Criterion Variance in Signal Detection Theory:

The Interactive Effect of Knowledge of Results and Task Difficulty on Binary Decision Tasks

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Arts

in

Psychology

at Massey University, Palmerston North,
New Zealand.

Robert T. Taylor

2010

Acknowledgements

To my supervisor, Dr. John Podd, I would like to express my deepest gratitude for his tireless support and guidance, his endless supply of Signal Detection resources, and his belief in the research. His tutelage, and patience, in refining my Signal Detection knowledge has been invaluable.

Thank you to Malcolm Loudon for all his help in the programming of the computer application. His support and skills was an asset that this research could not have done without.

To my fellow postgraduate students, thank you for providing me a sounding board to bounce ideas off, for providing advice, and an avenue for venting frustration in general.

To my family and friends, thanks are due for their understanding and support, and for putting up with my continual absenteeism.

And finally, thanks are due to my partner Steph. Her relentless support and encouragement made the research process that much easier, and was pivotal in the completion of this research.

Abstract

Within traditional Signal Detection Theory (SDT) experiments decision noise is very rarely considered, with researchers clinging to the assumption that the decision criterion has no associated variability. This assumption is incorrect. Furthermore, two factors contribute to criterion fluctuation: task difficulty and the type of knowledge of results (KR) delivered to the observer. The accepted standard in SDT experiments is to provide veridical trial-by-trial feedback ($TTKR_e$). This type of KR may adversely affect observer performance when the decision task is difficult, as the KR may appear highly inconsistent to the observer. The present study hypothesised that providing KR relative to the optimal criterion location ($TTKR_i$) would minimise criterion fluctuation. The present Criterion Variance Model (CVM) assumes that the decision criterion in SDT is subject to fluctuation. Two hypotheses were derived to test the model: a) contrary to the assumption of SDT, the decision criterion in a signal detection task is a variable rather than a fixed value on the decision axis, and is present within binary discrimination tasks; and b) There will be an interaction effect between the type of TTKR provided and the difficulty level of the task. Specifically, $TTKR_i$ will enable more accurate decision making than $TTKR_e$, but only for a difficult decision task. Forty-four observers took part in a simple binary decision task, discriminating whether a presented tone was high or low in frequency (Hz). All tones were easily discriminable from each other; thus, the experiment was free from sensory noise. Task difficulty was manipulated by varying the degree of overlap between the high and low distributions, from which the high and low tones were sampled. As predicted by the CVM, performance in a difficult decision task was affected by the type of KR provided. Observers who received $TTKR_e$ performed less well than observers who received $TTKR_i$ in the more difficult version of the task. Despite mean criterion location measures across groups approaching zero – the optimal location – criterion fluctuation was evident when observer error distributions were analysed. Furthermore, the degree of criterion fluctuation was large, and was associated with the level of task difficulty. A major caveat was the lack of a no KR condition. Consequently, the degree to which observers utilised the KR could not be fully assessed. Additionally, the number of tones may have been too small, possibly encouraging observers not to use the KR provided in a consistent manner. Further research should incorporate a no KR condition and increase the number of tonal stimuli while ensuring the tones are still separated by 3 or 4 JNDs. Despite these design issues, the results highlight the potential detrimental effects of veridical KR on performance, particularly under conditions of high uncertainty.

Table of Contents

Table of Appendices

List of Figures

List of Tables

Introduction and Overview

Signal Detection Theory (SDT) has been successfully applied to various disciplines within the field of psychology, most prominently within psychophysics. However, despite the many applied fields that have availed themselves of the techniques SDT has to offer, one principle remains the same, that it is a psychophysical approach to measuring accuracy in the presence of uncertainty (MacMillan & Creelman, 2005).

Though psychology has enjoyed much success with the application, it is curious that it took some time for psychology to become aware of the theory. During the early 20th century detection theory was mostly a product of communications and engineering, with it playing a pivotal role in the development of RADAR (Radio Detection and Ranging). Through the emission of electromagnetic waves, target objects (signals) could be detected; however, one inherent problem with the system was that external noise and unwanted signals corrupted the information as it was sent back to the receiver. This interference affects the signal to noise ratio. The higher the ratio, the more reliable the system is in detecting the target signal. The fundamental tenet is that signals must be detected against a background of noise, and in some instances the signal is barely detectable (Petterson, Birdsall, & Fox, 1954; Pierce, 1980).

Within psychology the basic tenets remain the same; that is, target stimuli must be detected against a background of noise. However, many of the fundamental aspects within SDT had been laid long before its formalisation. Detection and threshold (e.g., Thurston, 1927a) theories were commonplace within psychophysics, concerned primarily with the ability of an observer to detect signals against noise, and formed the foundation upon which SDT was built. Consequently, it is not surprising that SDT share some similarities with earlier detection theories, as many of the fundamental assumptions were retained within the SDT framework (Green & Swets, 1966; McNicol, 1972). Though SDT was first applied to vision experiments during this period (Tanner & Swets, 1954), its formalisation as a psychological theory was not in place until 1966 with Green and Swets' publication *Signal Detection Theory and Psychophysics*.

Since Green and Swets' (1966) seminal publication research has proliferated within psychology. Many decisions must be made under conditions where complete certainty does not exist. Rules govern our decision processes in the hope of making consistently accurate judgements, but they can also lead to erroneous decisions. Many applied fields typify this paradox, for example, radiography, recognition memory, or psychopathology diagnostics. However, what SDT seeks to assess is the ability of an observer to make accurate decisions about events where the evidence available to do so is incomplete. Prior to the development

of SDT, the ability of an observer to discriminate between signal and noise events was assessed without too much concern for how the observer's mere willingness to respond 'signal' or 'noise' biased the decision. SDT is a theory that provides a way of assessing this potential response bias independent of the observer's true degree of sensitivity. In simpler terms, an observer's response bias is essentially independent of sensitivity (the ability of an observer to detect a signal; Green & Swets, 1966), and concerns the propensity of an observer to favour signal or noise responses irrespective of what the true state of the event is (an observer may be biased to responding noise even though the events are signals). How observers do this is by using a cut point, called a criterion, to base their decisions upon. This criterion delineates the conditions under which an event is considered either a signal or a noise, and is the basis upon which an observer's decision rules are derived.

SDT assumes that the criterion the observer uses to decide whether a signal or a noise event has occurred is always fixed somewhere along the decision axis - the range of possible events that could occur. This assumed stability precluded any perceived role the criterion may have had in observer's sensitivity; it was assumed that because it was fixed it could not have been adversely affecting the observer's performance. However, research (e.g., Larkin, 1971; McNicol, 1975) unequivocally points to the fact that the criterion is not fixed during a detection task. Moreover, this 'criterion variance' impinges upon sensitivity estimates (McNicol, 1972); the greater the fluctuation in the criterion placement the greater the observer's sensitivity is underestimated. The principle is that shifting the criterion is not conducive to consistency in decision making because the decision rules are frequently changing, and this ultimately affects performance in detection tasks. Furthermore, there is reason to believe that the degree of criterion fluctuation is some function of task difficulty. Task difficulty is associated with increased levels of uncertainty in the environment, essentially making decisions harder to make. If there is more room for error, there is more chance that the criterion will shift. Effectively, as task difficulty increases, so might the degree of criterion variance.

The intuitive solution to this problem is to offer feedback, or knowledge of results, to help the observer make their decisions. Typically the observer is told when they have made correct and incorrect decisions. Informing the observer when errors have been made is assumed to minimise criterion fluctuation. Paradoxically, feedback may further reduce decisional accuracy, by increasing criterion variability through influencing the observer's response bias. Consequently, the criterion is relocated each time the observer is told they are incorrect. Furthermore, research (e.g., Larkin, 1971) has illustrated that this occurs on a trial by trial basis, meaning that the criterion may be undergoing shifts each time an event occurs. If this is so then there are implications for SDT studies that rely on feedback to train

observers to adopt certain criterion locations. Furthermore, the more difficult the detection task, the more variance the criterion may undergo when feedback is delivered. Feedback is a source of evidence which relays back to the observer the veridical state of the events. Uncertainty affects the feedback for the events in a way analogous to lowering the signal to noise ratio, and makes it difficult to detect true signals. By implication the reliability of the feedback under uncertain conditions is also compromised. Essentially, the feedback conveys this uncertainty, is incomplete, and affects the decision process.

Despite the effect that feedback can have on detection performance, historically SDT has never deemed feedback problematic, and continues to employ feedback in many SDT applications. For the reasons stated above, it might be better to treat feedback as an independent variable in SDT experiments. Furthermore, the effect it has on the criterion maybe to create additional variance, meaning the criterion also must be regarded as a variable. When these variables are free, they can create spurious results. Additionally, they can interact with other variables, such as task difficulty, to further exacerbate this effect. The conditions under which such interactive effects happen need investigating.

The present study attempted to show first, that criterion fluctuation does exist and produces spurious estimates of discriminability. Second, that the degree of criterion variability is influenced by both knowledge of results on each trial and by the difficulty level of the task. Both task difficulty and type of feedback were treated as independent variables. There was no sensory component to the task, as all stimuli were 100% discriminable from each other. The task was concerned with how feedback, particularly under conditions of uncertainty, affected the observer's accuracy to discriminate between two classes of events. Thus, a simple binary discrimination task was used. The quality of the feedback was manipulated in order to demonstrate that different types of feedback provide different results under difficult conditions. The interaction between task difficulty and feedback has far reaching implications in many rule governed fields as it highlights the pivotal role feedback plays in decision accuracy. In fact, any field that requires decisions to be made in the face of uncertainty will have to assess the type of feedback that is used. A better understanding of the role of feedback is imperative in improving decisional accuracy and consistency.

Chapter I

An Overview of Signal Detection Theory

The primary focus of the present research is that of the decision criterion and the effects decision noise has on its stability. This overview of SDT will use as a focus a simple discrimination task. The observer must decide whether a signal or noise event has occurred when provided with an amount of information that does not allow these decisions to be made with certainty. Inherently, some decision errors are bound to occur.

Perceptual judgement and decision tasks are said to have two fundamental processes associated with them: a sensory stage (detecting the events) and a decision response stage (Larkin, 1971; McNicol, 1975). Additionally, Green and Swets (1966) posit three fundamental elements of any binary decision task: (1) two possible states of the world, (2) the evidence received, and (3) the decision proper. In classical research the two possible states are signal + noise and noise alone. In more applied research it may be whether a tumour is present on an MRI, or whether a previously learned face or a new face was presented in a recognition task (MacMillan & Creelman, 2005). In the present study, the evidence presented is a clearly audible pure tone that has been sampled from one of two overlapping distributions of tones: high or low tones.

The following chapter provides an overview of the critical aspects of SDT as they apply to the present research.

SDT Fundamentals

In the traditional sense, a *signal* defines any target stimulus or event that had to be detected against a background of interference, usually called *noise*. The *signal to noise ratio* (SNR) maps the relative strength of the signal against the background noise. When the SNR is high, the signal should be relatively easy to detect; as SNR decreases, discriminating the signal from noise becomes more difficult.

Within psychology the most basic conceptualisation of the detection task involves an observer detecting whether a tone was present against a background of white noise. This type of design is known as the *yes/no* task, and simply requires the observer to decide on each trial whether a signal was present. Variations on this design include old/new faces (recognition memory), or high/low tone pitches (Green & Swets, 1966; MacMillan & Creelman, 2005). SDT also employs two alternative designs. The *rating scale* task is a refinement of the simple yes/no design, allowing observers to rate on a graded scale how confident they were that a signal was present. These ratings provide more information about

the response than a simple yes/no procedure. The ratings can be used to produce a function referred to as a receiver operating characteristic curve, discussed later (Green & Swets, 1966). Finally, the *two-alternative, forced-choice* task (*2AFC*) is similar to the yes/no design, except this time the observer is presented with two stimulus intervals, one of which contains the signal. After presentation the observer must decide which interval contained the signal (McNicol, 1972). In principle, the number of alternatives is unlimited, though it is the 2AFC task that is almost always used.

Experimental paradigms typically manipulate signals through increasing or decreasing their strength, thus altering the SNR. Trials may present stimuli that incorporate both signal + noise, or just noise alone. Stimuli then can represent two possible states of the world: noise (*n*), and signal + noise (*s*). After the stimulus presentation, the observer must then make the best possible decision (either '*S'* or '*N'*), which often reflects a statistical decision based on the likelihood that a presented stimulus favours state *s* or *n*[1] (McNicol, 1972).

Though in the traditional psychophysical sense, noise literally referred to background static, or white noise (McNicol, 1972), noise can also refer to any form of distraction or extraneous stimuli (e.g., lures in a face recognition task) that serve to mask the strength of the signal. This occurs at the sensory stage; typically noise of this kind is referred to as *external noise*, and alters the discriminability of the target. Because it can reduce the effective strength of the signal, it may create perceptual errors that cause observers to misinterpret a stimulus. For this reason it is also referred to as *sensory noise* – noise associated with the perception of the target. This type of noise affects the observer's *sensitivity* – the ability of the observer to discriminate a signal from noise (Green & Swets, 1966; McNicol, 1972). Though traditional psychophysical enquiry was particularly concerned with the sensory stage, as mentioned earlier often sensory noise was investigated without any concern for how the observer's decisions affected it. Consequently, the decision stage also needs investigating, and is the area under investigation in the present study.

Within the decision stage, the source of noise is *internal noise*. This reflects the fact that the internal system is inherently noisy. Neuronal firing rate, general brain activity, and cognitive processes provide a noisy background against which observers try to make accurate internal representations of the stimulus just observed (McNicol, 1972). A decision must be made based upon this representation. However, cognitive load and all manner of variables associated with an individual's makeup can alter this representation. These representations affect the decision rule that the observer uses in order to make a judgement, and affect the

[1] SDT convention states that lower case letters (e.g., *s, n*) refer to both the state of the world, or a stimulus event associated with that state, whereas upper case letters refer to response alternatives, e.g., *S* = responded 'signal'; *N* = responded 'noise'.

accuracy of the decisions when the representation has been distorted. For this reason internal noise can also be regarded as *decision noise* – noise associated with making a decision about a particular stimulus. This type of noise can affect the observer's *response bias* – the tendency to favour a particular response, or state of the world.

The Sensory Stage

Decisions are incumbent upon the perceived stimulus event. Conceptually, stimuli occupy some value along a sensory continuum, typically called the *decision axis*. Alternatively, stimuli are sometimes referred to as *decision, or evidence, variables* (hereafter denoted *x*; Kadlec, 1999; Stanislaw & Todorov, 1999). The implication is that stimuli vary by some magnitude, whether it is loudness, familiarity, length, etc., generally increasing in magnitude along the decision axis.

Above the sensory continuum there exists a psychological decision space (Rosner & Kochanski, 2009), from which stimuli can adopt any number of values within the continuum range. When noise and signal stimuli are sampled from the decision space they are assumed to have been sampled from either a signal or noise distribution. Decision variables, like many random variables, are represented as distributions upon the decision space rather than discrete points on the continuum (McNicol, 1975). A simplifying assumption is that these distributions are normally distributed with equal variances - *N(0,1)*. For convenience this assumption is accepted throughout the following chapters, and all examples refer to the equal variances case, as it directly applies to the research at hand.

In an ideal world the two distributions would not overlap; thus *x* would be truly representative of each hypothetical state. However, these distributions do not neatly reside at opposite ends of the continuum. Consequently, the distributions overlap, ultimately rendering some decision variables ambiguous as certain magnitudes of *x* will invariably reflect either a signal or a signal + noise state (Figure 1; Stanislaw & Tororov, 1999).

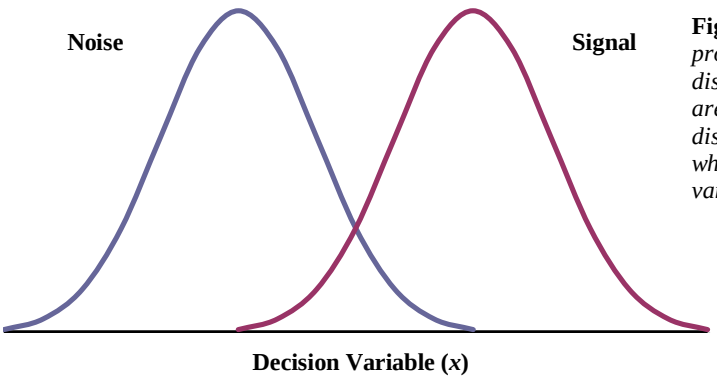


Figure 1: *Varying frequencies of x produce two overlapping Gaussian distributions with equal variances. The area beneath the point where the two distributions cross reflects the area where x could be a noise + signal variables, or simply noise alone.*

More simply, when the distributions overlap this alters the SNR, and ultimately uncertainty is increased which means that, for example, events that appear to be signals may actually be noise. It is assumed in detection theory that the noise distribution remains fixed at one location (commonly standardised as $\mu = 0$, $\sigma = 1$), while the mean of the signal distribution can be moved along the decision axis, varying the degree of overlap. The more the densities overlap, the closer their means become (Figure 2). The degree of separation between the means reflects both a source of sensory noise, and an index of sensitivity.

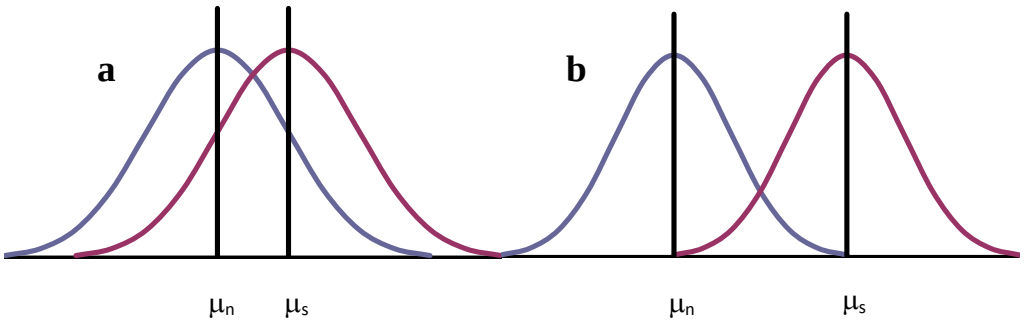


Figure 2: *(a) the degree of overlap is significant resulting in an attenuated distance between the means, and a increase in task difficulty; (b) the degree of overlap is reduced resulting in an increased distance between the means, and a decrease in task difficulty.*

Sensitivity is summarised as the distance between the means of the two distributions and estimated by a measure called *d′* (pronounced dee-prime; McNicol 1972):

$$d' = \frac{\mu_s - \mu_n}{\sigma_n} \qquad \text{(Eq. 1}^{2,3}\text{)}$$

d′ is a standardised unit that measures the distance of the signal mean from the noise mean. In Figure 2a the signal mean is one standard deviation above the noise mean, thus $d' = 1$, whereas in Figure 2b the signal mean is three standard deviations away, so $d' = 3$. *d′* is a parametric statistic; though non-parametric measures do exist, such as d'_e and A' (MacMillan & Creelman, 2005; McNicol, 1972). When *d′* is high, the SNR is also high. Sensitivity can also be estimated by measuring the area under the ROC curve, a topic that will be explored shortly. Typically, under equal variances, both parametric and non-parametric sensitivity indices produce similar results, yet differ when this assumption is violated. *d′* is robust under the equal variance assumption; therefore, it is retained for the purpose of the present

[2] *d′* is similar to a standardised *z* score, if the equal variance assumption is met then the denominator of the equation can be cancelled out.
[3] A glossary of SDT equations can be found in Appendix A.

investigation. Measures for the unequal variances case are beyond the scope of the present research, and as previously mentioned all examples assume equal variances.

Detection tasks also yield two other important statistics: the *hit rate* (HR) and the *false alarm rate* (FAR). *Hits* refer to the number of correctly identified signals, while *false alarms* refer to the number of noise events incorrectly identified as signals. These can be summarised as:

$$HR = n(S) / N(s) \quad \text{(Eq. 2a)}$$

$$FAR = n(S) / N(n) \quad \text{(Eq. 2b)}$$

where *n(S)* is the number of signal responses, and *N(s)* is the number of signal events. Alternatively, *N(n)* is the number of noise events.

Essentially, the HR and FAR are estimates of probabilities associated with increasing values of *x*. The HR assumes the area under the signal distribution to the right of *k* (Figure 3, diagonal stripes), while the FAR assumes the area under the noise distribution to the right of *k* (Figure 3, shaded grey). In this example *k* refers to the criterion, the cut point at which the decision switches from noise to signal. The criterion will be more fully explained in the next section.

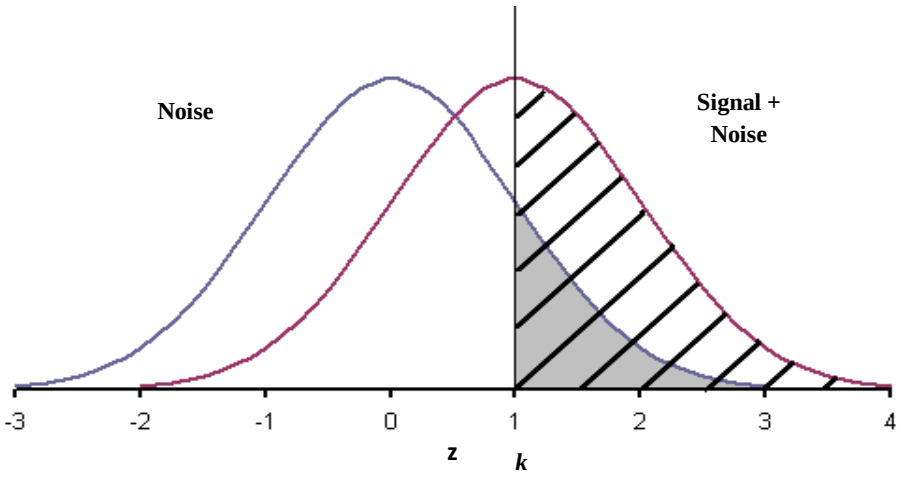


Figure 3: *A cut point has been assumed along the decision axis at z = 1. For this value of x the HR is the area under the signal density marked with diagonal stripes, whereas the FAR is the area under the noise density shaded in grey. In this example d′ = 1.*

Starting at the right of the decision axis and moving left, the HR and FAR probabilities are cumulative for decreasing values of *x*. They can be expressed as conditional probabilities:

$$HR = P(S \mid s) \quad \text{(Eq. 3a)}$$

$$FAR = P(S \mid n) \quad \text{(Eq. 3b)}$$

where *P(S|s)* is the probability that a signal response (*S*) will be made given a signal event (*s*) has occurred. Alternatively, *P(S|n)* is the probability that a signal response will be made given that a noise event has occurred (*n*). Furthermore, the HR and FAR can be converted into standardised *z* scores[4]. The *z*(HR) and the *z*(FAR) can then be used to calculate *d′* (MacMillan & Creelman, 2005):

$$d' = z(HR) - z(FAR) \qquad \text{(Eq. 4)}$$

Graphically, sensitivity can be expressed by plotting the HR and FAR values for any given point along the decision axis while holding *d′* constant (Verde, MacMillan, & Rotello, 2006). These points will produce a curved plot called a *Receiver-Operating Characteristic* (ROC curve. Green & Swets, 1966; MacMillan & Creelman, 2005; see Figure 4, on the next page). Theoretically, stimuli are continuous and can potentially adopt any value, thus an infinite number of HR and FAR points can be plotted for a fixed level of sensitivity. This implies continuous underlying sensory distributions[5], and ROC plots under these assumptions produce smooth curves. In actual practice this is not the case, and values of *x* represent discrete values of magnitude. The present study uses ROC curves where the discrete points are joined by straight lines, as the underlying distributions are discrete probability distributions.

The ROC curve is an *isosensitivity curve*. All points on the curve (except points 0,0, and 1,1) yield the same *d′* value. Shifting of the criterion along the decision axis from right to left can be used to generate the ROC curves. Theoretically, *d′* is independent of the criterion; the latter's value is determined by a place on the ROC curve. Thus, an observer's criterion position may fall below the negative diagonal (a "strict" position), or above (a "lax" position). The great strength of SDT is that it provides independent estimates of an observer's sensitivity and response bias.

