinom(fs[i+1],n2s[i+1],F,log=T)
   }
}
```

```
#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-ewlm.RData")
```

A.17. Chapter 5: Marginal Likelihood for EWLM (no TTKR)

```
# Assign tones to low and high vectors based upon TTKR values
z = as.vector(tone)
low = matrix(ifelse(r==0,z,0),400,nsubs)
high = matrix(ifelse(r==1,z,0),400,nsubs)

for(i in 2:400) for(j in 1:nsubs) low[i,j] =
ifelse(low[i,j]==0,low[i-1,j],low[i,j])
for(i in 2:400) for(j in 1:nsubs) high[i,j] =
ifelse(high[i,j]==0,high[i-1,j],high[i,j])

nsim = 1e5

D1=runif(nsim,.5,1); D3=runif(nsim,2,3)
G=matrix(runif(nsim,0,1),nsim,ntrials)
lag = matrix(seq(0,399),nsim,ntrials, byrow=T)

ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   lows=matrix(low[,j],nrow=nsim,ncol=ntrials,byrow=T)
   highs=matrix(high[,j],nrow=nsim,ncol=ntrials,byrow=T)
   Mn = Ms = array(dim=c(nsim,ntrials))

   C=0.5
   H=1-pnorm(C-D1)
   F=1-pnorm(C)
   ll[1,,1,j]=dbinom(hs[1],n1s[1],H)
   ll[2,,1,j]=dbinom(fs[1],n2s[1],F)
   Mn[,1] = lows[,1]
   Ms[,1] = highs[,1]

   for(i in 1:(ntrials-1)){
      Mn[,i+1]=apply(lows[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Ms[,i+1]=apply(highs[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Mn[,i+1]=ifelse(is.na(Mn[,i+1]),0,Mn[,i+1])
      Ms[,i+1]=ifelse(is.na(Ms[,i+1]),0,Ms[,i+1])
      C = 0.5*(Mn[,i+1]+Ms[,i+1])
      H=1-pnorm(C-D1)
      F=1-pnorm(C)
```

```
      ll[1,,i+1,j]=dbinom(hs[i+1],n1s[i+1],H,log=T)
      ll[2,,i+1,j]=dbinom(fs[i+1],n2s[i+1],F,log=T)
   }
}

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   lows=matrix(low[,j],nrow=nsim,ncol=ntrials,byrow=T)
   highs=matrix(high[,j],nrow=nsim,ncol=ntrials,byrow=T)
   Mn = Ms = array(dim=c(nsim,ntrials))

   C=1.5
   H=1-pnorm(C-D3)
   F=1-pnorm(C)
   ll[1,,1,j]=dbinom(hs[1],n1s[1],H)
   ll[2,,1,j]=dbinom(fs[1],n2s[1],F)
   Mn[,1] = lows[,1]
   Ms[,1] = highs[,1]

   for(i in 1:(ntrials-1)){
      Mn[,i+1]=apply(lows[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Ms[,i+1]=apply(highs[,1:(i+1)]*G[,i+1]^(i-lag[,1:(i+1)]),
           1,sum)/apply(G[,i+1]^(i-lag[,1:(i+1)]),1,sum)
      Mn[,i+1]=ifelse(is.na(Mn[,i+1]),0,Mn[,i+1])
      Ms[,i+1]=ifelse(is.na(Ms[,i+1]),0,Ms[,i+1])
      C = 0.5*(Mn[,i+1]+Ms[,i+1])
      H=1-pnorm(C-D3)
      F=1-pnorm(C)
      ll[1,,i+1,j]=dbinom(hs[i+1],n1s[i+1],H,log=T)
      ll[2,,i+1,j]=dbinom(fs[i+1],n2s[i+1],F,log=T)
   }
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-ewlm.noKR.RData")
```

A.18. Chapter 5: Marginal Likelihood for PGDM

```
nsim = 1e5
D1=runif(nsim,0.5,1); D2=runif(nsim,2,3)
C1=rep(0.5,nsim); C2=rep(1.5,nsim)
G=runif(nsim,0,1)

H=1-pnorm(C1-D1)
F=1-pnorm(C1)
```

```
ll=array(dim=c(2,nsim,ntrials,nsubs)) # rate x sim x trial x
subject

# Condition 1
for(j in 1:nsubcon[1]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   # Pure Error-Correction Model
   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+G*(xs[i]-C1)
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

C1=rep(0.5,nsim); C2=rep(1.5,nsim)
H=1-pnorm(C1-D1)
F=1-pnorm(C1)

# Condition 2
for(j in 1:nsubcon[1]+nsubcon[2]){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   # Pure Error-Correction Model
   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+G*(xs[i]-C1)
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

C1=rep(0.5,nsim); C2=rep(1.5,nsim)
H=1-pnorm(C1-D1)
F=1-pnorm(C1)