Figure 4 (on the next page) illustrates that as sensitivity increases (*d′* = 3), the bow of the ROC pushes up into the left hand corner. When sensitivity decreases (*d′* = 1), the bow become shallower, and recedes toward the chance line – the positive diagonal (i.e., *d′* = 0; chance performance). The HR and FAR therefore change in response to sensitivity. Ideally the HR would be 1 and the FAR would be 0 (MacMillan & Creelman, 2005), yet this is not possible due to the overlap in the distributions of signal and noise events.

[4] Use the standard equation for standardised z: $x-\mu/\sigma$.
[5] For the purpose of illustration the convention of assuming continuous densities has been adopted. Thus, smooth ROC curves are depicted.

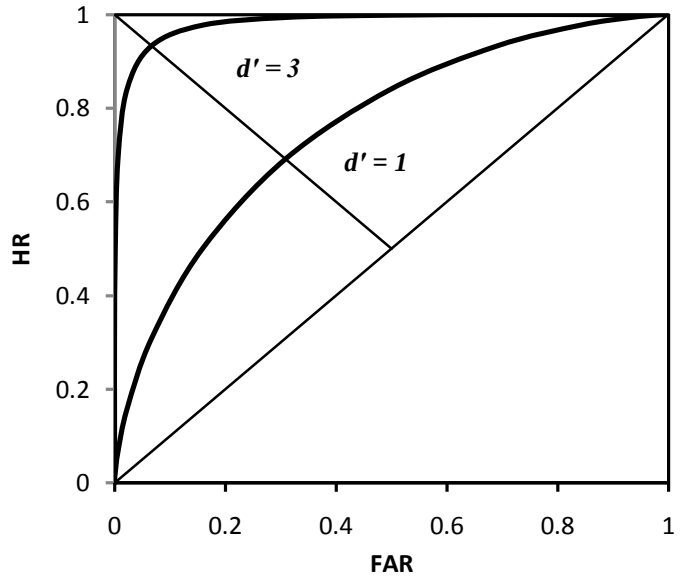


Figure 4: *The ROC curve has been plotted for all values along the decision axis for both d′ = 1 and d′ = 3 conditions. As sensitivity decreases the bow becomes shallower and recedes toward the chance line – the positive diagonal.*

Alternatively, plotting the *z*(HR) and *z*(FAR) produces a linear transformation referred to as the *z*ROC curve (Figure 5). *z*ROC curves are useful in making predictions about how much the FAR increases when the HR increases. Under the equal variances assumption the line increases as a function of the signal and noise variances. If both variances are equal at 1 then slope of the *z*ROC curve will increase at unity; i.e., have a slope of 1. Furthermore, the distance between the chance line and any point on the line with unit slope is a measure of *d′* (McMillan & Creelman, 2005).

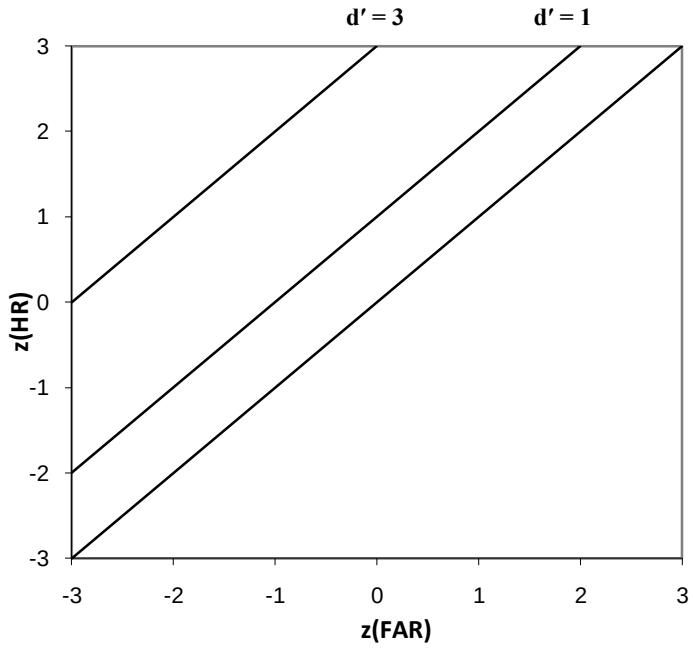


Figure 5: *The zROC curve has been plotted using the z(HR) and the z(FAR), producing a linear plot. The distance between the chance line and the plotted line for any value of z is equal to d′.*

Finally, the *area under the ROC curve* (AUC) is a measure of sensitivity, and is equal to the proportion of correct responses obtained by an unbiased observer in a 2AFC task (Verde et al., 2006). As the AUC diminishes, so does sensitivity. The AUC can be measured by a

variety of statistics, most commonly by A_z, A_g, and A' (MacMillan & Creelman, 2005; Verde et al., 2006). A_z is a parametric measure of the proportion based upon the standard normal curve, where:

$$A_z = \phi(d_a / \sqrt{2}) \qquad \text{(Eq. 5)}$$

This equation applies mostly to the equal variances cases. However, a non-parametric equivalent is A_g, which is the same as McNicol's (1972) *P(A)* measure, and approximates the AUC using geometric shapes created by casting lines down from each ROC curve point to the FAR axis, and then summing the trapezoid areas. The equation is expressed:

$$A_g = \frac{1}{2}\sum (FAR_{i+1} - FAR_i)(HR_{i+1} + HR_i) \qquad \text{(Eq. 6)}$$

Alternatively, another non-parametric measure can be used when only a single HR and FAR point is known. A' is a conservative estimate of the area under the curve which estimates the smallest possible AUC for that point:

$$A' = \frac{1}{2} + \frac{(HR - FAR)(1 + HR - FAR)}{4HR(1 - FAR)} \qquad If\ HR \geq FAR \qquad \text{(Eq. 7a)}$$

$$A' = \frac{1}{2} + \frac{(FAR - HR)(1 + FAR - HR)}{4FAR(1 - HR)} \qquad If\ HR \leq FAR \qquad \text{(Eq. 7b)}$$

Distinctions between Sensitivity Measures and ROC Functions

Essentially, the ROC function is generated by plotting all the HR and FAR values that correspond to the decreasing value of *x*. The points are thus generated from a fixed cut-off (*criterion*) point at specific *x* values. The curve then corresponds to the degree of overlap between the two distributions. This represents the second, and perhaps most contentious, assumption of SDT. Because the criterion is assumed not to fluctuate, this produces a curve that would be obtained by the *ideal observer* – the observer whose criterion is fixed at one location. Such an ROC curve is referred to an *implied* or *theoretical ROC* curve (MacMillan & Creelman, 2005). The curve reflects the theoretical index of sensitivity, d'_{th}, and indicates the sensitivity that would be obtained by the optimum, unbiased observer. In the present study the ROC curve and the corresponding d' values (shown in Figure 12, p. 39) are optimal estimates (thus d'_{th}) for the unbiased observer based on the discrete probability functions created for this present study.

For this reason it is assumed that given any fixed stimulus condition (e.g., $d' = 1$, or $d' = 3$), the real observer also adopts a fixed criterion position for the decision task, though it may

not always be optimal. If the observer assumes a conservative criterion (biased to responding '*N*'), fixing the criterion at this point will still produce a point that falls on the theoretical ROC curve. This is because response bias is independent of sensitivity, and no matter where the observer locates the criterion, as long as it is fixed their performance will mirror that of the ideal observer. An *empirical ROC* curve (MacMillan & Creelman, 2005) can be constructed by asking the observer to relocate their criterion to an alternative fixed position. Assuming a fixed position has been adopted, the observed point generated from the HR and FAR should fall on the theoretical ROC curve.

A second distinction can be made through the calculation of *d′* using the observed HR and FAR of the observer, by converting these values into *z* scores. In such a case sensitivity is expressed as d'_{ob} – the observed discrimination. This is calculated by using the HR and FAR obtained within the experiment and produces a datum point that can be plotted and compared to the theoretical ROC curve. Furthermore, the observed discriminability of the observer can be compared with that of the ideal observer by using the statistic, η (eta; Green & Swets, 1966):

$$\eta = (d'_{ob} / d'_{th})^2 \qquad \text{(Eq. 8)}$$

In summary, the sensory stage is primarily concerned with how well an observer can detect a particular stimulus. Sensitivity is described as the distance between the means of the two overlapping distributions. The distance is measured using *d′*. The smaller the *d′* value, the more the distributions overlap, indicating the task of discriminating *s* from *n* is relatively difficult. Furthermore, graphical representations of sensitivity have also been introduced (e.g., ROC and *z*ROC curves). Their shape informs us of the sensitivity, but also hints at the shapes of the underlying sensory distributions. Such plots are generated by passing a criterion along the decision axis. Finally, a distinction was drawn regarding the differences in sensitivity when it is calculated using theoretical versus observed HR and FARs. The role of the criterion in the observer's decision procedure is the focus of the following section.

The Decision Stage

Decisions are inherently statistical (more specifically, probabilistic) in nature, and are based upon perceived events. We almost always refer to the likelihood or probability that an event occurred without giving much thought to the mathematical constituents. In the present research stimuli can be randomly drawn from one of two states. Each variable has attached to it its own probability of occurring. These probabilities are referred to as *a priori* probabilities (Green & Swets, 1966) – the probability of a variable being drawn from one of the two states prior to the event occurring. The *a priori* probabilities are represented as *P(s)*

or *P(n)*; recall that *s* is a signal event, and *n* is a noise event; therefore, these probabilities are the probabilities that a signal or noise event will occur.

The observer is presented with some evidence and must use it to decide whether it represents evidence for an *s* or *n* event occurring. Such evidence can supplement the *a priori* probabilities to better the decision process. This results in conditional *a posteriori* probabilities - the probability of each state variable occurring conditional upon the event that has just occurred (Green & Swets, 1966). For the probability that an observer responds '*S'*, given a particular stimulus event, the *a posteriori* probabilities become *P(S | s)* and *P(S | n)* – the estimated HR and FAR. Decisions can be summarised as a conditional probability that take into consideration both *a priori* and *a posteriori* probabilities. When formulated it is expressed by Bayes' theorem:

$$P(S \mid s) = \frac{P(S)P(s \mid S)}{P(s)} \qquad \text{(Eq. 9)}$$

Ultimately, decisions reflect the propensity for an observer to respond in a particular way, given the evidence available. Detection theory suggests a method for assessing how biased an observer is when faced with a decision task – i.e., do they favour one state over the other? Recall that in SDT it is assumed that the observer will select a fixed location along the decision axis to place their criterion. The *criterion (k*[6]*)* reflects a subjective cut off point that the observer adopts in order to decide when a stimulus is sufficiently high in magnitude to warrant an '*S'* response (Stanislaw & Todorov, 1999). Typically the decision rule will take the form $x \leq k$ = *'N'* and $x > k$ = *'S'*. Consequently, any stimulus equal to or falling to the left of *k* will prompt an '*N'* response, while all stimuli to the right will produce an '*S'* response (Figure 6).

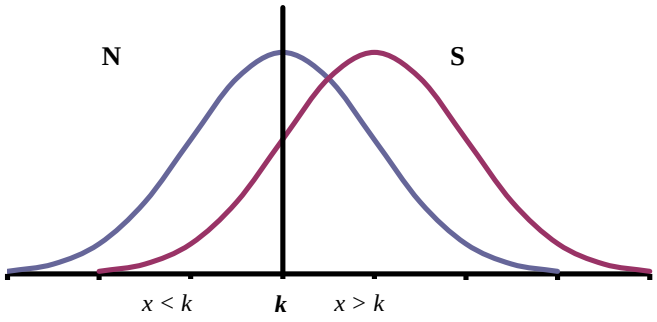


Figure 6: *The criterion splits the decision axis into 'S' and 'N' responses. Any stimulus magnitude equal to or falling to the left of k will produce an 'N' response, while any stimulus falling to the right of k will produce an 'S' response.*

[6] The symbol *k* has been adopted for the value of the criterion so as not to confuse it with the response bias measure, *c*.

Ideally, *k* would be positioned where the HR is maximised, and the FAR minimised. This position is known as the *optimal criterion* (Green & Swets, 1966; McNicol, 1972), and is the point where the two distributions cross over. In Figure 6 the criterion is to the left of where the distributions cross, meaning that the criterion is biased to some degree. Any criterion provides two pieces of information: (a) the position where the observer decides the stimulus reaches sufficient magnitude to respond '*S'*; and (b) the probability of the observer responding a particular way – the *response bias*. A common measure of response bias is the *likelihood ratio* (β), if continuous variables are assumed. This can be investigated by taking the probability densities of the *s* and *n* distributions where *k* cuts the two distributions. The densities at the criterion point are denoted *f(x)* (MacMillan & Creelman, 2005) for each distribution. The likelihood ratio is expressed mathematically as:

$$l(x) = \beta = \frac{f(x \mid s)}{f(x \mid n)} \qquad \text{(Eq. 10)}$$

for *s* and *n* distributions that are normally distributed with equal variances.

If $\beta = 1$ then the observer is considered to be unbiased. In this instance *k* has been placed right in the middle of the two distributions where they cross over – the optimal criterion location (see Figure 7, criterion *b*). As the criterion is moved leftward – becoming more *lax* (criterion *a*) – the value of β decreases and reflects a bias toward '*S'* responses ($\beta = 0.63$). Conversely, if β is shifted rightward – becoming *stricter* (criterion *c*) – its value increases and reflects a bias toward '*N'* responses ($\beta = 1.58$).

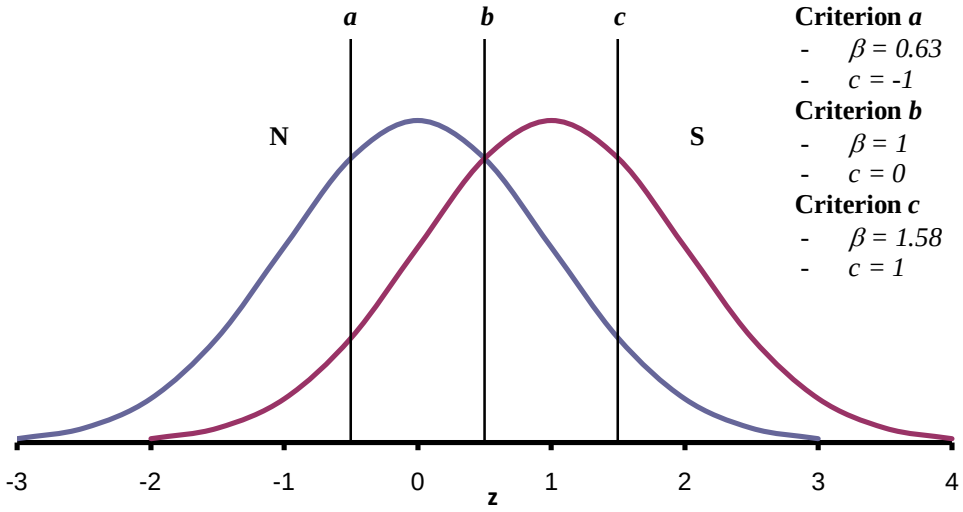


Figure 7: *Varying the criterion location yields different β values. Criterion (a) assumes a lax position where β = 0.63; criterion (b) is optimal where β = 1; and criterion (c) assumes a strict position where β = 1.58. Bias can also be measured using c. Measures of c have also been provided (see Eq. 11).*

A more intuitive method for investigating response bias is to again begin by converting the HR and FAR into *z* scores. *Criterion location (c)* measures the distance between where the observer has located their criterion in relation to the optimal criterion. Using Figure 7, the optimum criterion (β = 1) is set at *z* = 0.5. Criterion (*a)* is situated at *z* = -0.5, 1 S.D. below the optimum criterion, thus *c* = -1. *c* is mathematically expressed as (MacMillan & Creelman, 2005):

$$c = -\frac{1}{2}[z(HR) + z(FAR)] \qquad \text{(Eq. 11)}$$

where

$$z(HR) = z(\mu_s) - c \qquad \text{(Eq. 12a)}$$

$$z(FAR) = z(\mu_n) - c \qquad \text{(Eq. 12b)}$$

The optimal criterion position is *c* = 0 ($c = 0 \Leftrightarrow \beta = 1$) for the unbiased observer, and deviations from this point represent the degree of response bias. As *c* shifts left of the optimal point it becomes negative and represents a bias towards '*S'* responses (criterion *a*: *c* = -1), whereas if *c* is moved to the right it assumes a positive number and reflects a bias toward '*N'* responses (criterion *c*: *c* = 1). Furthermore, plotting an observer's HR and FAR can also illustrate the degree of response bias. ROC curves identify the optimal criterion by the negative diagonal line (see Figure 8). If the observer's point falls on this line then an optimal criterion has been adopted (criterion *b*). If the point falls either side of optimal this reflects a response bias. The ROC plot below indicates the relative points for the three criteria in Figure 7. Note how a lax criterion (shifting *k* left) produces a point to the right of the negative diagonal (criterion *a*) on the ROC curve. The reverse occurs for a strict criterion (criterion *c*).

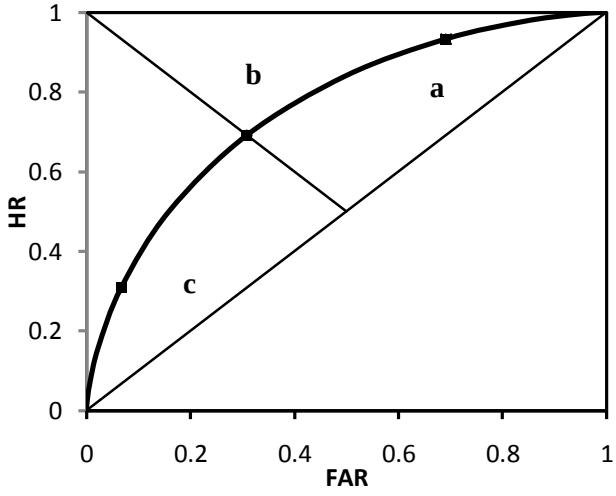


Figure 8: *Plotting the HR and FAR relative to the criterion location produces a point on the ROC curve that illustrates the degree of observer response bias. Criterion (a) reflects a lax position whereas criterion (c) reflects a strict position. Criterion (b) is located at the optimal point.. This ROC curve presented here is the result of normally distributed signal and noise densities with equal variances.*

This section has reviewed the role of the criterion in response bias. Two statistics are commonly used to measure response bias: c and β. SDT assumes that once a criterion is adopted it remains fixed for the duration of the experimental session. As will be shown in the following sections, this assumption has been met with much criticism, and reflects one of the major issues under investigation in the present research. The following sections review the criticisms levelled at SDT for assuming that the criterion remains fixed (no variability along the decision axis).

Chapter II

Criticisms of SDT and the Issue of Criterion Variance

SDT explicitly takes into account the decision processes involved when an observer is involved in a discrimination task. The basic theory makes two simplifying assumptions:

(a) the perceptual impressions of the stimuli over trials can be represented by Gaussian signal and signal + noise distributions having equal variances;
(b) within any condition a criterion is placed somewhere along the decision axis, occupying a fixed and unvarying position that splits the decision space into responses based upon stimulus strength.

Both assumptions have been contested; however, it is with assumption (b) that this present study concerns itself.

The Evidence for Criterion Variance

The fixed criterion assumption has been contested since the early 1960s (e.g., Speeth & Matthews, 1961), yet it was not until Tanner (1961) investigated the implications of non-zero criterion variance in signal detection that researchers started to investigate the issue more systematically. Until then, criterion variance was dealt with by ignoring it, or assimilating the criterion noise with perceptual noise (McNicol, 1975; Mueller & Weidemann, 2008). This practice was particularly common within simple yes/no designs (Triesman & Faulkner, 1985). Effectively, no specific attention was given to decision noise at all. Rating tasks have received much more attention with regard to criterion fluctuation and indeed the vast majority of the literature focuses on such tasks.

The reason for this is that in a rating task the observer is effectively adopting multiple criteria. Rather than asking observers to make a simple binary judgement, they categorically rate their confidence about their decision. For example, an observer may be asked to rate an event into the following categories: *certain signal; probable signal; probable noise; certain noise* (McNicol, 1972). Any number of categories could theoretically be used. The criteria act as boundaries between the adjacent categories (see Figure 9b, on the next page); the more categories there are, the more criteria there are, and the harder it becomes to maintain the criteria locations.

The traditional SDT assumption is that the criterion has zero variance, and is a specific type of probability function called a Dirac delta function (Rosner & Kochanski, 2009). SDT typically represent this function as a single vertical line in the decision space (e.g., see Figure 6), reflecting the mean of the criterion density, thus:

$$f(x;\mu,\sigma^2 = 0) = \delta(x-\mu) \qquad \text{(Eq. 13)}$$

Under the Dirac assumption the *z*(FAR) and *z*(HR) are calculated:

$$z(HR) = \frac{c-\mu_s}{\sigma_s} \qquad \text{(Eq. 14a[7])}$$

$$z(FAR) = \frac{c-\mu_n}{\sigma_n} \qquad \text{(Eq. 14b)}$$

The calculation simply standardises the distance of the criterion from the distributional mean, with the denominator only incorporating the standard deviation of the sensory distribution. This means the computation of *d′* is free from any criterion variation. However, criterion variance, like any other random variable, can be conceptualised as a Gaussian random variable (see Figure 9 below; McNicol, 1972; Wickelgren, 1968).

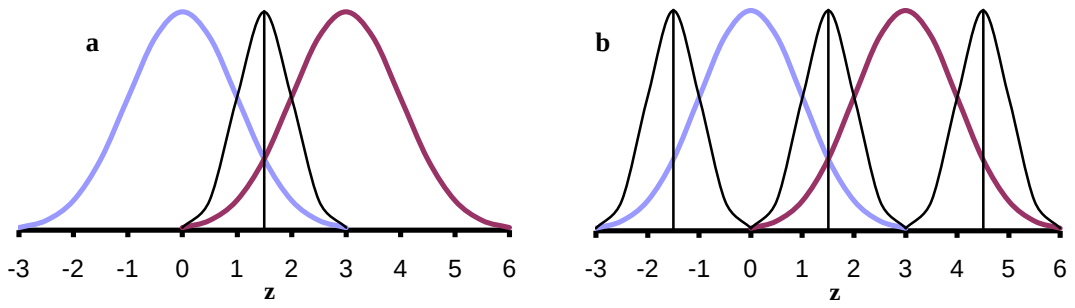


Figure 9: *Examples of binary (a) and rating (b) style tasks with regard to criterion and associated variability around the mean criterion position. Rating style tasks have multiple criteria, thus more variance is evident.*

Like all variables the criterion shifts around, and thus variance is created. The effect of criterion variance can be examined by allowing for the variance of the criterion to be accounted for in the computation of *z*(HR) and *z*(FAR). Such expressions are commonly referred to as Thurstone's Law of Categorical Judgement (Thurstone, 1927b, McNicol, 1972), and allow for non-zero criterion variance; where

$$z(HR) = \frac{\mu_c-\mu_s}{\sqrt{\sigma_c^2+\sigma_s^2}} \qquad \text{(Eq. 15a)}$$

$$z(FAR) = \frac{\mu_c-\mu_n}{\sqrt{\sigma_c^2+\sigma_n^2}} \qquad \text{(Eq. 15b)}$$

and σ^2_c is criterion variance.

[7] Under the equal variance assumption, the denominator can be cancelled out.

This additional variance must then be accommodated in the calculation of *d′*, where

$$d' = \left[\frac{\mu_c - \mu_s}{\sqrt{\sigma_c^2 + \sigma_s^2}} \right] - \left[\frac{\mu_c - \mu_n}{\sqrt{\sigma_c^2 + \sigma_n^2}} \right] \qquad \text{(Eq. 16)}$$

The inclusion of the criterion variance value in the denominators of Equation 16 has the effect of spuriously reducing the *d′* value, effectively compromising estimates of sensitivity. Such effects are most readily recorded within rating tasks where multiple criteria produce greater variance. However, the effects are still present within binary tasks, as this present study will illustrate. Rosner and Kochanski (2009) evaluated the validity of the existing Thurstonian equation and found it to be flawed. The law is based upon the assumption that multiple criteria hold a serial order that remains unchanged (for example see Figure 9b). This is referred to as the *absolute order constraint* (Triesman & Faulkner, 1985) and supposes that criteria maintain their distance apart and vary when changes are made. Thus, each criterion mean acts as a boundary for the adjacent rating category, holding its general position but not overlapping with the adjacent distribution.

Rosner and Kochanski (2009) found that Thurstone's law, when applied to rating tasks, produced negative theoretical probabilities. The law seemingly violates a fundamental principle of independent samples. If the criteria were independent then they are decreasingly likely to move together in an ordered fashion; rather they would vary independently, interchange locations, and overlap. Triesman and Faulkner (1985) provided some evidence for interchanging criteria location, though their results were not conclusive. Rosner and Kochanski quantified this effect, leading to a corrected version of the law that better suits rating style detection tasks (for a comprehensive review see Rosner & Kochanski, 2009). However, the pertinent issue is that the addition of criterion variance can inflate the variances of both the signal and noise distributions, effectively pushing the distributions closer together, and altering the degree of overlap (McNicol, 1972), thereby distorting results.

For example, Benjamin, Diaz, and Wee (2009) have illustrated how criterion variance can distort *z*ROC functions in recognition memory. Criterion variance, when coupled with unequal variances in the underlying sensory distributions, has an interactive effect. If $\sigma^2_s > 1$ then an increase in criterion variance will increase the slope. Conversely if $\sigma^2_s < 1$ then an increase in criterion variance will decrease the slope.

Lopez-Bascuas (2008) also documents distortions to linear ROC plots in the presence of increased criterion variance using speech and non-speech signal densities. While the non-speech signals appeared to conform to the Gaussian assumption, speech signals produced

aberrant ROC curvees, indicating non-Gaussian densities. However, it was hypothesised that non-zero criterion variances could account for this. Linear ROC curve transformations departed from linearity quite significantly, confirming the presence of criterion variation. Varying deflections in the lines also suggested that the criterion variance was not equal for all criterion locations in the rating task. Analogous to signal and noise densities, several independent criteria may not possess equal variances.

It has generally been accepted that trying to separate criterion noise from sensory noise is a difficult task (McNicol, 1972). Rosner and Kochanski's (2009) corrected law of categorical judgment may allow researchers to measure the amount of noise added by criterion variation, yet the assimilation of criterion and sensory noise does not address the source of the internal noise. Rather it measures the net effect on the estimates of the observer's sensitivity. Yet seeing that the addition of criterion variance can create spurious results, it is necessary to identify the mechanism through which this noise is introduced into the system. Specficially, what is happening in order to create this variance? Several models and hypotheses have been forwarded to account for criterion variance.

Models of Criterion Variance

Larkin (1971) provided one of the early models of criterion variance. Given the inherent probabilistic nature of decisions, the model conceptualised criterion variance as a "random walk" - an error correction model (McNicol, 1975) that suggests observers slowly shift their criterion after errors are made in such a way that reduces the probability of the same error occurring again. Larkin demonstrated that the probabilities associated with the sensory stimuli affected the decision stage, but not the sensory stage (see Chapter 1; recall that decisions relate to the probability of a stimulus event occurring). By manipulating the probabilities of the sensory stimuli (the frequency with which a stimulus is presented during the course of the experiment), criterion shifts were induced, affecting only the selection of the response alternative, and leaving the sensory stage unaffected. When uncertainty exists about which event has occurred (either signal or noise), observers shift their criterion, trying to eliminate uncertainty.

Essentially, this creates a source of noise over and above sensory noise, that is, decision noise. As the criterion is shifted in response to stimuli frequencies, the criterion's variance generates a distribution around the mean criterion position. The parameters of the criterion's distribution are then conditional upon the event probabilities, meaning that the event probabilities can be used to estimate the criterion distribution and predict the most likely response (Larkin, 1971).

It has been accepted within the SDT literature that rating tasks produce greater total criterion variance as a result of maintaining multiple criteria. The number of categories that an observer uses is directly related to the number of criteria that must be maintained. Maintaining several criteria creates increased cognitive load, which ultimately affects performance by destabilising the criteria. Though Clark and Mehl (1973) hypothesised that rating style tasks can produce lower *d′* values, they eschewed any suggestion that increased cognitive load precipitated criterion variance. However, recent research has shown that cognitive factors do impinge upon decision-making (e.g., Benjamin et al., 2009).

Triesman and Faulkner (1984a, 1984b) provide a model of criterion variance that synthesises both statistical and cognitive elements. The theory rests on sequential dependencies – the mapping of past experiences and signal frequencies, such that the position of the criterion on trial n_{+1} is dependent on information gained from trial *n*. This represents a largely probabilistic approach to criterion setting, but also requires past events and altered criterion values being recorded as memory traces. The development of the model involved the review of three existing models for sequential effects.

The *Linear Additive Learning* (LAL) model assumes that an observer will pick a reference criterion location, and alter the criterion relative to this point as a function of the response on the previous trial. During the inter-trial intervals the criterion has a natural tendency to drift or decay back to this reference point at a constant rate. The *Exponential Additive Learning* (EAL) model works in much the same way as the LAL model, the only difference being that criterion decay is exponential rather than constant. The additive models simply use the value of the criterion at any one time to reflect the effect of past events. The criterion is shifted due to responses on previous trials; thus, it is shifted to maximise certain responses given the past events. The third model is the *Independent Trace* (IT) model. The fundamental premise of this model is similar to that of the learrning models described above insofar as the criterion decays back toward the reference point at a constant rate, yet the IT model posits that each shift is individually recorded. This implies that each criterion shift can be linked to a series of memory traces; recorded events that trace the movement of the criterion (see Triesman & Faulkner, 1984a, for complete review).

The basic difference between the models is the way the past events are represented. The additive models simply use the value of the criterion at any one time to reflect the effect of past events, whereas the IT model places emphasis on the memory traces that have lead to the current position.

It makes intuitive sense that a system would utilise all sources of available information to enhance decisions, and for this reason the IT model set the foundation from which Triesman

and Faulkner (1984a) developed their model. The result was a theory of criterion setting that uses sequential dependencies to maintain the criterion at the optimal level. The model uses a long-term criterion setting process, and short-term processes that adjust the criteria. It incorporates two stages: (a) establishing a reference criterion and stabilising it at a reference location; and (b) using systems that make fine adjustments in response to environmental/experiential changes. These two short-term systems are referred to as the *tracking* and *stabilising* system.

The tracking system adjusts to changes in the external world. For example, recent detections will tend to relax the criterion, though increasing the risk for false alarms. Conversely, recent rejections will cause the criterion to become stricter, causing the system to adopt a more conservative approach. The stabilising system, like the tracking system, uses incoming information to refine the location, yet in a somewhat historical fashion. It does so by recalling the number of detections made over past encounters. For example, if a sustained series of signal responses is observed the criterion is regarded as too lax and is moved rightward (moving the criterion rightward along the decision axis causes the criterion to become stricter, moving the criterion leftward relaxes the criterion); a sustained series of noise responses causes the criterion to relax. Memories of past events, and past criterion locations are essential resources and reflect independent sources of information that influence each criterion shift.

These mechanisms push the criterion closer to an optimal position (Schoeffler, 1965), stabilising it as best as possible until further evidence suggests a move is necessary. The model states that signal frequency, past experience, and memory all contribute to criterion shifts. Additionally, the two short-term processes may function in unison. Triesman and Faulkner (1984a, 1984b) have illustrated that these internal processes serve as adaptive behaviours that attune our decision processes to changes in the environment, and occur over and above that of sensory effects. Our environment is dynamic and continually changing – signals come and go. Thus, vigilance and monitoring are effective strategies which aide the decision process, and facilitate the response of the criterion to such dynamic factors.

Brown and Steyvers (2005) investigated the effects of dynamic factors on criterion shifts. Using a lexical decision task, targets and distracters were manipulated such that their difficulty varied over trials within the same condition. This is analogous to altering the degree of overlap between the signal and noise distributions continuously, resulting in no fixed level of difficulty within a condition. Ultimately, this forces observers to alter their behaviour within the condition, and is known as a *context effect*. More specifically they looked at a type of context effect called a *mirror effect*. In SDT, the mirror effect defines

changes in both the HR and FAR as task difficulty alters. Assuming a fixed, optimal criterion, as the distracter distribution (noise) is moved closer toward the target distribution (signal), the shift will alter the FAR, but leave the HR unchanged (recall the FAR is the area under the noise curve to the right of the criterion location). However, mirror effects imply that the HR must also change when a change in difficulty is made. How this occurs is that when the noise distribution is shifted, the original criterion location no longer sits at the point where the two distributions cross; that is, it was no longer optimal. The original criterion thus becomes redundant, and must be relocated to the new optimal position. Consequently, this relocation alters the HR, demonstrating that changes in HR are accounted for by this shift in criterion placement.

In dynamic settings where task difficulty is frequently changing, the criterion must also shift in response to these changes. Brown and Steyvers (2005) were interested in how quickly these changes occur. Larkin (1971) asserted that the criterion shifts slowly in response to error, and accordingly Brown and Steyvers' study hypothesised that there would be a lag in criterion shift as the observer adjusts to the change. This may be somewhat analogous to the stabilising system proposed by Triesman and Faulkner (1984a). For example, if the noise distribution was moved closer to the signal, the observer, whose criterion is stabilised at its original optimal point, would start to respond '*S*' more often due to a lax criterion position. As the observer tracks their responses, a series of '*S*' responses would be noted, suggesting that the criterion may be overly lax. The result would typically be a shift rightward to a more optimal position. Evidently, no additional feedback (e.g., indicating when the observer was correct or not) was delivered to the observers, generating reliance on purely environmental cues (e.g., past responses, sequential dependencies) in the updating of their criterion. Brown and Steyvers detected a lag of approximately 14 trials in the updating of the criterion. However, they did not explore the mechanism that might possibly underlie the shift, although they allude to the criterion-setting model of Triesman and Faulkner (1984a, 1984b). Sequential dependencies offer a plausible process through which the observer updates the criterion in response to environmental cues.

One caveat in regard to Brown and Steyvers' (2005) study is the assumption that the observers were able to maintain an optimal criterion. How this was possible is not discussed, though it seems unlikely that it would have been fixed at optimal, if fixed at all. If sequential dependencies were at play then there would be a drift back and forth from the reference criterion. Given that no additional feedback was employed, it seems likely that this is how the observers were maintaining their criterion, and how they updated it. This point is not addressed in their report. Furthermore, if Brown and Steyvers are to draw parallels with the research of Triesman and Faulkner (1984b), then they have to relinquish the optimality

assumption. However, the lag parameter is useful in this instance as a measure of criterion change in response to environmental factors, for example, how many rejections or detections are needed before the criterion is shifted, and provides converging evidence for the cognitive processes underpinning criterion variation. Interestingly, Brown and Steyvers conclude that the lag in shifts can potentially be remedied by the addition of feedback. However, as the present study shows, the type of feedback provided must be taken into account.

One of the most succinct models of criterion variance has been compiled by Mueller and Weidemann (2008). The model centres on the principle that signal and noise prior probabilities affect the decision processes, and not sensory aspects, a principle first introduced by Larkin (1971). Sensory variables are mediated by the either the strength of the signal, or by the degree of overlap between the distributions; therefore, the stimulus probabilities can be manipulated without affecting either of these sensory components (Mueller and Weidemann, 2008). Furthermore, if the fixed criterion assumption holds, then manipulating the stimulus probabilities would preclude an effect on the decision process, as no criterion fluctuation would be induced. Under such conditions the empirical ROC curve would be identical to the theoretical ROC curve, resulting in the observed HR and FAR values falling on the theoretical ROC curve (see Chapter 1).

Balakrishnan (1998a, 1998b, 1999) cites violations of the assumption that ROC plots generated using altered stimulus probabilities produce similar curves. He argues that they do in fact change shape under such conditions, yet attributes the distortion to changes in the perceptual distributions, not decision noise. He used rating-style tasks under varying stimulus probabilities to generate ROC curves. While there is some evidence to suggest that rating-type tasks induce additional criterion shifts in response to signal probability, Balakrishnan (1999) found that rating-based measures show no shift in criterion when stimulus probabilities were manipulated. His evidence relied on a function called Ur_k – which measures the divergence between the cumulative density probabilities for both the signal and noise distributions at criterion point k (Balakrishnan, 1998b). When the respective HR and FAR points are plotted a ROC-type curve is generated that peaks at the central confidence point – the optimal criterion location. As stimulus frequency is manipulated, the peak is assumed to shift as a function of criterion fluctuation. However, Balakrishnan observed no such shifts in the peak. This led him to hypothesise that confidence criteria do not change in response to stimulus probability manipulations, remain fixed at an equal-likelihood point, and therefore, the distortions must be sourced at the sensory stage, not at the decision stage (Balakrishnan & MacDonald, 2002). However, both Larkin (1971) and Triesman and Faulkner (1984b, 1984b, 1985) demonstrate that decision criteria do shift in

response to stimulus probability, and leave sensitivity intact; the complete antithesis of what Balakrishnan is postulating.

Mueller and Weidemann (2008) attribute the violations that Balakrishnan (1998a, 1998b, 1999) had observed to two factors: decision noise and the confidence rating procedure. They dichotomise decision noise into *classification noise* – the mapping of internally represented percepts to binary decisions; and *confidence noise* – the mapping of internally represented percepts to classification responses. The effects of each type of noise may differentially affect the ROC functions, with exacerbated distortions occurring when confidence noise is greater than classification noise. They further hypothesised that the lack of evidence for criterion shifts in the Balakrishnan studies can be addressed by the presence of increased confidence noise, which, when greater than classification noise, can mask shifts in the peak Ur_k function. Their model posits that both stimulus probability and type of decision noise can affect criterion placement.

They formulated an extension to the traditional SDT model called the *Decision Noise Model* (*DNM*). The model assumes that decision noise can mask criterion shifts – especially if confidence noise is greater than classification noise. When decision noise is low, true shifts in the criterion can be detected using Ur_k. They demonstrated that equal levels of confidence and classification noise produced identical ROC plots, yet when confidence noise was increased the corresponding ROC curve altered, causing the confidence ROC curve to shift. Lower levels of decision noise were indeed conducive to detection of criterion shifts using Ur_k. However, when increased decision noise was introduced it appeared to stabilise the peak of the Ur_k function at the medial confidence point. This gave the illusion of a zero shift; thus confirming the hypothesis that decisional noise masks shifts in the criterion.

Mueller and Weidemann (2008) used ROC curves to assess the effects of stimulus frequency and decision noise. In their study the theoretical ROC curve was labelled as the *distal stimulus ROC* curve (DS-ROC). This measures the relationship between stimulus intensity and the internal representation. Like traditional SDT it maps the overall sensitivity of the condition. The empirical ROC curve was called a *confidence ROC* curve (C-ROC). This measures the observer's overt response, using rating-based measures similar to the example discussed earlier (see p. 17)

Noise in either stage should create discrepancies in the curves– thus perceptual noise should distort the theoretical DS-ROC while decisional noise should affect the C-ROC. Mueller and Weidemann (2008) hypothesised that if the C-ROC deviates significantly while the DS-ROC remains unchanged across manipulations of probability, then it is decision noise that is at play. Furthermore, they analysed the deviations of the C-ROC from DS-ROC curves for both

confidence and classification noise. When the equal variance assumption is met, equal decisional noise results in identical C-ROC and DS-ROC plots. Even though decision noise is present in these instances, the results mimic traditional SDT analysis whereby decision noise is incorporated into the measure of sensitivity; thus, the plots are identical. As confidence noise was increased the C-ROC function started to change, while the DS-ROC curve remained intact. This implies that the observed deviations in the ROC curve are attributable to decisional noise, not perceptual noise. These results lend further weight to the theories of Larkin (1971) and Triesman and Faulkner (1984a, 1984b).

Further analysis indicated that had the perceptual distribution been affected by stimulus probability manipulations, then the changes in the DS-ROC would have also been mapped onto the C-ROC, confounding the degree to which the decision noise was affecting the observer. However, such an effect was not observed, leaving the perceptual distributions unaffected and isolating decision noise as the only factor accountable for ROC curve deviations. The DS-ROC curve remained unchanged across all conditions. The ROC curve analysis the authors used met the equal variance assumption, thus, no analysis of the effects of decision noise under the unequal variances assumption was conducted. If the variances of the signal and noise distributions are unequal, this may also add noise into the system, and further distort the ROC curve shape. To address this, Mueller and Weidemann (2008) showed that while unequal variances indeed affected the ROC curve, the addition of decisional noise induced shifts over and above those created by unequal perceptual distributions. More simply, decision noise adds further noise into the system.

The above results imply that confidence-based tasks induce an inordinate amount of criterion variance. Mueller and Weidemann (2008) noted that the standard deviation of the confidence criteria was approximately three times as large as that of the classification criteria. This suggests that the confidence-based tasks may be less reliable than classification tasks, an intriguing notion as the bulk of SDT literature focuses on the confidence/rating task. Additionally, confidence ratings appeared to induce a higher proportion of sequential dependencies, creating additional noise through cognitive mechanisms. While adjustments to the criterion placement potentially benefit the observer, it does so at the expense of additional noise as the observer tracks past events. Cognitive processes further impinge on the inherent probabilistic nature of decision tasks, perpetuating criterion fluctuation. It becomes increasingly clear that many of the discrepancies in the estimation of observer performance rest at the decisional stage, and need systematic assessment. Though noise is expected to exist within the binary decision context, the inflation in decision noise within rating tasks converges with research that reports depressed d' values for rating style tasks (e.g., Clark & Mehl, 1973; Schoeffler, 1965).

In light of studies such as those described earlier, criterion variance can no longer be ignored in SDT analysis. While the research discussed so far focuses on rating tasks, there is every reason to believe that criterion variance has equally deleterious effects in binary decision tasks. Even a single criterion density can inflate the variance of the signal and noise densities (see Thurstone's Law, Chapter 1). Not only is the binary discrimination task the most simple to investigate criterion variance, but this kind of task is very frequently used. The present study aimed to find out by how much a discriminability estimate such as d' is affected by criterion variance, and under what conditions this variance alters.

Criterion variation may be mitigated to some extent through the introduction of feedback. Brown and Steyvers (2005) alluded to the fact that feedback (indicating correct or incorrect decisions) may speed up the lag in updating the criterion; in fact, feedback is commonplace within SDT experiments. Despite its pervasiveness within SDT research designs, the effect feedback has on estimates of sensitivity has not been investigated in any great depth. The present study is one of only a few studies that has systematically evaluated the role of feedback (knowledge of results) on criterion variability.

Chapter III

The Role of Knowledge of Results in Criterion Variance

Intuitively, any form of feedback would serve to enhance one's performance, informing of error and instigating change. However, some SDT studies have shown that feedback does not always improve performance. These counterintuitive findings have led to, albeit limited, investigations into the effects of feedback on performance. Feedback involves the observer being told whether a correct decision has been made upon their response. Furthermore, feedback may be qualitative (e.g., "correct" or "incorrect"), or quantitative (e.g., "that was two seconds too fast"; Salmoni, Schmidt, & Walter, 1984). Such feedback is also referred to as knowledge of results (KR), which from this point will be the term used.

Traditionally, a psychophysical task might include KR or not. KR typically was not considered as an independent variable; therefore, its potential confounding effects were never entertained. In many instances where criterion variability has been suspected, the role of KR has not been investigated. Research lends support to the idea that the criterion will drift naturally with no KR (see Triesman and Faulkner, 1984a, for a review on criterion drift models), yet the additional effects of KR on this drift have not been investigated to any great extent. However, Salmoni et al. (1984) suggested that the frequency and precision of the KR have an effect on the task, such that altering aspects of the KR may have differential effects on the observer. The pressing issue surrounding KR is the type that is provided. Very few studies have investigated such effects; consequently the literature is thin. This present investigation shows that KR type has differential effects upon estimates of classification accuracy through affecting criterion variation.

Evidence for Knowledge of Results and the Effects on Criteria

Schoeffler (1965) implicated KR in the suppression of d' estimates. The suppression in sensitivity is thought to be mediated by KR continually causing shifts in the criterion. Schoeffler argues that when an observer is informed that they have made a false alarm, their criterion is likely to shift rightward (become stricter) in order to reduce the probability of this error occurring again. Conversely, when the observer is informed that they made a miss, they are likely to loosen their criterion and become slightly laxer. Larkin's (1971) correction model bears striking similarities to Schoeffler's hypothesis. Therefore, the criterion reacts to KR given explicitly much the same way as environmental factors alter the criterion, and sequential effects can be induced. Since Schoeffler's theory, a modest body of research has accumulated that has looked at the effects of KR on task performance.

Clark and Greenberg (1971) investigated the effects of KR on recognition memory tasks and response criterion (l_x). The effects of stress were also under investigation, and whether an interactive effect existed with KR in lowering measures of sensitivity. Stress is thought to influence criterion placement in its own right. Higher stress levels are assumed to underpin a more lax criterion placement, mediated by fear of missing something important. Yet lower levels of stress suppose greater confidence, and so a stricter criterion is adopted.

Clark and Greenberg (1971) used a simple 2x2 design (Stress: high or low; KR: present, absent). As hypothesised, KR suppressed *d′* across both high and low stress conditions. KR also appeared to affect initial criterion placement, with those in the KR condition setting stricter criteria but relaxing them over the course of the trials as KR was provided. Conversely, those who received no KR set a laxer criterion, only to become stricter over the trials. Schoeffler (1965) hypothesised that the criterion would shift to a near optimal position with KR, and the evidence here lends support to this assertion.

Unfortunately, like many methods sections of the day, no explicit mention of the KR was made by Clark and Greenberg (1971). This is not surprising, given that KR was seen as a fixed aspect of the design itself, needing no special mention. KR is usually provided on a trial by trial basis, indicating whether the response made was correct or not. This type of qualitative feedback is known as trial-by-trial knowledge of results for events ($TTKR_e$).

The criterion will tend to move toward a region that reflects the observer's decision aim, for example, maximise expected values or maximise correct responses (Green & Swets, 1966). Once in this region the criterion will fluctuate around a mean position from trial-to-trial (McNicol, 1975). However, the degree and magnitude of this fluctuation is problematic, and can be shown to correlate with the type of KR provided. The very nature of SDT tasks ensures that the evidence presented will not allow all decisions to be made with certainty. Therefore, the KR regarding these stimuli will be contradictory in many instances, because certain types of evidence can indicate that either a signal or a noise event has occurred. If such TTKR is reliably used, it should be apparent that the continual updating of the criterion in response to events KR increases the magnitude of fluctuation, thus spuriously lowering measures of sensitivity.

McNicol (1975) investigated the effects of biased KR on absolute judgements of loudness. The feedback was biased so that the observers received feedback that pushed their response bias toward either a lax or strict position. The study indicated that observers' responses varied on a trial-by-trial basis in an attempt to follow the feedback. This can be attributed to shifts in the criterion location. Therefore, the *d′* value was suppressed.

Ryan and Fritz (2007) investigated the effects of KR and erroneous KR on performance in a mental timing task. Erroneous KR refers to feedback that is inaccurate to varying degrees; therefore, the feedback may be reliable 70%, 60%, or 50%, etc., of the time. Though not a traditional SDT task, the authors noted that both reliable KR and erroneous KR affected performance, memory, and decision thresholds, with erroneous KR producing further decrements in performance. Ryan and Fritz acknowledged that all KR pushed decision thresholds to near midpoint (central criterion location) and stabilised performance, synonymous with effects seen in previous research (e.g., Clark and Greenberg, 1971; Larkin, 1971; McNicol, 1975; and Shoeffler, 1965), but noted that erroneous KR induced greater fluctuation in the criterion location.