# Condition 3
for(j in 1:nsubcon[3]+sum(nsubcon[1:2])){
   print(j)
   hs=h[,j];fs=f[,j]
   rs=r[,j];ss=s[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   # Pure Error-Correction Model
   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C1 = C1+G*(xs[i]-C1)
```

```
      H=1-pnorm(C1-D1)
      F=1-pnorm(C1)
   }
}

C1=rep(0.5,nsim); C2=rep(1.5,nsim)
H=1-pnorm(C2-D2)
F=1-pnorm(C2)

# Condition 4
for(j in 1:nsubcon[4]+sum(nsubcon[1:3])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   # Pure Error-Correction Model
   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

C1=rep(0.5,nsim); C2=rep(1.5,nsim)
H=1-pnorm(C2-D2)
F=1-pnorm(C2)

# Condition 5
for(j in 1:nsubcon[5]+sum(nsubcon[1:4])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]

   # Pure Error-Correction Model
   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

C1=rep(0.5,nsim); C2=rep(1.5,nsim)
H=1-pnorm(C2-D2)
F=1-pnorm(C2)

# Condition 6
for(j in 1:nsubcon[6]+sum(nsubcon[1:5])){
   print(j)
   hs=h[,j];fs=f[,j]
   n1s=n1[,j];n2s=n2[,j]
   xs=x[,j]
```

```
   # Pure Error-Correction Model
   for(i in 1:ntrials){
      ll[1,,i,j]=dbinom(hs[i],n1s[i],H,log=T)
      ll[2,,i,j]=dbinom(fs[i],n2s[i],F,log=T)
      C2 = C2+G*(xs[i]-C2)
      H=1-pnorm(C2-D2)
      F=1-pnorm(C2)
   }
}

#subject marginal likelihood
tmp=apply(ll,c(2,4),sum,na.rm=T)
marginalsS=array(dim=c(dim(tmp)[2]))
for(i in 1:(dim(tmp)[2])) marginalsS[i]=mean(exp(tmp[,i]-
max(tmp[,i])))