Both McNicol (1975) and Ryan and Fritz (2007) report shifts in the criterion when KR was reliable and events based, that is, the KR was veridical, indicating that a signal had occurred when one was presented, and the same for a noise event. Conversely, Han and Dobbins (2008) report evidence that fully correct KR had little effect on the criterion location in recognition memory. The authors document minimal criterion variance for those observers who received KR when compared to those who received no KR at all. In fact the data points for both conditions fell on the same ROC curve irrespective of whether KR was present or not.

However, once biased KR was introduced an increase in criterion variance was observed. Additionally, changes in the ROC curve shape were observed, suggesting that criterion fluctuation was displacing the function. Han and Dobbins (2008) held sensitivity constant so as not to confound the decision noise with that of sensory noise. The authors trained the observers to adopt either a strict or a lax criterion, and observed criterion shifts in both conditions, with ROC curve points falling away from the curve. These changes in ROC curve, while indicative of the unequal variances inherent in recognition tasks, also reflect variance that can be ascribed to a fluctuating criterion.

Criterion fluctuation can be reliably attributed to memory processes (e.g., Benjamin et al., 2009). Additionally, KR can interact with memory traces to induce criterion shifts. Ell, Ing, and Maddox (2009) investigated the effects of delayed KR in a rule-based category learning task. Essentially, the participants had to learn and maintain criterion locations that helped them assign specific patterns to their correct category. KR was provided to enhance learning of the appropriate categorisation. The number of criteria to learn and the length of KR delay (either immediate or delayed by 5 seconds) were manipulated in the experiment. What Ell et al. found was an interaction between number of criteria and KR delay, observing exacerbated drift as a function of increased criterion numbers and delayed KR. Working memory was

thought to have been taxed, affecting learning through expending all working memory capacity.

$TTKR_e$ provides the observer with information regarding which distribution the stimulus was sampled from, seeking to train a participant to adopt a specific criterion location. This largely assumed component of the research design has far greater implications when TTKR is viewed as an independent variable. It may interact with other variables, such as working memory and stress, to further attenuate estimates of sensitivity. Despite research starting to show that this is the case, $TTKR_e$ remains the KR of choice in most SDT applications.

The Interaction of KR and Task Difficulty, and the Introduction of Optimal KR

The nature of $TTKR_e$ is such that the observer can be presented with apparently discrepant information. Due to overlap between the underlying distributions, a stimulus of a certain magnitude may at one time be sampled from the signal distribution, whereas a few trials later it may be sampled from the noise distribution. When the observer receives such contradictory information this induces criterion fluctuation. Rather counter-intuitively, research tends to support the fact that no KR can be better than TTKR (e.g., Han & Dobbins, 2008; Lee & Zentall, 1966), probably for this very reason.

While the 1970s established a tentative interest in criterion variability, Podd (1975) realised that $TTKR_e$ may actually increase criterion variability compared to no KR at all. This is because evidence presented on some trials could be indicative of an *n* or an *s* event. Podd reasoned that providing information relative to the optimum observer would provide better qualitative feedback, reducing the amount of criterion variability. Trial by trial ideal knowledge of results ($TTKR_i$) informs the observer of what the optimal response would have been, and is delivered relative to the optimal criterion position (see chapter 1). Therefore, any stimulus that falls to the right of the optimum criterion, no matter which distribution it was sampled from, should always be responded as '*S*', whereas all stimuli that fall to the left are responded as '*N*'. For example, presently the optimal criterion in the $d'_{th} = 3$ condition is located between magnitudes 10 and 11 (see Chapter 4, Figure 11, p. 38). The KR will inform the observer that any stimulus at magnitude 11 or greater is '*H*', and any stimulus falling at magnitude 10 or less is '*L*'. Consequently, observers receive information based on a single cut-off point which will maximise the percentage of correct decisions. This type of feedback represents a novel approach to providing KR to observers, but requires that an optimal decision rule can be derived.

Unlike conventional SDT methods Podd's (1975) investigation used a series of 21 tones, each of which was perfectly discriminable from the next (see Chapter 4). The investigation

examined the differential effects that $TTKR_e$ and $TTKR_i$ had on a binary discrimination task, requiring the observer to discriminate whether the tone was either a "high" tone or a "low" tone. The binary task presents a concise method to assess the effects of KR, eliminating the effects of multiple criteria, and more importantly, eliminating the possibility that criterion variability could be due to noise at the sensory level.

$TTKR_i$ provides the observer with more consistent feedback compared to $TTKR_e$; for this reason it was hypothesised that $TTKR_i$ would stabilise the criterion at a near optimal level. Podd's (1975) results confirmed that $TTKR_i$ enhanced performance in the discrimination task, pushing the observers' criterion closer to that of the optimal criterion. $TTKR_e$ reduced estimates of accuracy in classifying the tones as high or low, purely as a result of the greater criterion fluctuations it induced.

These results shed new light on the existing SDT literature. What Podd (1975) illustrated was that the established standard of using $TTKR_e$ in detection tasks may actually confound the results. However, a weakness in Podd's study was that the observers were trained to adopt a sub-optimal criterion. In fact they were trained to perform worse than chance. Effectively, when feedback was introduced a major shift from a sub-optimal location to near optimal location occurred. What this created was a large shift as the criterion relocated, thus confounding the local effect of criterion fluctuation. Consequently, the effects of KR incorporated both criterion fluctuations and the initial criterion inertia.

Richards-Ward (1992) extended upon the research of Podd (1975) by investigating the interaction between the two types of KR and task difficulty. In a departure from Podd's method, Richards-Ward (1992) used visual stimuli by way of line lengths instead of auditory tones. The fundamental task was still that of binary discrimination, yet moved the sensory component from that of audition to vision. Like Podd's research, it was hypothesised that $TTKR_e$ would produce decrements upon estimates of sensitivity, more so than that of $TTKR_i$, and that KR would interact with task difficulty to further affect estimates of sensitivity.

Task difficulty was manipulated by increasing the degree of overlap between the underlying distributions in the model. The d' value in the easy condition was $d'_{th} = 3$ (little distribution overlap), and $d'_{th} = 1$ (a lot of distribution overlap). As task difficulty increased, Richards-Ward (1992) illustrated that both optimal and real observer estimates decreased. This showed that the more difficult the task, the more criterion fluctuation can be expected, and the more $TTKR_e$ can further affect estimates of observer accuracy. However, though Richards-Ward did not train the observers to sub-optimality, they were required to set their criterion as maximally lax in the first KR trial. Though not sustained there for long, the requirement

induced a large criterion shift. It is unclear what effects that this had on the overall effect of KR.

Richards-Ward (1992) partially replicated the effect of $TTKR_e$ lowering estimates of observer accuracy, compared to those observers who received $TTKR_i$. The real observer ROC curve points did reliably fall away from the ideal observer ROC curve. However, this was more pronounced in the easy condition. Surprisingly, an *increase* in decision accuracy was observed in the difficult $TTKR_e$ condition, the complete antithesis of what should have happened. It seems likely that observers were not following the KR all the time. Had they been, this improvement in performance could not have occurred. In the present study, instead of training observers to biased positions, they were not trained to adopt any particular criterion. Instead it was assumed that observers would take a common-sense approach and set a criterion near the middle of the range of tones.

The two unpublished reports of Podd (1975) and Richards-Ward (1992) contain valuable information that has not been introduced into mainstream literature. Their results are empirically valuable and provide a means through which previous accounts of reduced estimates of sensitivity can be explained. Given the weight of information, the present investigation attempts to both refine the methodologies previously used by Podd and Richards-Ward, and replicate the findings that types of feedback interact with task difficulty to lower estimates of observer discriminability.

The present research, therefore, is not a traditional psychophysical study where an observer discriminates between signal and noise stimuli. Rather it is concerned solely with the decision-making abilities of the observer. Consequently, the normal sensory component of the SDT task was eliminated, so as to assess only the decisional stage and the effects of the fluctuating criterion. The underlying densities in the theoretical model retain Gaussian form and equal variances, and remain fixed at one of two overlapping positions, to provide two levels of difficulty. These known parameters also allow for the optimal criterion to be established, thus allowing for the delivery of ideal KR.

The modifications made to the present design provide increased empirical robustness, and thus allow for more substantiated effects. The specific details of how this will be achieved are the topic of the next chapter.

Chapter IV

The Present Research

The present Criterion Variance Model (CVM) specifically identifies and assesses contributors to criterion variance. This model focuses on the criterion variability caused by a noisy decision process, and suggests a method for reducing this noise. Failure to eliminate the effects of decision noise reduces the apparent accuracy of the decision maker, as outlined in Chapter 2[8].

Two factors in particular contribute to criterion variability: task difficulty and type of KR. How a change in task difficulty affects criterion variability has received virtually no attention in the literature. One unpublished study (Richards-Ward, 1992) out of the Massey University lab did vary task difficulty while examining the effects of KR on performance in a task very similar to the present one. This study was briefly discussed in the previous chapter. To recap, Richards-Ward (1992) demonstrated $TTKR_i$ improved performance compared to $TTKR_e$; however, this effect was more pronounced in the easier condition. Furthermore, observer performance improved when $TTKR_e$ was introduced into the hard condition, despite performance being poorer than the observers who received $TTKR_i$. This finding suggested that observers were perhaps not consistently using the KR. While Richards-Ward (1992) provided some evidence that $TTKR_i$ improved performance under difficult conditions, the results are equivocal. The major impetus behind the present research was to replicate and extend the findings of Richard-Ward by clearly demonstrating that a) criterion variance increases with task difficulty; and b) that different types of KR is required when the decision task is difficult if decision noise is to be minimised.

Task difficulty can be investigated by varying the degree of overlap between the S and N distributions (see Figure 11, p. 38). As the overlap increases it becomes increasingly difficult to distinguish *s* and *n* events. The probability of a value of the evidence variable (*x*; see Chapter 1, p. 6) indicating an *s* or *n* event approaches equality with increasing difficulty. So, for example, in a relatively easy task where the overlapping distributions are the N and S3 distributions shown in Figure 11, stimulus frequency 8 has a probability of 0.19 for the N distribution, and a probability of 0.01 for the S3 distribution. Clearly, for this value of the evidence variable one would maximise correct decisions by responding *"N"*. However, for stimulus frequency 8 in the harder condition (0.15), while the response *"N"* remains the best, there is now a much increased probability that stimulus frequency 8 represents a value

[8] Although most of the literature on criterion variance does not have a direct bearing on the present study, it was necessary to describe the methods that have already been taken to investigate criterion variance. The present study takes a markedly different approach.

sampled from the S1 distribution. The implication of an increase in the S1 probability for stimulus frequency 8 is that an observer may response *"S"* on one trial and be correct because it had been sample from the S1 distribution. However, the observer may be incorrect on a subsequent trial when provided with exactly the same evidence, because the evidence had been sample from the N distribution.

The above argument holds true for many values of the evidence variable in the difficult condition compared to the easy condition. Therefore, the error rate increases, and as the observer attempts to maximise the number of correct responses, adjustments to the criterion are likely, inducing criterion fluctuation. Both Larkin (1971) and Mueller and Weidemann (2008) demonstrated that stimuli probabilities affect the location of the criterion. Clearly, if a stimulus has almost an equal chance of being sampled from the N or S distribution, this will affect the decision rule that the observer operates by, and increases the difficulty in discriminating whether the evidence was a *s* or *n* event.

Intuitively, the introduction of KR is a step that should serve to mitigate the effect of criterion variability. $TTKR_e$ is frequently used in many SDT designs, often with the implicit assumption that the KR will help keep the observer "on track". The role of KR as an independent variable in its own right has received little attention; accordingly, the CVM addresses the possibility that specific types of KR may increase, rather than decrease, criterion variance. The nature of $TTKR_e$ is that contradictory KR can be provided by relaying back to the observer which distribution an event had originated from. This type of KR provides evidence for the veridical nature of the event. The inherent problem with such KR is that it can be contradictory for the stimuli that have increased probabilities of being sampled from both N and S distributions. This confuses the observer who then tries to eliminate error by shifting the criterion so as to minimise the chance of making the same mistake again (Larkin, 1971). In a difficult discrimination task $TTKR_e$ is especially ambiguous for the reasons described earlier. In fact, it is possible that this type of KR results in a poorer performance than receiving no KR at all.

The best way to minimise these errors is to provide a decision rule that precludes uncertainty. This can be achieved by relaying KR back to the observer relative to the optimal criterion (see Chapter 1, p. 14). This type of KR is called $TTKR_i$, and the decision rule is relative to a fixed location on the decision axis. Though errors cannot be avoided, $TTKR_i$ maximises the percentage of correct decisions, which maximises observer performance. This type of KR will stabilise the criterion and reduce fluctuation. This means that every time a particular event occurs, for example frequency 8, that same event will always be regarded as noise irrespective of which distribution it was drawn from, because it falls to the left of the

criterion. On this basis $TTKR_i$ should improve observer accuracy. Based on the arguments for task difficulty and type of KR, the CVM makes two predictions:

1. Contrary to the assumption of SDT, the decision criterion in a signal detection task is a variable rather than a fixed value on the decision axis, and is present within binary discrimination tasks (Hypothesis 1); and,
2. an interaction between the type of TTKR provided and the difficulty level of the task. Specifically, $TTKR_i$ will enable more accurate decision making than $TTKR_e$, but only for a difficult decision task (Hypothesis 2).

In summary, the CVM predicts that that $TTKR_e$ and $TTKR_i$ will have approximately equal affect on performance in an easy binary decision task. For a hard discrimination task $TTKR_i$ will enable better performance than $TTKR_e$ because the former induces less criterion fluctuation than the latter. Figure 10 summarises these predictions.

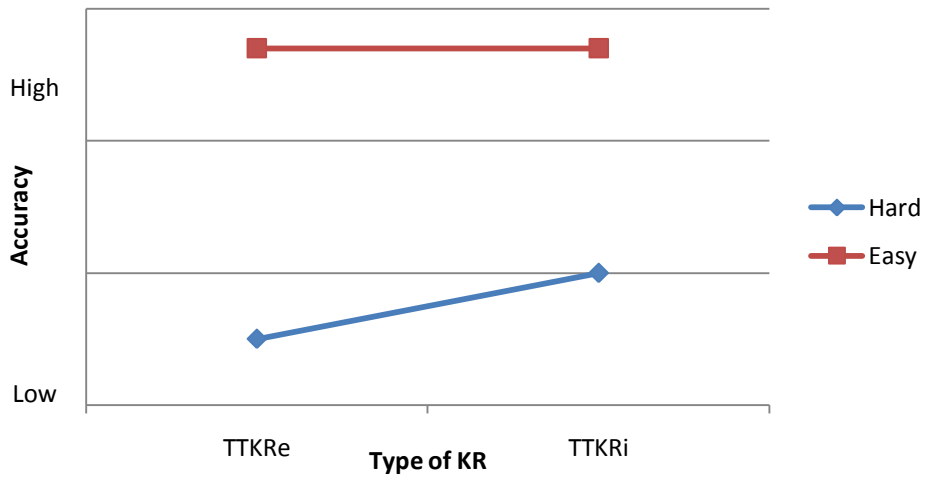


Figure 10: *The interactive predictions for the current research. In a hard decision task $TTKR_i$ is expected to improve observer accuracy compared to that of $TTKR_e$. However, in an easy decision task the type of KR is expected to have little, or no, effect.*

In order to test the model, specific requirements had to be met. The present research departs from classical SDT in two significant ways. Primarily, the concern of the research was to investigate the decision process in a binary discrimination task, using SDT methods and statistics. Second, all tonal stimuli were easily discriminable from each other. The auditory stimuli used in Podd's (1975) original study were retained, as the tones had already been subject to discriminability analysis.

The underlying distributions in the model are approximations to the standard normal curve, calculating probability mass functions for each stimulus magnitude, thus, $P(X=x) \sim N(0,1)$. This produced good approximations to Gaussian distributions with equal variances. Each sampling distribution (either high or low) consisted of 200 tones spread over 14 discrete

tonal frequencies with unit standard deviation. The trial sequence consisted of 400 tonal presentations (200 "high" and 200 "low"), which were randomly ordered by sampling without replacement.

Discriminability between high and low tones was set at two levels of difficulty (d'_{th} = 1, d'_{th} = 3; see Figure 11, p. 38). In both conditions the "low" distribution was fixed at $\mu = 0$, $\sigma = 1$, with the mean lying between tonal frequencies 7 and 8. The degree to which the "high" distribution overlapped with the low distribution was manipulated for each condition, and quantified by the measure d'. In the d'_{th} = 1 condition the high tone mean sat one standard deviation above the low tone mean ($\mu = 1$, $\sigma = 1$), with its mean located between tonal frequencies 8 and 9. The d'_{th} = 3 condition saw the high tone mean move three standard deviations above the low tone mean ($\mu = 3$, $\sigma = 1$), with its mean located between frequencies 13 and 14. Figure 11 illustrates the overlapping distribution for each condition. The ROC curves for each condition were generated by passing the criterion from right to left along the decision axis for the distributions in Figure 11, and can be seen in Figure 12 (p. 39). The ROC curves are therefore the theoretical curves for the ideal observer, and represent the maximum level of performance obtainable in the present tasks, given the degree of uncertainty introduced by the overlapping high and low tone distributions.

The area under the ROC curve (AUC) also provides a measure of discriminability (MacMillan & Creelman, 2005). The AUC can take on values between 0.5 (chance performance and complete distributional overlap) and 1.0 (complete distributional separation). The measures used to calculate the AUC for the study's theoretical ROC curves was the non-parametric A_g (see Equation 6, p. 11; MacMillan & Creelman, 2005). For the two levels of difficulty, A_g for the ideal observer was 0.78 for d'_{th} = 1 condition and 0.98 for the d'_{th} = 3 condition[9].

Response bias was also assessed by using the measure c (see Equation 11, p. 15). Bias measures ideally should meet two standards (McMillan & Creelman, 2005, p. 39): (a) the measure should depend monotonically upon the HR and FAR in the same direction, and (b) should be independent of the sensitivity index. For these reasons c is a better candidate as a measure of bias than β (MacMillan & Creelman, 2005). Where the negative diagonal in the ROC space intersects the ROC, c = 0; correspondingly then, the optimum criterion in the d'_{th} = 3 condition is located between frequencies 10 and 11. The d'_{th} = 1 optimum criterion was located between frequencies 8 and 9. All $TTKR_i$ was given relative to these criterion positions.

[9] The distributions were modelled on the standard normal curve so a parametric measure could have justifiably been used.

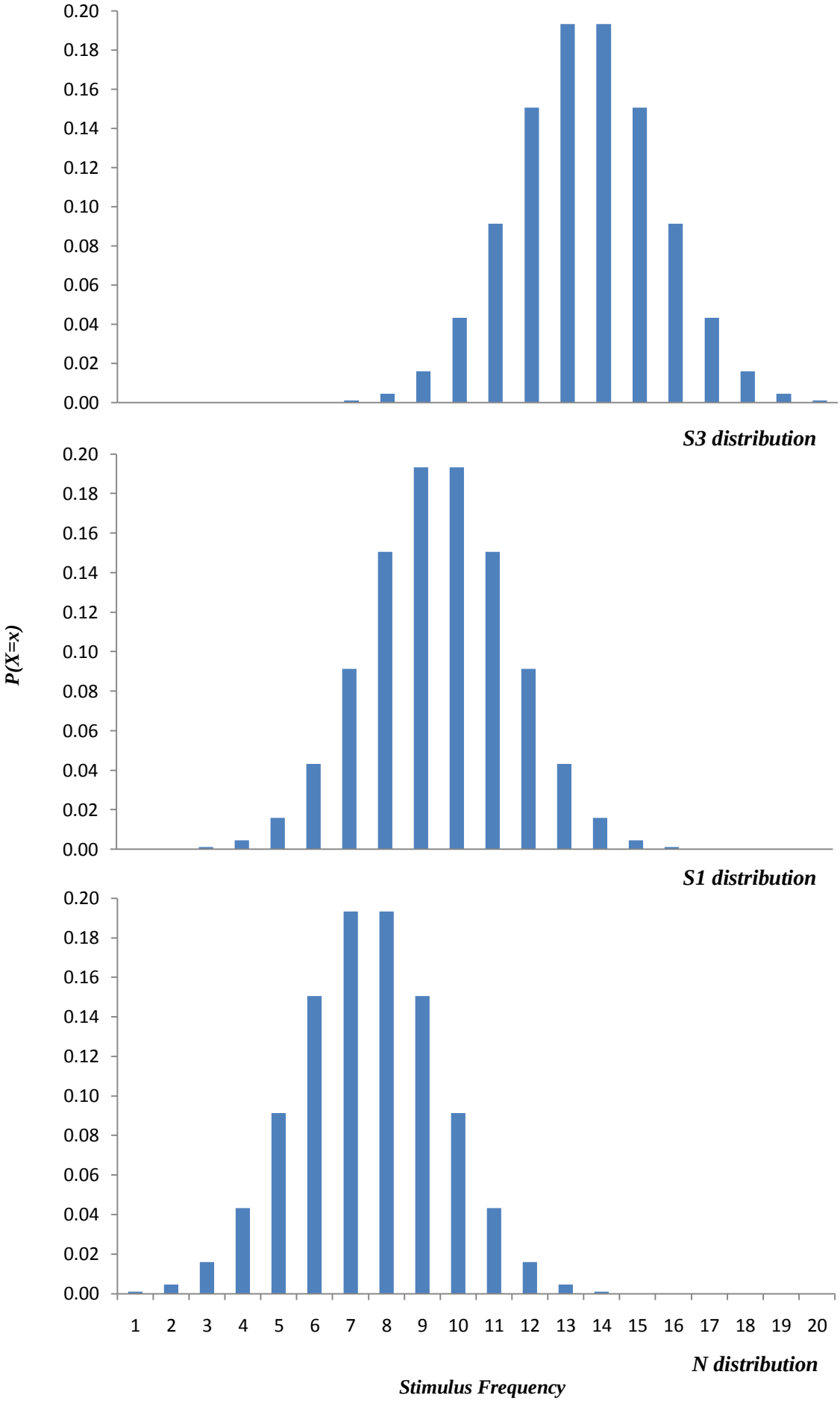


Figure 11: *Probability distributions for the present research, showing degrees of overlap for both levels of difficulty. N = noise distribution; S1 = signal distribution d'= 1; S3 = signal distribution d'= 3.*

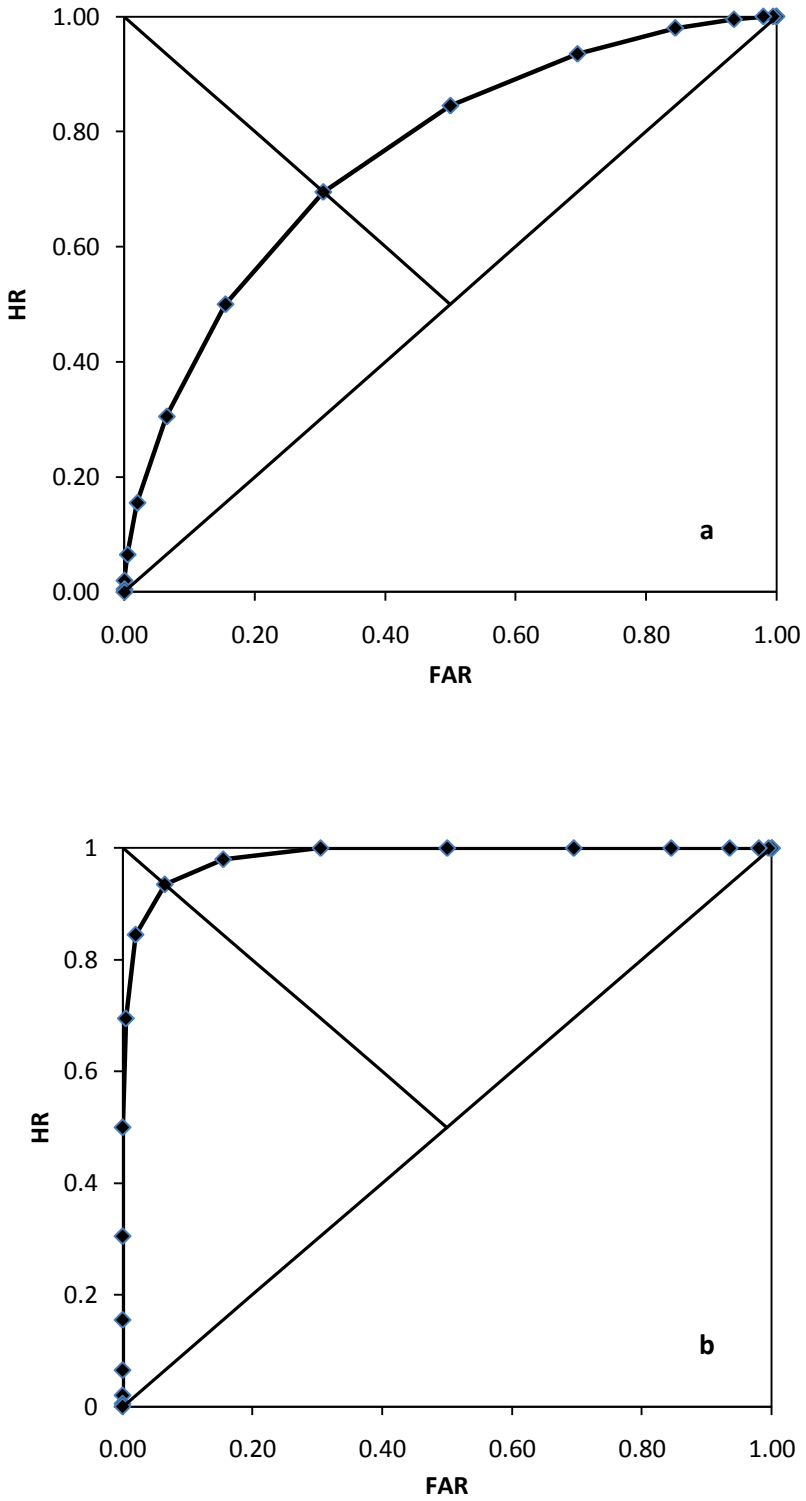


Figure 12: *Theoretical ROC functions for both levels of difficulty;* ***a)*** $d'_{th} = 1$; ***b)*** $d'_{th} = 3$.

Method

Pilot Investigation

The original design for the present investigation used a d'_{th} = 0.5 as the hard condition. It was hoped that by making the task very difficult the criterion would fluctuate maximally under such conditions.

Originally, 100 training trials were completed in order to familiarise the observer with the trial sequence. Furthermore, the trials were used to train the observer to adopt an optimal criterion position. An issue with Podd's (1975) and Richards-Wards' (1992) research was that the criterion variability observed may have been confounded with an overall shift in the criterion. To alleviate this problem the present study initially trained the observer, using $TTKR_i$, to acquire a criterion position near optimal. The hypothesised effect was that when $TTKR_e$ was delivered, the criterion would become less stable and decay from the optimal position, particularly so in the d'_{th} = 0.5 condition. However, during the pilot runs this was not the case.

Observer accuracy improved with the introduction of $TTKR_e$ in the d'_{th} = 0.5 condition. It appeared that the $TTKR_i$ in the difficult condition was difficult to follow because so many of the tonal frequencies could have been sampled from either distribution. Therefore, observers abandoned using the feedback altogether. In order to combat the contradictions the observers appeared to revert back to the decision strategy used during the training.

Accordingly, it was decided to push out the high distribution in the difficult task a further 0.5 of a standard deviation, producing a d'_{th} = 1 difficult condition. The KR in this condition was a little less likely to appear inconsistent to the observer. However, observers still failed to consistently adjust their responses according to the KR provided.

In an attempt to overcome this problem, the training was reduced to 50 trials, with no feedback given at all. Instead, the observers were simply informed that higher tones generally indicated the high distribution has been sampled from, and vice versa for the lower tones. No KR was provided. In addition, in order to encourage consistent use of the KR, a payoff matrix was introduced. Observers were awarded 2c for every correct decision made, but penalised 2c for every incorrect decision. Observers then had the potential to earn $8 over and above their reimbursement for participation. It was hoped that such a measure would ensure observers would consistently utilise the KR provided. Additionally, the final set of instructions stressed the importance of using the trial-by-trial feedback to obtain the best monetary payoff.

In summary, for the actual experimental trials training was reduced to 50 trials with no feedback, the hard condition relaxed to $d'_{th} = 1$, and a payoff matrix introduced to encourage consistent use of the feedback.

Main Study

Observers

Forty-four observers with normal hearing agreed to participate in this study with informed consent. The sample consisted of 26 women (59%) and 18 men (41%) with an age range from 19 to 50 years (*M = 23.2; SD = 5.5* years). Observers were randomly allocated to one of four experimental conditions. The majority of observers were Massey University students, mainly recruited by flyers advertising the study. Informed consent was obtained from each participant and the investigation was approved as a low risk study by the Massey University Human Ethics Committee.

Apparatus and Stimuli

The experimental stimuli consisted of 20 discriminable tones. The 20 tones retained for this study ranged from 382 Hz to 1187 Hz, separated by 19 equal steps of 42.4 Hz. The frequency values were greater than 12 JNDs but less than 16 JNDs apart (50% level), ensuring that all tones were easily discriminable from the next (Shower & Biddulph, 1931). A table of the tonal frequencies can be found in Appendix B. Tones were set and played at a level that the observers found comfortable[10].

Tones were computer generated wave files that were programmed into the stimulus presentation programme. The tones were generated using a Hann function – also known as a 'Hanning Window'. The Hann function is a specific type of tapering window function used in signal processing. The function tapers the ends of a sampled region – in this case the tone – to bring it smoothly up from and down to zero. The effect is the reduction of unwanted noise at the extremes of the sample (Blackman & Tukey, 1959).

For each trial a tone was either sampled from the '*L*' or '*H*' distribution, and was done so on a random basis. No more than three low or high events could occur consecutively in the sequence. This reduced the possibility of sequential effects occurring. Trial presentations were controlled by a programme into which the trial randomisations were written. The programme produced outputs that recorded hits, false alarms, misses, and correct rejections,

[10] Unfortunately sound pressure levels could not be obtained for the headphones. Typically an artificial ear is used which measures pressure at the earpiece of the headphones. However, the unavailability of the equipment meant that sound pressure information could not be reported.

as well as HR and FAR. Additionally, the observers response ('H' or 'L') and the KR provided ('H' or 'L') was available for each trial.

Verbal instructions were recorded and played back open air so both observers could listen simultaneously (two observers were run at a time using separate computers). On screen instructions were written in 12 point Arial font. Observers were stationed between 35 – 55cm from the screen. All tones were played through Panasonic RP-HT161E-K stereo headphones.

Design

The study used a 2x2 factorial between-subjects design. There were four experimental conditions into which observers were randomised. Two independent variables were manipulated: task difficulty (d'_{th} = 1 or 3), and type of KR ($TTKR_e$ or $TTKR_i$). Recall that $TTKR_e$ provided feedback in relation to the distribution that a stimulus was sampled from (either high or low), whereas $TTKR_i$ provided feedback in terms of the ideal observer's decision rule. $TTKR_i$ informed the participant of what the best response for each stimulus would have been. Two dependent measures were obtained: d'_{ob} for the observer's accuracy across all conditions, and A', which estimates the AUC using a single HR and FAR (see Equation 7a and 7b, p. 11). A measure of response bias was also obtained using c.

Each trial lasted 3500 ms and consisted of tone presentation (200 ms; including rise-fall time), response interval (2000 ms), feedback (300 ms), and inter-trial interval (1000 ms). Figure 13 shows the sequence of events for each trial. The response interval duration allowed 2000ms for a response to be made; if no response was made a "no response" was indicated on the output file. As soon as the observer had made a response, the feedback was immediately given. Therefore, trials differed in length, with 3500ms being the absolute maximum trial length. This trial format was the same for both main and training trials.

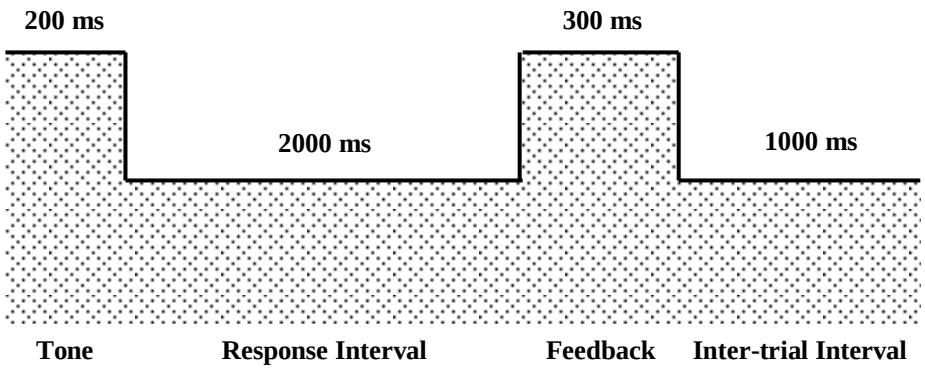


Figure 13: *Trial sequence and corresponding times.*

The response box was centred on the computer screen, and consisted of a "high" response button (on the left) and a "low" response button (on the right). A green light flashed above the response button that should have been selected on that trial. All trials were completed in a single session, with two observers being run through the experiment at a time (using separate computers). All observers completed 50 training trials before moving on to the main block of 400 trials. Observers then received either events or ideal KR relative to the level of difficulty they were randomised to (d'_{th} = 1, or 3). Observers were given a short break after completing the first 200 trials.

Procedure

Observers first read an information sheet (Appendix C) and were encouraged to re-read the sheet if necessary. They were then told that the experiment would take no longer than 45 minutes. They were also informed that they would be reimbursed $10 for their time, and that they could withdraw at any stage during the experimental phase; however, payment would not be received unless the study was completed.

All observers had been randomised to a particular condition prior to commencement of the trials. Upon arrival the observers signed a consent form (Appendix D) and then were assigned to the computer with the appropriate programme loaded on. Instructions were present on the screen but the observers were asked to turn their attention to the experimenter so that the verbal instructions could be played (see Appendix E, for verbal instructions). The verbal instructions introduced the pay off matrix, highlighting the potential to earn an additional $8 by earning 2c for every correct decision, but also emphasising that an incorrect decision resulted in a 2c penalty. It was again reiterated that using the TTKR would yield the best outcome. Once the verbal instructions had finished the observers were asked to put their headphones on and read the training instructions on the screen (see Appendix E, for training instructions).

After reading the instructions the observer was instructed to click on the "continue" box on the screen, upon which the full range of tones were played. Immediately after the tones had played the response box was centred on the screen, with the two feedback boxes filled yellow. This colour indicated that the observer had two seconds before the beginning of the first trial. The observers then completed the 50 training trials.

Next, main trial instructions were brought up on the screen (see Appendix E, main trial instructions). The main trial instructions introduced the role of the green feedback light and further reiterated the KR should be used to enhance decision making. Clicking the "continue" box centred the response box on the screen, with the yellow feedback boxes again providing

a 2 second warning prior to trial commencement. Observers were then presented with the first 200 trials, after which a message box popped up indicating that the observer was half way through, and a break was recommended. Breaks were no longer than 5 minutes, after which the observers completed the final 200 trials.

The experimenter then calculated the observer's earnings by taking the number of correct decisions (Hits + Correct Rejections) and subtracting the number of incorrect decisions (False Alarms + Misses). This amount was then added to the $10 already paid for participation. Observers were thanked for their time and an offer extended to ask any questions about the study, as well as the opportunity to read the final report.

Chapter V

Results

There were 11 observers in each of the four experimental conditions. Preliminary analyses were undertaken to assess whether gender and time of study had any effect on the dependent variables. The gender distribution across experimental conditions was relatively even, except for the d'_{th} = 1 events condition. Analysis revealed no significant relationship between gender and condition ($p = .90$, *Fisher's exact test*[11]). Furthermore, the time of day (am or pm) that the experimental trials were run had no effect on performance, $\chi^2 (1) = 0.53$, $p = 0.47$.

Observer Performance

Table 1 summarises the mean values for the primary dependent variables, d'_{ob} and A'. Mean hit and false alarm rates are also provided along with mean *c* values.

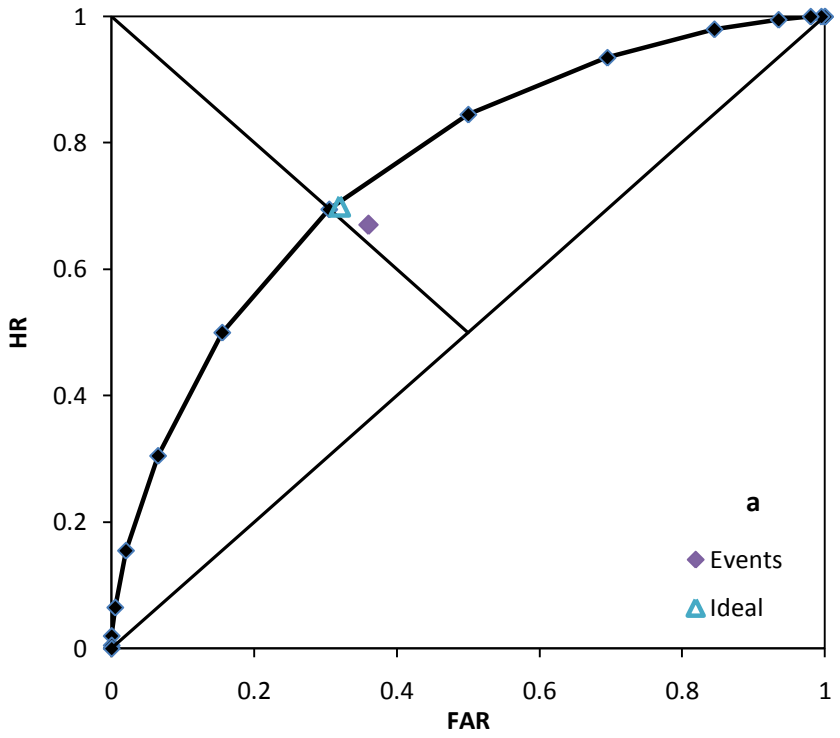
Table 1: *Mean values for dependent measures across independent variables.*

	d' = 1				*d' = 3*			
	Events		*Ideal*		*Events*		*Ideal*	
HR	0.67	(0.02)	0.7	(0.03)	0.88	(0.04)	0.88	(0.04)
FAR	0.36	(0.03)	0.32	(0.03)	0.14	(0.04)	0.14	(0.04)
d'	0.80	(0.07)	0.99	(0.07)	2.25	(0.3)	2.25	(0.24)
A'	0.74	(0.01)	0.78	(0.01)	0.93	(0.02)	0.93	(0.02)
c	-0.03	(0.05)	-0.02	(0.07)	-0.05	(0.13)	-0.05	(0.13)

Note: Values in parentheses = SD.

Both d'_{ob} and A' varied across KR groups in the hard (d'_{th} = 1) condition, whereas they remained stable in the easy (d'_{th} = 3) condition, as predicted by the CVM. The mean performance for all groups is shown in Figures 14a and 14b (on next page) relative to the theoretical ROC curves. $TTKR_e$ in the hard condition produced an attenuated HR (*0.67; SD = 0.02*), and an increase in the FAR (*0.36; SD = 0.03*). $TTKR_i$ in the hard condition produced a near optimal HR (*0.7; SD = 0.03*) and FAR (*0.32; SD = 0.03*). The mean d'_{ob} value in the $TTKR_e$ group ($d'_{ob} = 0.8$, *SD = 0.07*) was noticeably lower than the $TTKR_i$ group ($d'_{ob} = 0.99$, *SD = 0.07*). When observer performance is compared to the ideal observer using η (a measure of observer efficiency; Eq. 8, p. 12), the $TTKR_i$ group ($\eta = 0.98$) performed much closer to optimal than the $TTKR_e$ group ($\eta = 0.64$).

[11] SPSS output for 2x4 contingency table exact test provides Fisher's Test, though it is assumed this is the Freeman-Halton extension of the 2x2 Fisher's exact test.



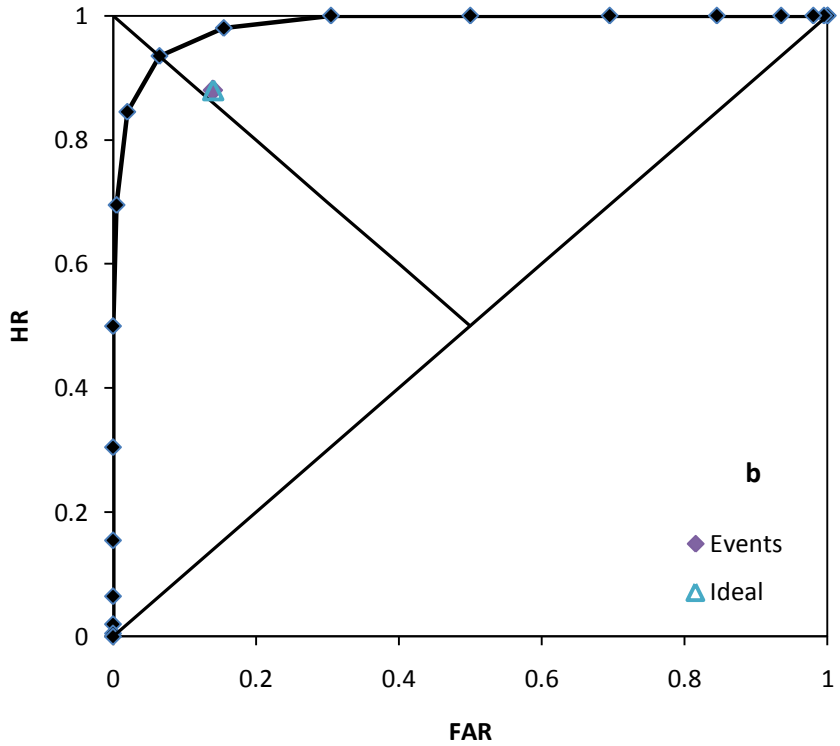


Figure 14: a) *Theoretical ROC for d'_{th} = 1 condition depicting events and ideal mean d'_{ob} values. Although the groups had a mean c close to optimal, there is a noticeable difference between the events and ideal KR group;* **b)** *Theoretical ROC for d'_{th} = 3 condition depicting events and ideal mean d'_{ob} values. In the easy version of the task, the criterion adopted was near optimal. As predicted by the CVM there is little difference in average performance for events and ideal KR.*

This difference is also evident for the *A′* values, where the $TTKR_e$ group (*A′ = 0.74, SD = 0.01*) produced a lower score than the $TTKR_i$ group (*A′ = 0.78, SD = 0.01*). $TTKR_e$ and $TTKR_i$ ROC points further illustrate the differences between groups (see Figures 14a and 14b). The $TTKR_i$ point indicates near optimal performance, whereas the $TTKR_e$ point falls below the theoretical curve. There existed some variation in *c* across conditions, though this was minimal with all values being close to *c = 0*, the optimal position. Both easy KR conditions provided identical HR (*0.88; SD = 0.04*) and FAR (*0.14; SD = 0.01*), yielding identical d'_{ob} and *A′* values across $TTKR_e$ (d'_{ob} *= 2.25, SD = 0.3; A′ = 0.93, SD = 0.02*) and $TTKR_i$ (d'_{ob} *= 2.25, SD = 0.24; A′ = 0.93, SD = 0.02*) groups (η *= 0.56* for both KR groups). Figures 15a and 15b below graph observer performance across KR groups for each condition. The easy condition produced a stable performance across KR groups for both *A′* and d'_{ob}. Observer accuracy improves in the $TTKR_i$ condition within the hard condition for both measures. Thus, there was an interaction between task difficulty and type of KR, as predicted by the CVM. Each hypothesis is investigated in the following sections.

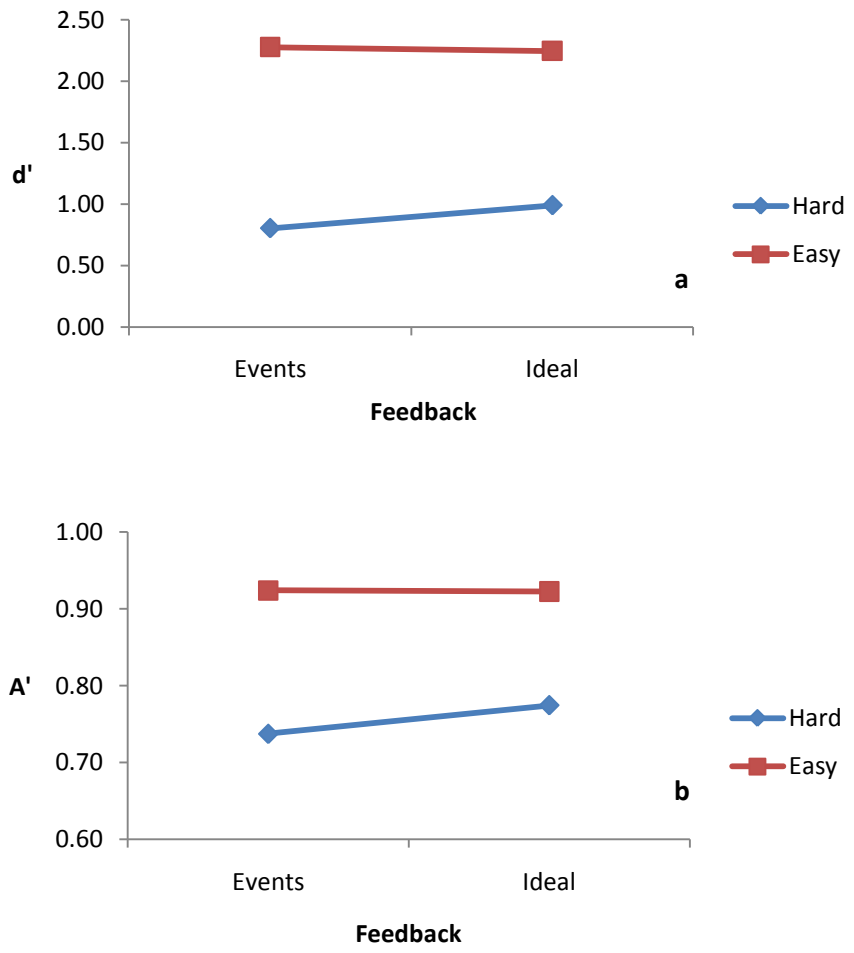


Figure 15: a) *Accuracy measure, d'_{ob}, across KR groups for both d'_{th} = 1(easy) and d'_{th} = 3 (hard) conditions. Performance was better in the hard ideal KR condition, compared to that of the events KR condition;* **b)** *Accuracy measures, A', across KR groups for both d'_{th} = 1 (easy) and d'_{th} = 3 (hard) conditions. Performance again was better in the hard ideal KR condition, compared to that of the events KR condition.*

Hypothesis 1(Criterion Fluctuation)

Hypothesis 1 predicted that the criterion will shift, even in a binary discrimination task. In SDT, the value of the criterion (*c* in the present study), once adopted, is assumed to remain fixed. In the present case, the optimal value for the criterion was *c = 0*. Indeed, as Table 1 (p. 45) shows, the mean *c* value for all conditions fell very close to zero, the optimal value. However, had decisions been optimally made the ROC points shown in Figures 14a and 14b would have coincided with the theoretical ROC curve.

The suboptimal values can be due to only one thing in the present study – criterion fluctuation. To investigate criterion fluctuation, the number of errors each observer made for each stimulus value were plotted. The spread and density of the error plots are indicative of criterion fluctuation, with increased spread and density indicating greater criterion fluctuation. For comparison, cases with similar criterion locations (*c*) were selected, but performance was either poor or good, relative to the condition the cases were drawn from. In cases where performance was poor, increased spread in the error distribution was expected, despite near optimal *c* values.