save(marginalsS, file="bf-pgdm.RData")
```

Appendix B

Experimental Stimuli

Table B.1:

Tonal stimuli used in experiment.

		Dist. Frequencies		
Tone	Hz	L	HH	HE
1	445.4	1	-	-
2	466.6	1	-	-
3	487.8	2	-	-
4	509.0	3	-	-
5	530.2	5	1	-
6	551.4	8	1	-
7	572.6	11	2	-
8	593.8	14	3	-
9	615.0	16	5	-
10	636.2	19	8	-
11	657.4	20	11	-
12	678.6	20	14	-
13	699.8	19	16	1
14	721.0	16	19	1
15	742.2	14	20	2
16	763.4	11	20	3
17	784.6	8	19	5
18	805.8	5	16	8
19	827.0	3	14	11
20	848.2	2	11	14
21	869.4	1	8	16
22	890.6	1	5	19
23	911.8	-	3	20
24	933.0	-	2	20
25	954.2	-	1	19
26	975.4	-	1	16
27	996.6	-	-	14
28	1017.8	-	-	11
29	1039.0	-	-	8
30	1060.2	-	-	5
31	1081.4	-	-	3
32	1102.6	-	-	2
33	1123.8	-	-	1
34	1145.0	-	-	1

Appendix C

Justifying Sensory Control

The tones used for present investigation were identical in range to those used by Podd (1975) and Taylor (2010). Prior analyses found the tones to be highly discriminable though upon reanalysis the number of JNDs separating adjacent tones was found to be in error. The stimuli consist of a series of 34 pure tones that ranged from 445.4 Hz to 1145.0 Hz, separated by 21.2 Hz steps. It had previously been determined that this constant separation equated to a difference of approximately 4-6 JNDs; as justified by Shower and Biddulph (1931; see also Podd, 1975). However, this is incorrect and the number of JNDs separating adjacent tones is greater, as will be shown next.

Weber's Law states that the difference threshold (ΔS) – or the amount of separation required between two stimuli in order for them to be discriminated - increases proportionally as the magnitude of the stimulus (S) increases, where the ratio between these two values (k) is a constant. Formally stated, $\Delta S/S = k$. For the present discussion ΔS can be approximated by $\Delta S = S_i - S_{i-1}$, where S_i is the i^{th} stimulus in the series and S_{i-1} is the immediately preceding stimulus. If k is known the difference threshold is equal to $\Delta S = kS$. For each increase in stimulus magnitude the proportion required to maintain a constant k should grow exponentially. This relationship holds for either S or ΔS, where $S_n = S_{n-1}e^k$ and $\Delta S = \Delta S_{n-1}e^k$. However, the fixed separation (i.e., 21.2 Hz) between all adjacent stimuli means that ΔS in increasingly linearly, not exponentially. This means that in fact k is not constant and is a *decreasing* function of ΔS, where $k = \Delta S.S^{-1}$. Unfortunately, what this also means is that the number of JNDs are a decreasing function of stimulus magnitude which presents a concern in justifying sensory control.

Reading off Shower and Biddulph's (1931, p. 279) table for Weber's fraction, the JND for monaural detection in the frequency range of 250 Hz – 2000 Hz, can range anywhere from .0355-.0079 at the 5db level to .0107-.0018 at the 80db level. It appears, though, that the original JND figures reported were derived from this table and are inappropriate for binaural detection. Instead, Figure 10 (Shower & Bidudulph, 1931) should have been consulted which provides the JNDs for binaural detection. Fortunately, the size of k remains relatively constant across the specified range; the only difference is that k is smaller in the binaural

case. By inference we can assume that for the binaural presentation of tones k will remain relatively constant as the sensation level increases. Reading off Figure 10 it was determined that $k \approx .003$, and the number of JNDs separating the tonal stimuli was estimated by simply dividing the difference between adjacent tones by k.

Reading values off a table would appear a rather imprecise method in establishing the size of a JND. Instead, we can borrow a rather convenient rule of thumb from the psychoacoustic literature that provides an alternative metric for the JND. The cent ($¢$) is a logarithmic unit that is typically used to measure musical intervals. A handy feature of the unit is that a single JND is approximately equal to $5¢$. A further convenience is that k and $¢$ are monotonic. Like k, the difference threshold increases proportionally with stimulus magnitude, with the difference required growing exponentially; this time where $S_n = S_{n-1.}2^{¢/1200}$. If we compare both $S_n = S_{n-1}e^k$ and $S_n = S_{n-1.}2^{¢/1200}$ it can be seen that they are simply transformations of each other, and rearrangement of both lead to the following equality

$$\frac{S_n}{S_{n-1}} = 2^{¢/1200} = e^k. \qquad \text{Eq. C.1}$$

Taking the natural logarithm of all terms results in

$$\ln\left[\frac{S_n}{S_{n-1}}\right] = \frac{¢}{1200}\ln 2 = k. \qquad \text{Eq. C.2}$$

Solving for $¢$ is a simple matter of algebra, where

$$¢ = 1200\,\frac{\ln\left[\frac{S_n}{S_{n-1}}\right]}{\ln 2} = 1200\frac{k}{\ln 2}. \qquad \text{Eq. C.3}$$

Given this relationship the JNDs inferred from Shower and Biddulph's paper can be directly compared. The number of cents separating the adjacent tones are plotted alongside the number of JNDs based upon the Figure 10 in Shower and Biddulph (1931). While the cent does not appear to be a good estimator of the JND in monaural detection, the curves are

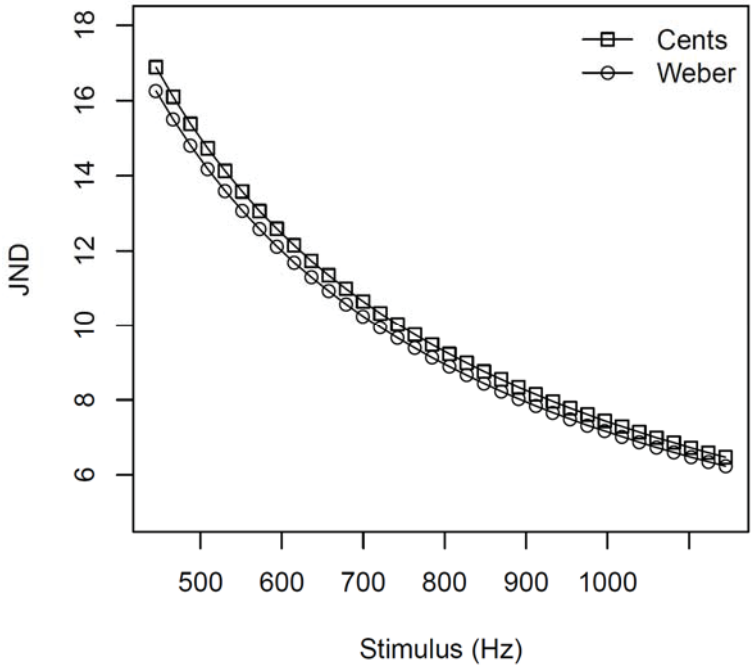

Figure C1: Just noticeable differences (JND) calculated using either Weber's Fraction (k) or the logarithmic cent ($¢$) transformation.

sufficiently close in the binaural case and it appears that there is good correspondence between the measures (see Figure C1). The crucial element to derive from the curves is that even in the higher tone regions where the number of JNDs is dropping there are still at least 6 JNDs between tones. This should satisfy the assumption of the sensory control.