These error response distributions also provide support for Hypothesis 2 as the degree of criterion fluctuation is influenced by the type of KR received and task difficulty. According to the CVM theory, $TTKR_e$, particularly in the hard condition, should produce increased criterion fluctuation. In assessing the validity of Hypothesis 1 an appeal to Hypothesis 2 is unavoidable; however, an assessment of Hypothesis 2 is made in the following section. Complete response distributions for select observers in each condition are displayed in Appendix H, which also contains the number of correct and incorrect decisions for each stimulus.

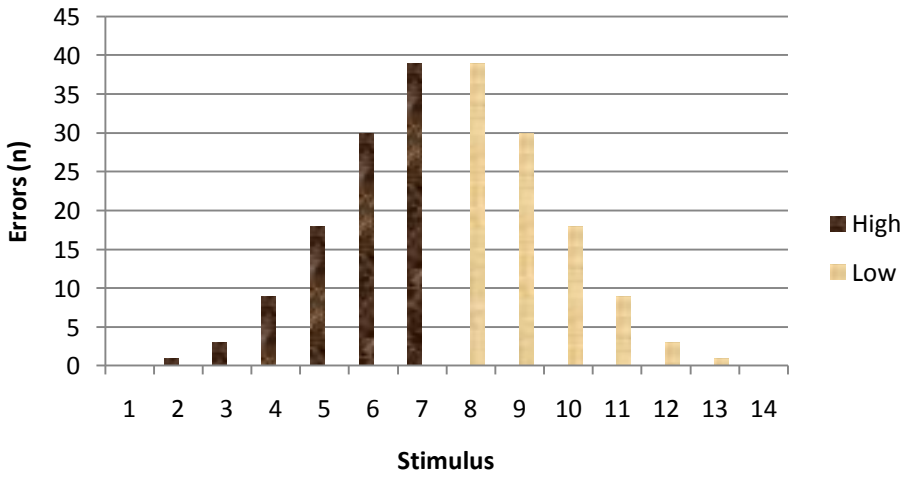


Figure 16: *Example of an ideal error distribution, with the optimum criterion located between tones 7 and 8. If the criterion is fixed then high errors should only fall to the left of the criterion, whereas low errors should only fall to the right. Errors should also reduce in frequency for tones further away from the optimum criterion.*

The error distributions present the errors an observer made in deciding whether a particular event was drawn from the high or low distribution. Two distributions of errors are generated; one for high errors and one for low errors. Low errors indicate that a false alarm occurred – deciding that a low tone was actually a high tone. High errors indicate that a miss occurred – deciding that a high tone was actually a low tone. Errors will naturally occur in the present decision task, even for the optimal observer. However, assuming the criterion has been fixed at the optimal location, an error response distribution (see Figure 16 on the next page for an example of an error distribution) will have two defining features.

First, the number of errors committed should remain confined to the area of overlap between the high and low distributions. This is because the overlapping area contains all tones that could be indicative of either a high or a low state. If an observer consistently used a fixed criterion then errors will not occur outside of this region. Furthermore, errors should occur more frequently for stimuli in the middle of the area of overlap, as these stimuli occur more frequently. Consequently, the distribution of errors would taper off significantly toward the ends of the area of overlap. Thus, fewer errors will occur toward the ends of the area of overlap because these stimuli occur less frequently. Increases in error frequency toward the tail regions, and the presence of errors outside the area of overlap, indicate a criterion that is fluctuating.

Second, the high and low errors create two distributions (one reflecting false alarms, the other misses) that span the area of overlap (tones 3 – 15 in the hard condition; tones 8 – 13 for easy condition). If the criterion is fixed then the high error distribution will be to the left of the criterion, and the low error distribution will be to the right of the criterion. Thus, the frequency of high errors will be greatest for tones just to the left criterion and will gradually decrease for tones further away from the criterion. For low errors the frequency will be greatest for tones just right of the criterion. By implication the high and low error distributions should not overlap if a fixed criterion is adopted. If there is overlap between the high and low error distributions, or if high and low errors are recorded on the opposing side of the criterion, then the criterion is fluctuating. This occurs because the criterion is shifted to another location, thus decision are based on a different decision rule. This alters the frequency of errors and changes the shape of the error distribution because the tones are judged by a different standard. For example, moving the criterion leftward (becoming lax) reduces the number of tones to the left of the criterion. Consequently, more tone to the right are considered high, therefore, more high errors are likely to occur. This would account for high errors occurring in lower tone frequency ranges. Illustrative cases of criterion fluctuation are presented next.

Condition: d'_{th} = 1, events. Observer 2 (d'_{ob} = *0.93; c = 0.00*) provided the highest d'_{ob} score in this condition. Observer 38 (d'_{ob} = *0.76; c = -0.06*) was one of the poorer performers in this condition.

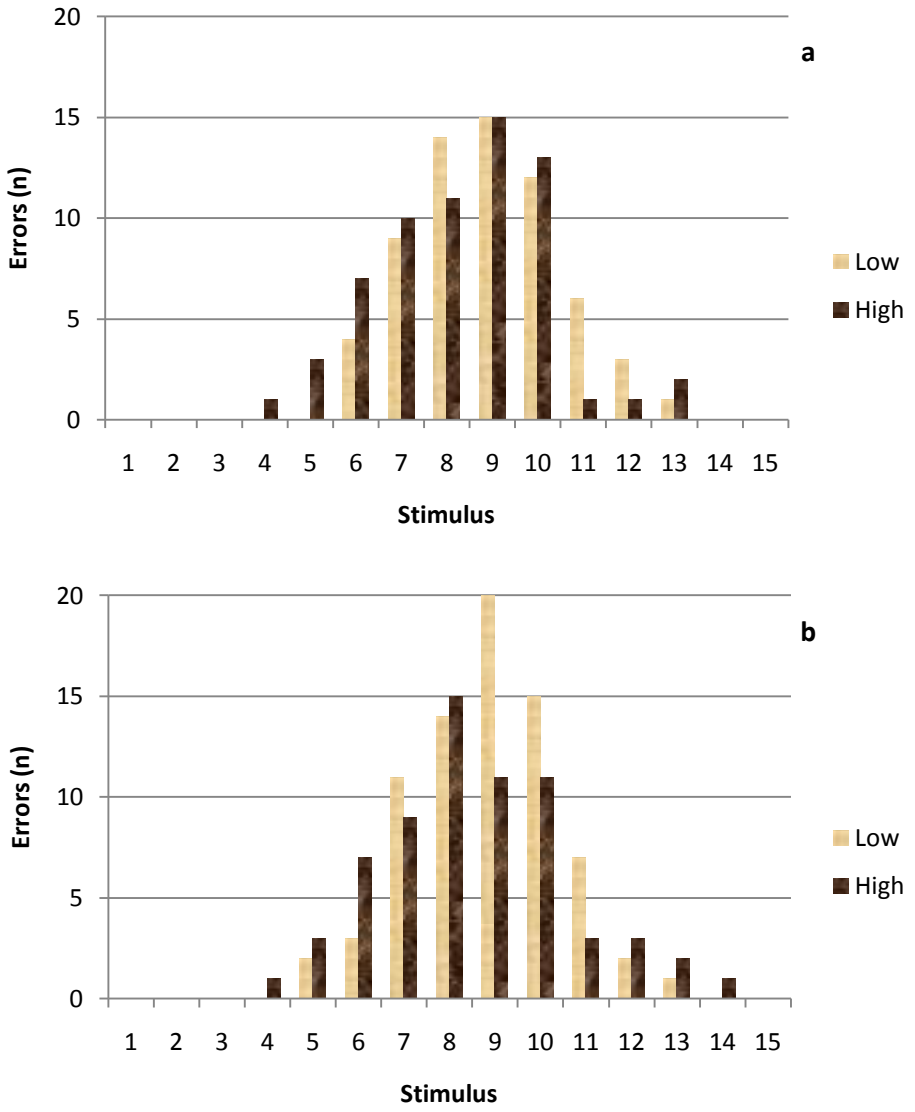


Figure 17: a) *Distribution of errors for Observer 2;* **b)** *Distribution of errors for Observer 38. Both distributions reflect errors made using $TTKR_e$ under hard conditions. The optimum criterion is located between tones 8 and 9.*

Figures 17a and 17b display the errors made by each observer. In both cases the high and low error distributions overlap significantly. Furthermore, high error frequencies were recorded as far up as tonal frequency 14 (Observer 38), though this was not outside the area of overlap. On average, the errors remain confined to the area of overlap in the stimulus distributions; however, the frequency of errors was greatly increased toward the end of the overlap area for both high and low errors, indicating that the criterion was shifting significantly. The frequency of high errors within the higher tonal region was more prevalent for Observer 38. Low errors also occurred well into the low range with some frequency for

both observers. These error distributions can only arise from a failure to maintain a consistent decision rule. Thus, they show how widely a criterion can fluctuate, even though in SDT the criterion estimate *c* is given by a single value, implying no such fluctuation.

Condition: d'_{th} = 1, ideal. Observer 27 ($d'_{ob} \approx 1$; $c = -0.02$) was one of the better performers in the $TTKR_i$ hard condition. Observer 24 ($d'_{ob} = 0.95$; $c = 0.04$) was one of the poorer performers in this condition.

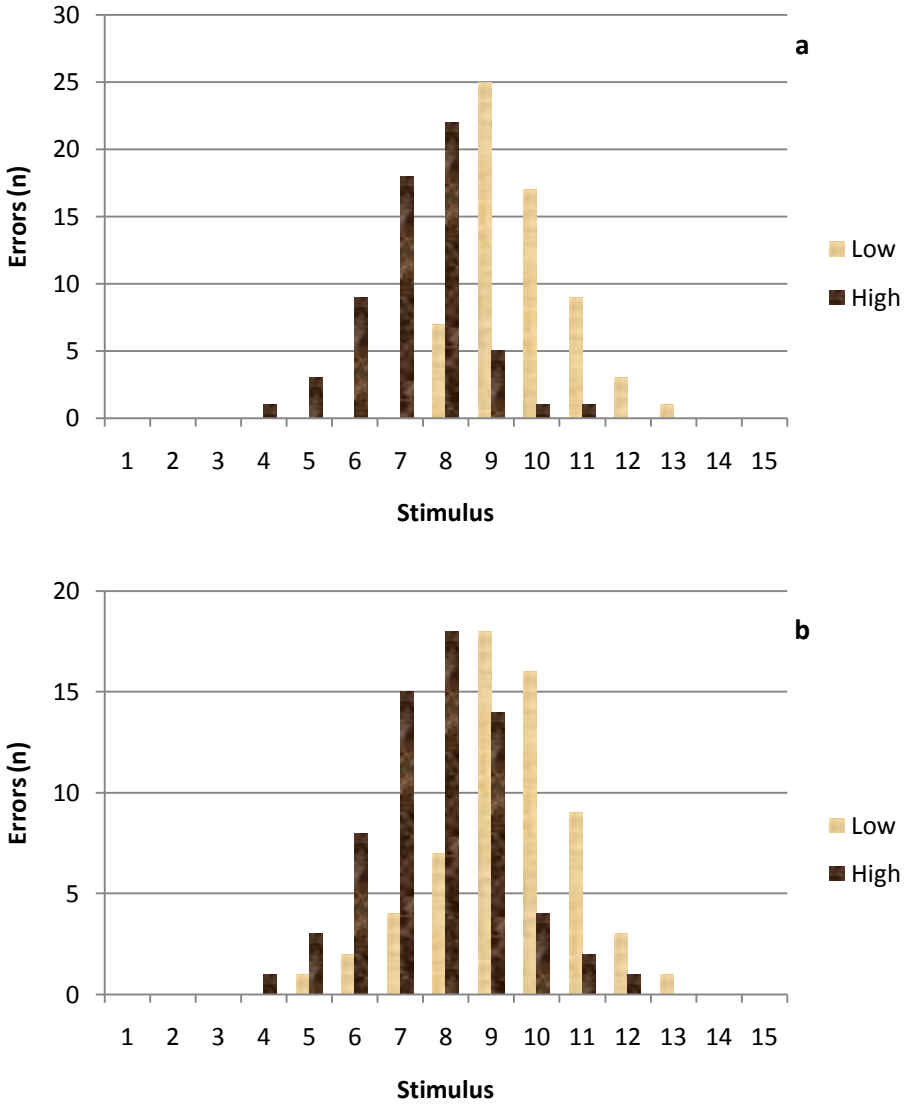


Figure 18: a) *Distribution of errors for Observer 27;* **b)** *Distribution of errors for Observer 24. Both distributions reflect errors made using $TTKR_i$ under hard conditions. The optimum criterion is located between tones 8 and 9.*

Figures 18a and 18b display the error distributions for each observer. For both observers the spread of errors remain confined to the area of stimulus distribution overlap, with errors less frequent toward the ends of the overlap. Observer 27 had an error distribution that best approximated what an ideal error distribution should look like. There was very little overlap in the high and low errors, sticking predominantly to the relative side of the criterion. There

was some overlap occurring around the criterion point, but the fluctuation did not move that far. Observer 24 showed increased errors around the midpoint; however, they were more frequent for tones further away from the criterion than Observer 27. The degree of error overlap was also greater, though there was some separation.

Condition: d'_{th} = 3, events. Observer 29 (d'_{ob} = *2.78; c = -0.02*) was the best performance in this condition, and indeed across all d'_{th} = 3 conditions. Observer 44 (d'_{ob} = *1.83; c = 0.00*) was the least accurate.

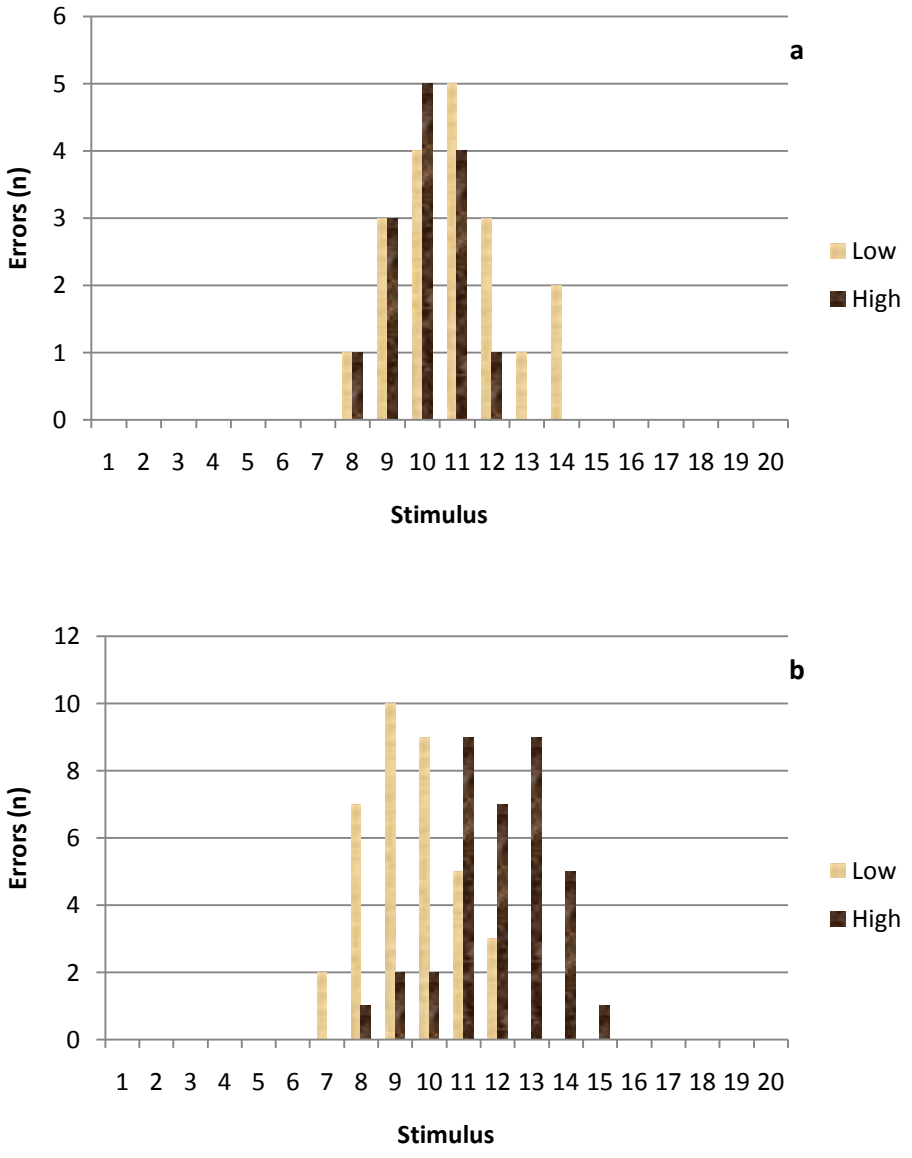


Figure 19: a) *Distribution of errors for Observer 29;* **b)** *Distribution of errors for Observer 44. Both distributions reflect errors made using $TTKR_e$ under easy conditions. The optimum criterion is located between tones 10 and 11.*

Figures19a and 19b display the error distributions for each observer. Increased criterion fluctuation is immediately apparent in Figure 19b. Though the high and low errors overlap significantly for Observer 29 (Figure 19a), errors remain largely confined to the area of

overlap, and were centrally located around the optimal criterion (between tones 10-11). However, had the criterion been fixed at optimal, high errors would only occur to the left of the criterion, and low errors to the right. Observer 44 had a wider fluctuating criterion, evidenced by increased error frequencies for tones toward the tails of the overlap, though errors largely remained confined to this area.

Condition: d'_{th} = 3 ideal. Observer 23 (*d'_{ob} = 2.54; c = -0.07*) provided the best performer in this condition, with Observer 42 (*d'_{ob} = 1.71; c = -0.04*) performing less well.

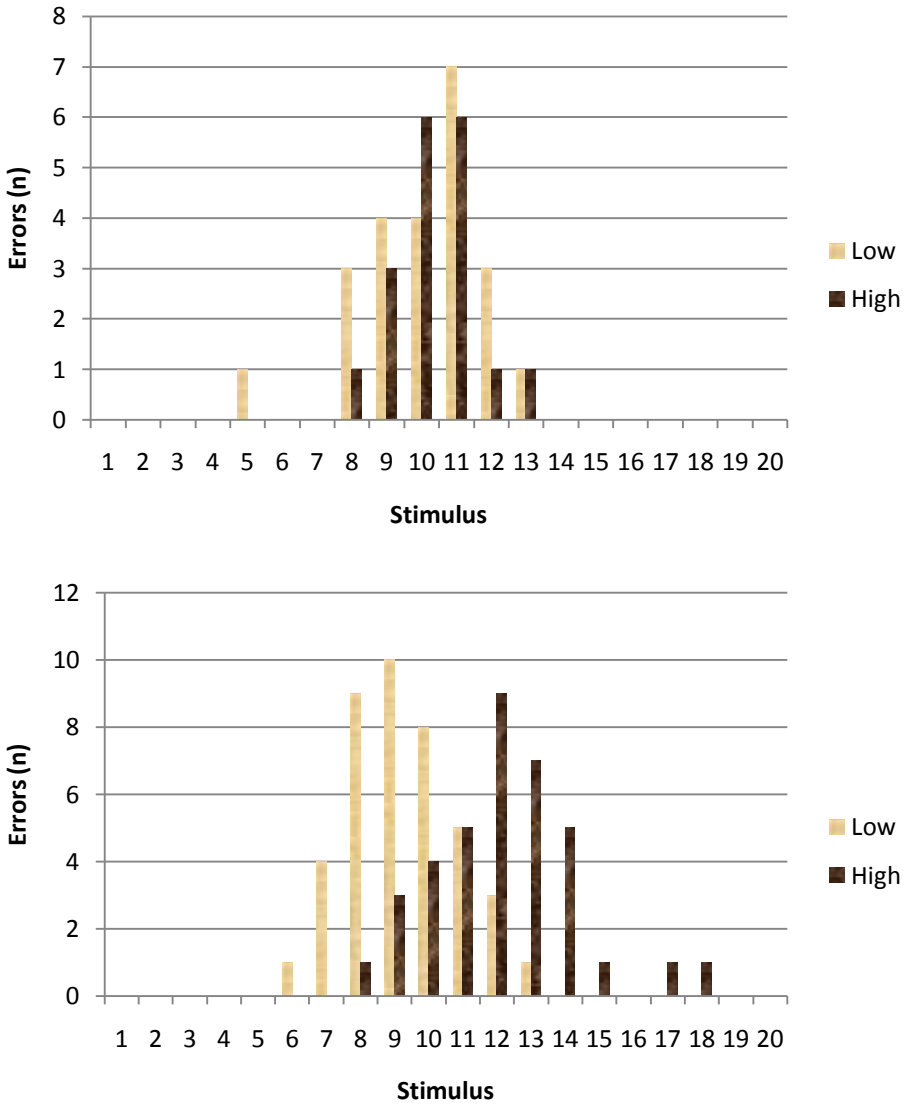


Figure 20: a) *Distribution of errors for Observer 23.;* **b)** *Distribution of errors for Observer 42. Both distributions reflect errors made using $TTKR_i$ under easy conditions. The optimum criterion is located between tones 10 and 11.*

Figures 20a and 20b display the error distribution for each observer. The distributions are very similar to the observers in the easy $TTKR_e$ condition. Observer 23 (Figure 20a) had a more attenuated error spread than Observer 42. There was overlap between the high and low distributions for Observer 23, but they remain confined to the area of overlap, and the

frequencies of the errors were located around the optimal criterion. The error distribution for Observer 42 illustrates a high degree of criterion fluctuation. In much the same fashion as the behaviours of the criterion for Observer 44, the criterion for Observer 42 in Figure 20b shifts over a surprisingly wide proportion of the decision axis.

The similarity of the error distributions and the performances in the easy condition were expected. The CVM suggested that under easier conditions criterion fluctuation would be minimised, and performance would enhance. What the observers in the easy condition illustrate though, is that if noise is entering the system then performance will be affected and errors will increase. For example, compare the error distributions and accuracy measures for Observer 44 ($d'_{ob} = 1.83$) and Observer 29 ($d'_{ob} = 2.78$). The error distributions graphically illustrate the role of criterion instability. Lower d'_{ob} values and overall accuracy were consistent with an increase in both spread and frequency of errors. These error distributions highlight the effect of criterion instability, and how it impacts upon observer accuracy, even in a binary discrimination task.

As mentioned previously, these error distributions provide support for Hypothesis 2, which states that task difficulty and type of KR will interact. $TTKR_i$ was expected to improve performance by limiting the amount of criterion fluctuation. To illustrate the difference between $TTKR_e$ and $TTKR_i$ a comparison was made between the distributions for Observer 38 ($TTKR_e$, Hard) and Observer 27 ($TTKR_i$, Hard). Figures 21a and 21b (on the next page) compare the error distributions.

The $TTKR_i$ distribution (Figure 21b, Observer 27) was less dense and reflects errors confined to predominantly the midpoint regions, and errors drop off for tones further away from this region. Furthermore the high and low errors remain largely separate. This is indicative of a criterion that remained in a relatively central location. In comparison, the $TTKR_e$ distribution (Figure 17a, Observer 38) was much denser due to high and low errors across the whole range of overlap and the frequency of errors increased for tones further away from the midpoint. This indicates that the criterion was fluctuating greatly. These distributions lend support to the CVM, demonstrating that criterion fluctuation under hard conditions can be reduced by the type of KR that is used. In this case $TTKR_i$ has reduced the amount of criterion fluctuation.

To summarise these findings, the plots of the error distributions not only indicate that criterion fluctuation existed in the present simple decision task, but that decision criteria can fluctuate over a very wide range. One can only assume that the decision process in a conventional signal detection task will generate similar fluctuations. The next section assesses Hypothesis 2, and analyses the interaction between task difficulty and type of KR.

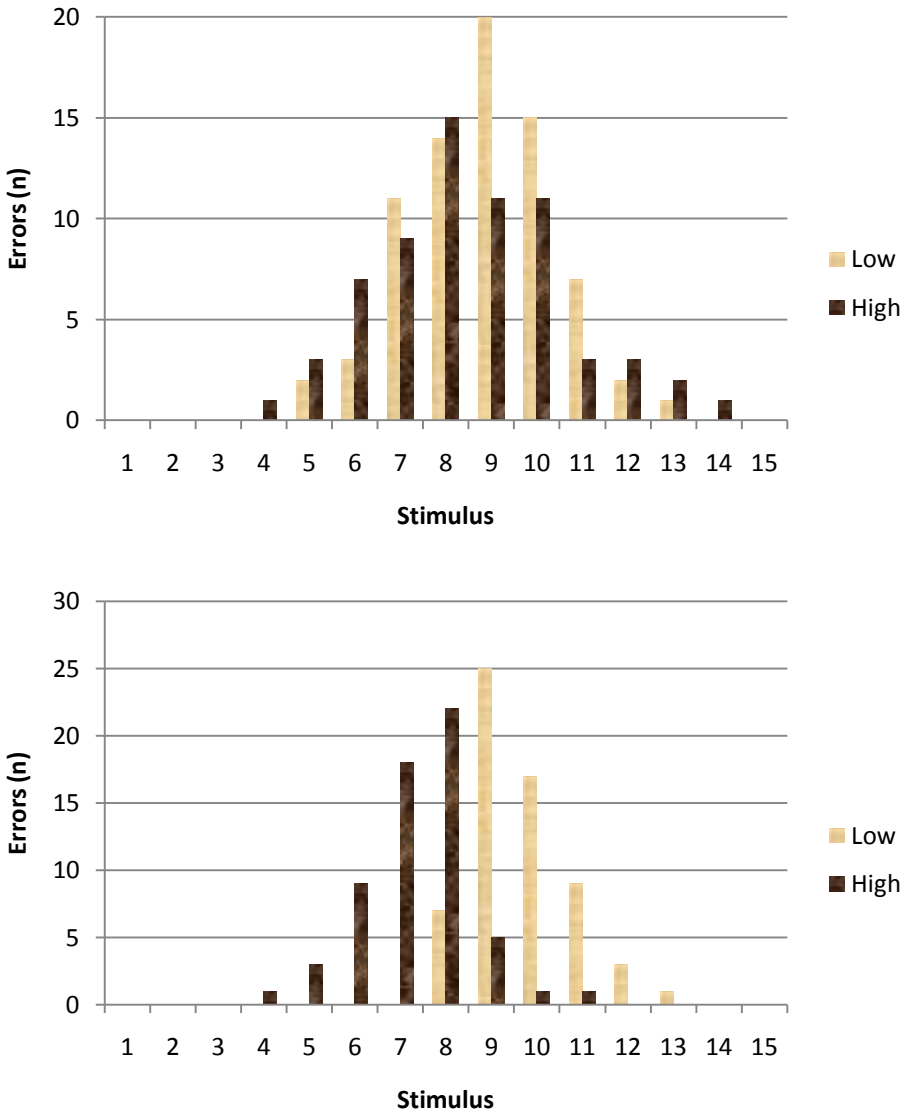


Figure 21: a) *Distribution of errors for Observer 38. This error distribution was generated using* $TTKR_e$ *under hard conditions;* **b)** *Distribution of errors for Observer 27. This error distribution was generated using* $TTKR_i$ *under hard conditions. The optimum criterion is located between tones 8 and 9.*

Hypothesis 2 (Interaction)

Hypotheses 2 predicted an interactive effect for type of KR and task difficulty. Specifically, $TTKR_i$ will enable more accurate decision making than $TTKR_e$, but only for a difficult decision task. $TTKR_i$ and $TTKR_e$ error distributions have already demonstrated that criterion fluctuation is reduced when $TTKR_i$ is delivered to observers who have to make a hard decision. Furthermore, the interaction is shown graphically in Figures 15a and 15b (p. 47), which approximate the prediction made in Figure 10 (p. 36). Separate ANOVAs, one for d'_{ob} and the other for A', were conducted to assess this interaction. All statistical analysis was conducted using *Statistical Package for the Social Sciences (SPSS)* version 17.0 (SPSS Inc., 2008).

ANOVA test (d'_{ob}) The interactive effect did not yield a significant result ($p \leq 0.5$), though the interaction bordered on significance, $F(1, 40) = 3.35$, $p = 0.08$, $\eta^2 = 0.006$ ($\eta^2_p = 0.08$) [12]. Only the main effect for task difficulty was qualified, $F(1, 40) = 524.3$, $p < 0.05$, $\eta^2 = 0.92$ ($\eta^2_p = 0.93$), as the main effect for type of KR did not reach significance, $F(1, 40) = 1.68$, $p = 0.20$, $\eta^2 = 0.003$ ($\eta^2_p = 0.04$).

The lack of a statistically significant result was likely due to inadequate statistical power (*SP*). When this was investigated, the statistical power was insufficient to detect both the main effect of KR type ($SP = 0.24$) and the interactive effect ($SP = 0.43$). Power for the effect of task difficulty ($SP \approx 1$) was more than adequate.

ANOVA test (A') For the non-parametric measure, *A'*, the interactive effect between task difficulty and type of KR was statistically significant, $F(1, 40) = 13.18$, $p < 0.05$, $\eta^2 = 0.01$ ($\eta^2_p = 0.25$). This interaction qualifies the main effect for type of KR, $F(1, 40) = 10.8$, $p < 0.05$, $\eta^2 = 0.01$ ($\eta^2_p = 0.21$). As expected task difficulty was significant, $F(1, 40) = 1023.1$, $p < 0.05$, $\eta^2 = 0.94$ ($\eta^2_p = 0.96$). This result indicates that $TTKR_i$ enabled better performance than $TTKR_e$. Moreover, the beneficial effects of $TTKR_i$ are stronger compared to $TTKR_e$ when the difficulty of the task is increased.

Post Hoc tests

Post hoc tests were conducted to evaluate the simple main effects.

t-tests (d'_{ob}) Mean differences between KR groups in the hard condition revealed a statistically significant difference, $t(20) = 6.43$, $p < 0.05$, d [13] $= 2.7$. As expected, task difficulty yielded a large effect size. The mean differences between the KR groups in the easy condition were not significant $t(20) = 0.28$, $p < .79$, $d = 0.1$. This result demonstrates that $TTKR_i$ enables better performance than $TTKR_e$ under difficult conditions, but not under the easy condition. These results verify the prediction made by the CVM.

t-tests (A') *t*-tests using the *A'* dependent measure support the results found for d'_{ob}. Mean differences between KR groups in the hard condition were statistically significant, $t(20) = 6.47$, $p < 0.05$, $d = 2.4$. However, mean differences between KR groups in the easy condition were not significant, $t(20) = 0.21$, $p < 0.84$, $d \approx 0$. These results confirm the

[12] SPSS provides partial eta squared as the estimate of effect size; however, because partial eta squared can theoretically account for more than 100% of the total variance, eta squared has been reported to account for variance. Eta squared was calculated using the SPSS ANOVA output (Levine & Hullet, 2002). See Appendix F on how to do this. However, partial eta has also been reported as many studies accept it as a measure of effect size.

[13] *d* is the effect size associated with group mean differences (Cohen, 1988).

interactive effect observed in the ANOVA test, and further support the theory that $TTKR_i$ enhances performance under difficult conditions.

Auto- Correlation Analysis

For the CVM theory to be properly tested, observers had to use the KR consistently. Though absolute attention cannot be expected, environmental factors may have been implemented by observers. In order to investigate such effects an auto-correlation analysis was undertaken.

Auto-correlation This type of correlation is a time series analysis in which a series is correlated with itself at specific time lags, or delays. The presence of sequential dependencies can be investigated using such analysis, and the degree to which previous responses relate to impending decisions can be estimated (see Chapter 2 for a brief review on sequential dependencies). In the present study each observer's response distribution was correlated with itself at specific trial lags (k). At a lag of zero ($k = 0$) the auto-correlation function (r_k; *ACF*) will be 1, but will decrease as the trial lag increases. The size of the ACF at specific lags indicates the degree to which prior responses influenced the response on the current trial. ACFs were recorded for $k = 0, k\text{-}1,...k\text{-}5$ lags. However, because significant correlations occurred only for *k-1*, correlations for lags *k-2* to *k-5* can be found in Appendix G.

Lag *k-1* correlations for all observers across all condition are shown in Table 2. Sequential dependencies went no further back than one trial. The analysis revealed a minimal presence of sequential dependencies for the hard events condition mean ($r_{k-1} = -0.09$), whereas the ideal condition had a slightly elevated ACF mean ($r_{k-1} = -0.13$). The easy condition produced higher ACFs, suggesting a slightly increased reliance on the trial *n-1* response for the trial *n* response compared to that of the hard group. The mean ACF between the events and ideal groups, $r_{k-1} = -0.024$, and $r_{k-1} = -0.023$, respectively, did not differ. The analysis indicates that sequential dependencies were more evident in the easy condition than in the hard condition, suggesting that in the easier condition there was some reliance on previous responses. This may indicate that the KR was not was not being followed all the time, though other factors may be responsible for the increased ACF values. In any event, the small sizes of the ACFs suggest that sequential dependencies had little effect on observer performance.

Table 2: *ACF values for all observers across all conditions.*

					TTKRe Hard									
	Ob.13	*Ob.28*	*Ob.9*	*Ob.2*	*Ob.38*	*Ob.8*	*Ob.21*	*Ob.20*	*Ob.19*	*Ob.4*	*Ob.37*	*M*	*SD*	*Range*
ACF	0.00	-0.11	-0.24	-0.06	-0.05	-0.11	-0.19	-0.25	-0.17	0.05	0.09	-0.09	0.11	-0.25 - 0.09
					TTKRi Hard									
	Ob.27	*Ob.24*	*Ob.32*	*Ob.14*	*Ob.26*	*Ob.5*	*Ob.7*	*Ob.10*	*Ob.40*	*Ob.34*	*Ob.41*	*M*	*SD*	*Range*
ACF	-0.02	-0.17	-0.09	-0.07	-0.07	-0.11	-0.22	-0.20	-0.16	-0.29	-0.04	-0.13	0.08	-0.29 - -0.02
					TTKRe Easy									
	Ob.29	*Ob.44*	*Ob.3*	*Ob.22*	*Ob.33*	*Ob.1*	*Ob.18*	*Ob.16*	*Ob.6*	*Ob.17*	*Ob.39*	*M*	*SD*	*Range*
ACF	-0.28	-0.35	-0.28	-0.22	-0.30	-0.19	-0.28	-0.23	-0.15	-0.25	-0.16	-0.24	0.06	-0.35 - -0.15
					TTKRi Easy									
	Ob.23	*Ob.42*	*Ob.12*	*Ob.11*	*Ob.36*	*Ob.15*	*Ob.31*	*Ob.35*	*Ob.30*	*Ob.25*	*Ob.43*	*M*	*SD*	*Range*
ACF	-0.23	-0.28	-0.28	-0.21	-0.31	-0.15	-0.26	-0.26	-0.16	-0.20	-0.16	-0.23	0.06	-0.31 - -0.15

Chapter VI

Discussion

The CVM assumes that the decision criterion in SDT is not a fixed, unvarying quantity; it is subject to fluctuation. The model proposes that two key factors can influence the degree of fluctuation: the difficulty level of the task, and the type of KR provided. To test the CVM, two hypotheses were derived:

1. Contrary to the assumption of SDT, the decision criterion in a signal detection task is a variable rather than a fixed value on the decision axis, and is present within binary discrimination tasks; and
2. an interaction between the type of TTKR provided and the difficulty level of the task. Specifically, $TTKR_i$ will enable more accurate decision making than $TTKR_e$, but only for a difficult decision task.

Hypothesis 1 was assessed by examining the spread and density of the error distribution plots for specific observers across all four experimental conditions. Despite mean *c* values across all conditions approaching zero, the error distributions demonstrated that criterion fluctuation existed. Moreover, the criterion fluctuation existed within both easy and difficult tasks. In most cases the criterion fluctuation varied over large portions of the decision axis. This was largely mediated by the task difficulty level and the type of KR. For the ideal decision maker, the criterion is fixed at the optimal value, irrespective of task difficulty. However, for a real observer, the present results show that the criterion fluctuates, even when a relatively easy discrimination task was in use, and where the KR had little or no effect. The results demonstrate that not only does criterion variance exist, but the degree to which it fluctuates, even within a simple binary task, is significant.

There was less criterion fluctuation for $TTKR_i$ than for $TTKR_e$ because the former is feedback in terms of the optimum decision rule (that will maximise the percentage of correct responses in the present case) while the latter is veridical, telling the observer which distribution was actually sampled from. In a difficult discrimination task $TTKR_i$ provides consistent feedback for each evidence variable, irrespective of which distribution was actually sampled. This is not the case for $TTKR_e$. Feedback can seem contradictory to the observer in cases where the evidence presented could indicate either that the high distribution or the low distribution had been sampled. As the KR delivers feedback relative to the true state of the events, the observer continually alters the criterion to avoid making similar errors (Larkin, 1971). Paradoxically, this leads to an increase in errors. It was shown that $TTKR_e$ causes the criterion to shift significantly as the observer attempted to incorporate

the KR into their decision rule. Conversely, the type of TTKR provided in an easy discrimination task had little or no effect on performance, simply because most evidence variables provide strong evidence of either a low or high event. The error rate is lowered and criterion fluctuation is minimised.

Schoeffler (1965) first suggested that $TTKR_e$ can be highly contradictory and may influence criterion fluctuation, though he produced no evidence in support of this assertion. Schoeffler argued that the contradictory nature of veridical KR may suppress estimates of observer accuracy. The present study has shown this to be the case. When task difficulty increases, the event probabilities increase for both high and low events, and the feedback becomes increasingly inconsistent. Both Larkin (1971) and Mueller and Weidemann (2008) demonstrated that decision noise is influenced by stimuli probabilities, and affects the observer error rate. The present results demonstrated that task difficulty influences criterion fluctuation by altering the probabilities associated with the evidence variables, which affected the consistency of the $TTKR_e$, and influenced the amount of errors that observers made.

Podd (1975) hypothesised that providing KR relative to the optimal criterion location would enhance observer performance. Podd demonstrated an increase in observer performance when $TTKR_i$ was provided compared to providing $TTKR_e$. Podd derived error plots similar to those found in Appendix H, and demonstrated that errors occurred across wide portions of the decision axis. When $TTKR_i$ was provided these errors were minimised significantly. The present results also demonstrate a reduction in errors when $TTKR_i$ was provided; however, the result was only found when task difficulty increased. Podd (1975) did not manipulate task difficulty in his research, though the CVM supported Podd's hypothesis.

In examining the CVM all sensory confounds were eliminated by making all the tonal stimuli easily discriminable from one another. This left only decision noise to consider in the reduction of observer performance. Variation in the criterion was shown to be considerable under certain circumstances. The implication of the present results is that similar decision noise will occur in an orthodox SDT task that produces both sensory and decision noise. Sensory noise and decision noise exist independently. Typically, orthodox SDT experiments ignore decision noise and focus solely on sensory factors. Thus, reductions in observer performance are only attributed to sensory factors. However, the present research, and indeed many other authors (e.g., Larkin, 1971), have demonstrated that decision noise spuriously affects SDT statistics. Furthermore, the amount of decision noise can be considerable, even in a simple binary decision task.

SDT needs to be modified to take account of decision noise, or a method is required to remove the effects of criterion fluctuation, and eliminate the underestimation of observer performance. In fact, one method for reducing decision noise does exist. Group Operating Characteristic (GOC; Taylor, Boven, & Whitmore, 1991) analysis effectively alleviates the problem of unique noise in psychophysical experiments. Unique noise is any extraneous source of noise outside of the experimental design, for example, heartbeat, criterion variability, and memory.

The technique is analogous to ROC analysis insofar as relative HR and FAR points are plotted for rating categories; however, the HR and FAR points in GOC analysis are generated from multiple observations. GOC curves are constructed by replicating a detection or recognition experiment multiple times. The experiment can be delivered to one observer many times or to a group of observers. In either case, the experiment is replicated several times, and the stimuli are rated on the same scale every time. The HR and FAR are consequently not calculated from a single rating of the stimuli, but on the sums of the ratings for the stimuli across replications (Taylor et al., 1991).

This technique serves to reduce the amount of unique noise in the system by taking the average performance and constructing ROC points from a group of observations. This serves to enhance sensitivity measures. Taylor et al. (1991) demonstrated that if the variance of the unique noise is large, then GOC analysis will lead to significant improvements in performance measures (for a complete review see Taylor et al., 1991).

However, this does not minimise the natural uncertainty that exists within some decision tasks. The present findings address this concern by demonstrating that the rules by which an individual operates can limit the amount of noise in the decision process. In many cases it is not feasible to replicate studies; thus, the refinement of decisions rules is a major contributing factor to the reduction of decision noise. This is not to discount the usefulness of GOC analysis. The existence of such a technique would benefit the validity, and accuracy, of psychophysical research. It is surprising that the technique is not more widely used.

Hypothesis 2 predicted an interactive effect between task difficulty and type of KR. Specifically, $TTKR_i$ will enable more accurate decision making than $TTKR_e$, but only for a difficult decision task. Already this hypothesis has received some support through evaluating Hypothesis 1, and the statistical analysis further supported the effects of KR and task difficulty. Mean accuracy measures (d'_{ob} and A') differed between KR groups in the difficult task condition, as predicted by the CVM. Furthermore, there was no effect for type of KR in the easy task condition. Though the interaction effect for d'_{ob} fell slightly short of significance (probably due to inadequate statistical power), both measures verified the

interactive effect. Post hoc *t*-tests for the simple main effects corroborated the interaction, yielding significant results and large effect sizes for the difficult task condition. The statistical analysis supports the hypothesis that task difficulty and type of KR affect observer performance.

As previously mentioned the alterations in event probabilities affect the nature of the KR; thus, when difficulty increases veridical KR becomes increasingly inconsistent as many of the evidence variables could indicate sampling from either distribution. However, the difficult task results make it very clear that a rule-based type of feedback can be much superior to veridical feedback. Judging events in relation to a fixed decision rule yielded higher performance estimates through minimising the movement of the criterion. $TTKR_i$ approximated the fixed criterion assumption, but demonstrated that even ideal feedback does not completely eliminate criterion fluctuation.

Prior to the present research, only one study had previously investigated the interaction between task difficulty and type of KR. Richards-Ward (1992) demonstrated that type of KR had an effect on performance when task difficulty varied. Unfortunately, his results were equivocal. While Richards-Ward did demonstrate that $TTKR_i$ improved observer performance in a hard discrimination task compared to $TTKR_e$, the effect was stronger for the easy discrimination task. Furthermore, when $TTKR_e$ was provided to observers in the difficult task condition their performance improved. The CVM demonstrated that under difficult conditions $TTKR_e$ was less conducive to observer performance, and observed minimal effects in the easy decision task. However, the CVM aligns with Richards-Wards' (1992) study insofar as $TTKR_i$ improved performance compared to $TTKR_e$.

The major implication the present results have is that the type of TTKR used becomes increasingly important as the task becomes more difficult. The KR that is usually provided in SDT is veridical ($TTKR_e$). This type of KR is likely to induce more and more criterion fluctuation as the difficulty level of the task increases. In difficult signal detection tasks (which reflects a low SNR), values of detectability parameters, such as d', and area under the ROC curve, will inherently have relatively small values due to the low SNR. These already minimised values can be further reduced by the criterion being free to fluctuate, and are affected more than the larger detectability values that accompany an easy task.

In other words, providing KR in terms of the actual state of affairs has increasingly adverse effects on detection parameters as the detection task becomes more difficult. Estimating the degree to which decision noise is affecting measures such a d' is problematic. If the degree of criterion fluctuation was constant then this would be possible; however, the present results demonstrated that criterion fluctuation was not constant, and varied across levels of difficulty.

Thus, $TTKR_e$ will not only cause spuriously low estimates of detectability, but the degree of suppression depends on task difficulty. It appears that when the SNR is small, detectability indices will be more affected than for larger SNRs.

Paradoxically, where there is no suitable model to produce the ideal decision sequence from, so that $TTKR_i$ cannot be given, it might be best to provide no feedback at all (Han & Dobbins, 2008), especially in difficult detection tasks. The differential effect that TTKR has on detection parameters cannot be ignored, despite it being given with a view to simply help motivate the observer. Feedback is an independent variable that may confound the outcome of a study, especially where the detection task has several levels of difficulty (e.g., see Green & Swets, 1966).

Strong support for the CVM was generated through the verification of both hypotheses. Not only does criterion fluctuation exist, but it exists in even the simplest of detection tasks, and varies when task difficulty varies. The results also show that a certain type of feedback can reduce the amount of criterion fluctuation, but that amount is partly dependent upon the difficulty level of the task.

Supplementary Analysis

Additional analysis was undertaken to be sure that the effects presently described were attributable to the type of KR and task difficulty, and not due to other factors. Another assumption made by SDT is that each trial is independent of the previous trial. That is, it is assumed that sequential dependencies (Triesman & Faulkner, 1984a) do not exist. To investigate whether any sequential effects were present an auto-correlation analysis was done. If sequential effects were present it needed to be assessed how, if at all, the different types of feedback and task difficulty level affected it.

Significant correlations were only observed as far back as the immediately preceding trial. While on the surface this suggests that sequential dependencies had some role in the decision strategy, the size of the correlations were not substantial enough to warrant any meaningful contribution. Mueller and Weidemann (2008) report that increases in sequential dependencies should occur when rating style tasks are used, and the relative lack of sequential dependencies in the present binary case is consistent with this.

There was some difference in the ACF across conditions. Lower ACFs were observed in the hard condition compared to the easy condition, though the difference was small. An auto-correlation will generally yield a correlation at a lag of one trial, but the correlation decreases for further lags. The fact that correlations existed at a lag of one trial was not surprising. The slight increase in the ACF in the easy task compared to the hard task could be expected

because the observer sequences are more similar due to the increased probability of an event truly supporting the high or low state. Furthermore, there was virtually no difference in the ACFs between the KR groups in the easy task condition. The upshot of this analysis is that sequential effects were not significant, and did not affect performance or the effects of task difficulty and type of KR.

Limitations

Data regarding how much the criterion fluctuated with no TTKR would have been useful. The lack of a no KR condition was a major limitation in the present study. A no KR condition would have provided a baseline from which to assess the degree of criterion fluctuation. Furthermore, the extent to which observers consistently used the KR could have been assessed. This assessment might have been accomplished by conducting a cross-correlation analysis, which correlates the observer sequence with the KR sequence. However, in the absence of a no KR condition this analysis was not possible. Consequently, estimates on the extent to which the KR was utilised were not possible. Such an analysis would have been particularly useful in cases where KR non adherence was suspected.

A few observers obtained perfect or near perfect scores (relative to the ideal observer). Where $TTKR_e$ has been provided it is unreasonable to accept that an observer can achieve this level of performance if the KR was being fully utilised. These cases suggest that the KR was not being used consistently. The effect of a few observers performing in this way weakened the results, though not enough to cause the CVM predictions to fail. The fact that some observers, irrespective of KR type, performed better when the KR was effectively disregarded heightens the need to incorporate a no KR condition into the experimental design.

A further issue that may have contributed to these odd cases was the complexity of the task. The task was not overly complex, and this may have allowed observers to maintain near optimal criterion locations without the aid of the KR. The task used tones that had 12 to 16 JNDs between each frequency; a large difference in tonal frequency between adjacent tones. Consequently, the tones were rather easy to discriminate as high or low tones. One way to potentially minimise this effect is to reduce the number of JNDs between adjacent tones.

The JNDs could be reduced to around 3 and 4, minimising the distance between the tones, but still leaving tones discernable from each other (ensuring no discrimination problems between adjacent tones). This would increase the number of tones per distribution, making the task more complex, and perhaps making the observer more reliant on the KR. The level

of task difficulty would remain unchanged, though the increased number of tones within each condition may induce observers to more faithfully use the KR.

There was also reason to suspect that the instructions unduly affected observer performance. They stressed that perfect performance was impossible, and that mistakes would be made. The observers were instructed to follow the KR as it would yield the most consistent results, and the highest earning potential. Yet some observer still chose to adopt a more subjective approach, preferring their decision strategy over the KR. The specificity and depth of the present instructions compared to previous research (Podd, 1975; Richards-Ward, 1992) was greater, and this may have alerted observers to some perceived deception, despite no actual deception being used in the experiment.

The specificity of the instructional content was an attempt to clarify the task requirement. Richards-Ward's (1992) study also provided evidence of KR non-adherence, and it was with this in mind the present study stressed the need to use the KR consistently. However, the instructions may have been overly specific, and the repeated pleas to be sure to use the feedback may have made some observers suspicious of deception.

In light of these limitations the inclusion of a no KR condition is paramount. In order to verify the effect of type of KR, and the address the degree of KR adherence, a no KR condition allows for this by acting as a control. The sample size used to assess the CVM was rather small, yielding non significant results in some cases. Finally, less emphasise needs to be placed on the instructional content; these need refining. Reversion to the original format and content of Podd's (1975) simple instructions may facilitate increased adherence, and minimise any biases.

Conclusion and Future Directions

The type of KR an observer receives impacts upon task performance. TTKR is not something one can use or not use as one pleases. It affects performance but does not do so in a consistent fashion, and has a greater influence when task difficulty increases. The standard of providing $TTKR_e$ in SDT experiments, and ignoring the KR as a variable in its own right, is called into question. Some strong support was provided for the CVM, though the model needs further testing and refinement in order to ensure strong and consistent findings.

The theories of Sheoffler's (1965) and Larkin's (1971) suggest that when KR is present the criterion will invariably shift toward optimality, and this indeed appeared to be the case. Though in order for optimal performance to be obtained the criterion needs to be stabilised at one location. Even when the criterion shifts toward the optimum point, there is likely movement around that point. Furthermore, shifts are required to reach that position. Thus,

the past shifts in the criterion bear upon the performance, and not its final location. The fact that criterion fluctuation was so prevalent even within a simple binary task bears heavily on SDT research designs. Fluctuation in the criterion translates into estimates of sensitivity or discriminability that are lower than they would have been if there was no criterion fluctuation.

This study has illustrated that criterion fluctuation exists, and can be minimised or exacerbated depending on the type of feedback provided and the difficulty level of the task. The independent existence and generation of decision noise from sensory noise only serves to further decrease estimates of observer performance. In traditional psychophysical designs where sensory abilities are of primary focus, noise is not only being generated by sensory variables, but also decision variables, which further suppress SDT statistics. The roles of KR and task difficulty have to be accommodated in order to reduce the amount of noise entering at the decision stage, and confounding estimates of sensitivity.

The CVM model has implications not only for SDT, but any other environment that operates on a rule-based system. Choices are made all the time with less than complete information. The world we live in is at most times highly unpredictable and confusing, yet we must make decisions amidst the uncertainty. Though feedback can enhance our decision making, this is not always the case.

This present study has shown that providing simple correct/incorrect answers is not always effective in improving performance. The inherent nature about decision making with less than complete information is that errors will occur; they cannot be avoided. The best that we can do is to minimise these errors in some way. $TTKR_i$ demonstrated that when KR is provided according to a fixed rule, the KR is more consistent and performance is improved. When uncertainty is high, making decisions based on the optimal rule will result in more correct decisions being made than when providing veridical feedback and its tendency to induce criterion fluctuation.

Whether it is the classroom or the hospital, using a consistent decision rule to help minimise errors may yield improved results compared to veridical feedback. The idea of using the ideal observer in gauging performance and accuracy is not uncommon (e.g., Green and Swets, 1966. Geisler (2003) is a more recent example of a model which incorporates an ideal observer); however, delivering KR relative to the ideal observer reflects a completely different approach to demonstrating criterion variance.

The CVM is one of few models that have investigated how different types of feedback and task difficulty may interact to affect performance. Despite several design issues, the present

results highlight the potential detrimental effects of KR on performance, particularly under conditions of high uncertainty. Future SDT studies need to remain mindful of such effects. In a conventional SDT study, both sensory noise and decision noise are present. The decision noise is very rarely considered with researchers clinging to the assumption that the decision criterion has no associated variability. This assumption is incorrect. The present study clearly shows that matters can be made worse by giving veridical trial-by-trial feedback, especially where a signal detection task is inherently difficult.

References

Balakrishnan, J. D. (1998a). Measures and interpretations of vigilance performance: evidence against the decision criterion. *Human Factors, 40,* 601-623.

Balakrishnan, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods, 3,* 68-90.

Balakrishnan, J. D. (1999). Decision processes in discrimination: fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 1189-1206.

Balakrishnan, J. D., & MacDonald, J. A. (2002). Decision criteria do not shift: Reply to Triesman. *Psychonomic Bulletin & Review, 9,* 858-865.

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

Blackman, R. B., & Tukey, J. W. (1959). "Particular Pairs of Windows." In *The measurement of power spectra, from the point of view of communications engineering.* New York: Dover.

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 587-599.

Clark, W. C., & Greenberg, D. B. (1971). Effect of stress, knowledge of results, and proactive inhibition on verbal recognition memory (*d'*) and response criterion (l_x). *Journal of Personality and Social Psychology, 17,* 42-47.

Clark, W. C., & Mehl, L. (1973). Signal detection theory procedures are not equivalent when thermal stimuli are judged. *Journal of Experimental Psychology, 97,* 148-153.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. (2nd ed.). Lawrence Erlbaum Associates: New Jersey.

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

Geisler, W. S. (2003) Ideal Observer Analysis. In L. Chalupa and J. Werner (Eds.), *The visual neurosciences,* 825-837. Boston: MIT press.

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

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: evidence for adaptive criterion learning. *Memory & Cognition, 36,* 703-715.

Kadlec, H. (1999). Statistical properties of *d'* and β estimates of signal detection theory. *Psychological Methods, 4,* 22-43.

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

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

Levine, T. R. & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28,* 612-625.

Lopez-Bascuas, L. E. (2008). Signal densities and criterion variance in speech and non-speech perception. *The Journal of the Acoustical Society of America, 123,* 3733-3736.

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

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.

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

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

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

Pierce, J. R. (1980). *An introduction to information theory: symbols, signals, and noise* (2nd ed.). New York: Dover.

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.

Richards-Ward, L. A. (1992). *The effect of knowledge of results and task difficulty on a binary discrimination task: the problem of criterion variability in signal detection theory.* Unpublished Honours Project, Massey University, Palmerston North, New Zealand.

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.

Ryan, L. J., & Fritz, M. A. (2007). Erroneous knowledge of results affects decision and memory processes on timing tasks. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1468-1482.

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.

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

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

Speeth, S. D., & Mathhews, M. V. (1961). Sequential effects in the signal-detection situation. *Journal of the Acoustical Society of America, 33,* 1046-1053.

SPSS for Windows, Rel. 17.0.0. (2008). Chicago: SPSS Inc.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behaviour Research Methods, Instruments, & Computers, 31,* 137-149.

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

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

Taylor, A., Boven, R., & Whitmore, J. (1991). Reduction of unique noise in the psychophysics of hearing by group operating characteristic analysis. *Psychological Bulletin, 109,* 133-146.

Thurstone, L. L. (1927a). Psychophysical analysis. *American Journal of Psychology, 38,* 368-389.

Thurstone, L. L. (1927b). A law of comparative judgement. *Psychological Review, 34,* 273-286.

Triesman, M., & Faulkner, A. (1984a). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68-111.

Triesman, M., & Faulkner, A. (1984b). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 119-139.

Triesman, M., & Faulkner, A. (1985). Can decision criteria interchange locations? Some positive evidence. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 187-208.

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

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

Appendices

Appendix A

Glossary of Signal Detection Equations

SDT Statistics

$$HR = n(S) / N(s)$$

$$FAR = n(S) / N(n)$$

$$z(HR) = \frac{c - \mu_s}{\sigma_s}$$

$$z(FAR) = \frac{c - \mu_n}{\sigma_n}$$

Sensitivity Measures

$$d' = \frac{\mu_s - \mu_n}{\sigma_n}$$

$$d' = z(HR) - z(FAR)$$

$$A_z = \phi(d_a / \sqrt{2})$$

$$A_g = \frac{1}{2}\sum (FAR_{i+1} - FAR_i)(HR_{i+1} + HR_i)$$

$$A' = \frac{1}{2} + \frac{(HR - FAR)(1 + HR - FAR)}{4HR(1 - FAR)} \qquad If\ HR \geq FAR$$

$$A' = \frac{1}{2} + \frac{(FAR - HR)(1 + FAR - HR)}{4FAR(1 - HR)} \qquad If\ HR \leq FAR$$

Response Bias Measures

$$l(x) = \beta = \frac{f(x \mid s)}{f(x \mid n)}$$

$$c = -\frac{1}{2}[z(HR) + z(FAR)]$$

Appendix B

Stimuli: Tonal Frequencies

Table B.1: *Tonal frequencies associated with the stimuli from low (tone 1) to high (tone 20). Frequencies are in Hz.*

Stimulus	*Tone Frequency (Hz)*
1	381.8
2	424.4
3	466.6
4	509.0
5	551.4
6	593.8
7	636.2
8	678.6
9	721.0
10	763.4
11	805.8
12	848.2
13	890.6
14	933.0
15	975.4
16	1017.8
17	1060.2
18	1102.6
19	1145.0
20	1187.4

Appendix C

Information Sheet

MASSEY UNIVERSITY
COLLEGE OF HUMANITIES
AND SOCIAL SCIENCES
TE KURA PŪKENGA TANGATA

The Effects of Knowledge of Results in a Binary Decision Task

INFORMATION SHEET

Greetings!

My name is Rob Taylor and I am a graduate student at the School of Psychology, Massey University. The research I am conducting is in fulfilment of my Masters degree, and am seeking to investigate the effects that task difficulty and feedback (knowledge of results) has on our decision-making abilities. The research is being supervised by John Podd, a lecturer here at Massey University.

Why is this Research Important?

Decision-making is an everyday occurrence, and sometimes occurs when limited information is available. Uncertainty affects our ability to make sound decisions, and often we rely on our own decision rules and feedback from our environment to aid us. This happens not only in everyday life, but also within clinical and medical fields where a diagnosis must be made. The fact is that feedback may alter our decision rules, and sometimes this can have a damaging effect.

The present study investigates how feedback affects decision-making when you are presented with insufficient information to make a decision with certainty.

Who are we Looking for?

The research requires the assistance of *HIGHLY MOTIVATED* individuals. The task requires a fair amount of vigilance and tenacity, and because it uses auditory stimuli, you should have normal hearing.

What will happen During the Experiment?

You will undertake an auditory discrimination task, in which a decision must be made about whether the tone presented was high or low. The tones will be delivered via a computer and you make your response with the mouse by clicking on one of two boxes on the computer screen. Feedback will be provided following the presentation of each tone. The experiment is about ***30 minutes long***, and you will have a short break after completing half the trials. For your time you will be reimbursed **$10**, ***subject to completion*** of the experiment.

All information that you provide, including individual results, will be anonymous and will be only used for the purpose of the present investigation. At no stage will any individual person be identified in any report pertaining to this study.

This project 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(s) named above are responsible for the ethical conduct of this research. If you have any concerns about the conduct of this research that you wish to raise with someone other than the researcher(s), please contact Professor John O'Neill, Director, Research Ethics, telephone 06 350 5249, email humanethics@massey.ac.nz.

Invitation

If you would like to participate in this study please let me know (via email) as soon as possible, and sign the provided consent form. Please bring the signed consent form with you when you come to do the study. To summarise your rights, please note that you are under no obligation to accept this invitation. If you decide to participate, you have the right to:

- decline to answer any particular question;
- withdraw from the study (before completion of the experiment);
- ask any questions about the study at any time during participation;
- provide information on the understanding that your name will not be used unless you give permission to the researcher;
- be given access to a summary of the project findings when it is concluded.

Contact

If you would like any further information or have any questions about the study please do not hesitate to contact John Podd or myself:

Rob Taylor
School of Psychology
Turitea, Palmerston North
taylor.rob@hotmail.com

Associate Professor John Podd
School of Psychology
Turitea, Palmerston North
Rm P3.33
Phone: +64 6 3569-099, Ext 2067
J.V.Podd@massey.ac.nz

Appendix D

Consent Form

MASSEY UNIVERSITY
COLLEGE OF HUMANITIES
AND SOCIAL SCIENCES
TE KURA PŪKENGA TANGATA

The Effects of Knowledge of Results in a Binary Decision Task

PARTICIPANT CONSENT FORM

I have read the Information Sheet and have had the details of the study explained to me. My questions have been answered to my satisfaction, and I understand that I may ask further questions at any time.

I agree to provide information to the researcher on the understanding that it is completely confidential. I understand that I am free to withdraw from the study at any time, or to decline to answer any particular question in the study.

I agree to participate in this study under the conditions set out in the Information Sheet.

Signature: ..... **Date:** .....

Full Name - printed .....

Appendix E

Instructions

Verbal Instructions: Played to participants prior to experiment.

Hello and Welcome.

We are investigating the effect of knowledge of results – feedback - on task performance. Your task is to make a decision about a tone you hear on each trial, deciding whether you think it is high or low. It is impossible to provide a correct answer on every trial because of the way the study is set up. The feedback you will receive on each trial is designed to help you perform at the best possible level; so please, pay attention to the green feedback light on each trial. This feedback light tells you what the correct response was by lighting up the box you should have clicked on. To score the maximum number of correct responses it is crucial that you use the feedback provided on each trial.

You are about to run through some training trials. These trials are to familiarise you with the pace and format of the trials. First of all you will hear the full range of tones used in the experiment. Please listen carefully. As soon as the tones have played you will be presented with the response box, which will have two yellow boxes lit up. This indicates that you have 2 seconds before the training begins. Use the mouse to click on the left box if you think the tone is low and on the right box if you think the tone is high. Initially, the trials will seem to be going very fast, but after a few trials you will get accustomed to the pace.

Remember! To maximise the number of correct decisions you make, you must use the feedback lights which tell you the correct response for each trial. The investigation is primarily concerned with how feedback affects our decisions, so please use the feedback.

To increase your motivation and to get you to pay attention to the feedback lights, during the main trials you will have the potential to earn an additional $8 over and above the $10 you will automatically receive for taking part in the study. The way to earn this is to maximise your percentage of correct decisions. For every correct decision you will earn 2c; however, for every incorrect decision you will lose 2c. The best way to maximise your earnings is to follow the feedback. Try to avoid making personal judgements about the tones, and adhere to the feedback, as it will yield the highest earning potential.

After 200 trials there will be a short break. Take a minute or two to compose yourself at this point. Maybe stretch and walk around a little. Upon clicking okay to continue, there will be no 2 second warning. Trials start straight off, so please be ready for this.

To score well, the experiment requires you to remain highly motivated and attentive throughout. If for any reason you need to stop, please press the spacebar to pause the trials. Once you are ready to commence, click the okay button.

Finally, here are a few tips when performing the task. Between trials, return the mouse arrow to between the two response buttons. This will allow minimal movement and less time to respond. Lastly, follow the feedback. You will not perform perfectly but following the feedback allows you to optimise your performance.

DO YOU HAVE ANY QUESTIONS?

Training Instructions: Displayed on screen prior to training trials. Words in bold appeared bold on screen.

These training trials will help you become familiar with the pace and format of the experimental trials. Please read the following instructions before beginning.

For the following **50** trials you will hear a series of tones. Your task is to decide whether the tone you hear is **high** or **low**. Two boxes will appear in the centre of the screen, one for a high tone decision and one for a low tone decision. Using your mouse, please click on the appropriate box after the tone has been played.

Upon clicking the "Continue" button below the range of possible tones will immediately play. Please be attentive as it is essential that you familiarize yourself with the range. Once the tones have played, two yellow boxes will appear in the response box. This indicates that the trials will begin in **2 seconds**.

During the trials you will have **2 seconds** after the tone is played in which to respond. Once you have responded the next trial will immediately begin.

Once the training trials are complete a box will appear. When this occurs please **raise your hand** to let the examiner know. With the researchers approval you may then continue onto the main trials.

Remember, these trials are simply designed to familiarize you with the pace and format of the main trials. Consequently, no feedback is offered during the training.

Do you have any questions before you start?

Main Trial Instructions: Displayed on screen prior to main trials. Words in bold appeared bold on screen.

The previous training trials familiarized you with the pace and format of the task. The following **400** experimental trials are exactly the same as the training trials; however this time the feedback **will aid you in optimizing your performance**.

Your goal is to make as many correct decisions as you can. The green feedback light will

flash above the response buttons to indicate whether the choice you made was correct. For example, if you click on the high button, and the light flashes over the low button, you have made an incorrect response.

It is essential that you understand these tasks are **very difficult**. Consequently you will not correctly categorize every tone on every trial. The feedback is designed to help you achieve the **best possible outcome**, so it is imperative that you take note of it.

The probability of a high tone, or a low tone, is 50% for each trial. This means that each trial is independent, and there is no benefit in trying to guess what tone may be presented next based on previous tones you have heard. Rather, **utilise the feedback to improve your decision making**.

As with the training trials, upon clicking the "continue" button you will be given a **2 second** warning before the trials begin. Additionally, you will again have **2 seconds** to respond after hearing the tone. After **200** trials a box will pop up indicating that you are half way through the trials, at which point you may rest, stretch, or get up for a couple of minutes. When you are ready to proceed click the "okay" button to commence the next series of trials.

Try and respond on every trial. If for any reason you are uncertain and have to make a guess, please try and spread your guesses equally across the high and low responses. And finally, it cannot be stressed enough the importance of following the feedback. Remember, the feedback light will help you achieve **optimal results**.

Do you have any questions before you start?

Appendix F

ANOVA Tables and Calculation of Eta Square

Table F.1: *ANOVA table for d'_{ob} with eta squared calculated.*

Source of Variation	***df***	***SS***	***MS***	***F***	η^2
Between					
Difficulty	1	20.51	20.51	524.31	0.92
Feedback	1	0.066	0.066	1.68	0.003
Interaction	1	0.131	0.131	3.35	0.006
Within (error)	40	1.565	0.039		0.07
Total	43	22.27			

Table F.2: *ANOVA table for A' with eta squared calculated.*

Source of Variation	***df***	***SS***	***MS***	***F***	η^2
Between					
Difficulty	1	0.311	0.311	1023.17	0.94
Feedback	1	0.003	0.003	10.79	0.009
Interaction	1	0.004	0.004	13.18	0.01
Within (error)	40	0.12	0.00		0.04
Total	43	0.331			

The calculation of eta squared

The default estimate of effect size provided by SPSS is partial eta squared. The issue with the estimate is that it effectively accounts for more than 100% of the total variance. Eta squared is the preferred alternative, and provides an estimate of the variance accounted for by a specific variable. Levine and Hullett (2002) provide the formula for eta squared:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

Conveniently, eta squared can be calculated from the SPSS output tables. All that is needed is the SS for the between groups variables, and the total SS, both of which are provided by SPSS. For example, using table C.1, difficulty accounts for 92% of the total variance, where:

$$\eta^2(difficulty) = \frac{20.51}{22.27} = 0.92$$

Appendix G

Auto-Correlation Tables

Table G.1: *Auto-correlations for all observers in condition $d'_{th} = 1$, events KR.*

Lag	*Ob.13*	*Ob.28*	*Ob.9*	*Ob.2*	*Ob.38*	*Ob.8*	*Ob.21*	*Ob.20*	*Ob.19*	*Ob.4*	*Ob.37*	*M*	*SD*	*Range*
1	0.00	-0.11	-0.24	-0.06	-0.05	-0.11	-0.19	-0.25	-0.17	0.05	0.09	-0.09	0.11	-0.25 - 0.09
2	-0.06	0.00	0.02	-0.09	-0.03	-0.01	0.14	-0.01	-0.08	0.00	0.07	-0.01	0.07	-0.09 - 0.14
3	0.11	0.09	0.04	0.00	0.01	0.06	-0.10	0.00	0.04	0.08	0.11	0.04	0.06	-0.10 - 0.11
4	0.05	-0.07	0.00	-0.01	0.04	-0.03	0.05	0.05	-0.03	-0.01	-0.08	0.00	0.05	-0.08 - 0.05
5	0.02	0.13	-0.06	-0.06	-0.06	-0.03	-0.02	-0.05	0.00	0.07	-0.08	-0.01	0.06	-0.08 - 0.13

Table G.2: *Auto-correlations for all observers in condition $d'_{th} = 1$, ideal KR.*

Lag	*Ob.27*	*Ob.24*	*Ob.32*	*Ob.14*	*Ob.26*	*Ob.5*	*Ob.7*	*Ob.10*	*Ob.40*	*Ob.34*	*Ob.41*	*M*	*SD*	*Range*
1	-0.02	-0.17	-0.09	-0.07	-0.07	-0.11	-0.22	-0.20	-0.16	-0.29	-0.04	-0.13	0.08	-0.29 - -0.02
2	0.02	-0.04	0.01	-0.07	-0.07	-0.11	-0.02	-0.01	-0.05	0.03	-0.04	-0.03	0.04	-0.11 - 0.03
3	0.15	0.08	0.10	0.15	0.14	0.09	0.04	0.01	0.07	-0.01	0.08	0.08	0.05	-0.01 - 0.15
4	-0.04	-0.09	-0.05	-0.05	0.01	0.01	-0.07	-0.04	-0.04	-0.08	-0.06	-0.05	0.03	-0.09 - 0.01
5	0.05	0.02	-0.02	0.04	0.03	0.01	-0.04	-0.03	-0.03	0.00	0.01	0.00	0.03	-0.04 - 0.05

Table G.3: *Auto-correlations for all observers in condition $d'_{th} = 3$, events KR.*

Lag	*Ob.29*	*Ob.44*	*Ob.3*	*Ob.22*	*Ob.33*	*Ob.1*	*Ob.18*	*Ob.16*	*Ob.6*	*Ob.17*	*Ob.39*	*M*	*SD*	*Range*
1	-0.28	-0.35	-0.28	-0.22	-0.30	-0.19	-0.28	-0.23	-0.15	-0.25	-0.16	-0.24	0.06	-0.35 - -0.15
2	0.01	0.03	0.01	0.08	0.06	-0.04	0.06	0.03	-0.05	-0.06	-0.05	0.01	0.05	-0.06 - 0.08
3	0.06	-0.01	0.02	0.03	0.03	0.00	0.08	0.10	0.07	0.10	0.09	0.05	0.04	-0.01 - 0.10
4	-0.01	0.06	0.04	-0.01	0.02	0.08	-0.01	0.00	0.07	0.05	-0.04	0.02	0.04	-0.04 - 0.07
5	0.05	-0.02	0.02	0.06	0.00	-0.05	0.05	0.06	0.03	0.01	0.05	0.02	0.04	-0.05 - 0.06

Table G.4: *Auto-correlations for all observers in condition $d'_{th} = 3$, ideal KR.*

Lag	*Ob.23*	*Ob.42*	*Ob.12*	*Ob.11*	*Ob.36*	*Ob.15*	*Ob.31*	*Ob.35*	*Ob.30*	*Ob.25*	*Ob.43*	*M*	*SD*	*Range*
1	-0.23	-0.28	-0.28	-0.21	-0.31	-0.15	-0.26	-0.26	-0.16	-0.20	-0.16	-0.23	0.06	-0.31 - -0.15
2	0.03	0.06	0.01	-0.04	0.06	0.01	0.06	-0.19	-0.05	-0.03	0.04	0.00	0.07	-0.19 - 0.06
3	0.07	0.05	0.05	0.07	0.04	0.11	-0.03	0.07	0.10	0.02	0.00	0.05	0.04	-0.03 - 0.11
4	0.01	0.02	-0.04	0.02	0.01	0.01	0.05	0.05	-0.01	0.07	0.04	0.02	0.03	-0.04 - 0.07
5	0.02	-0.04	0.02	-0.02	0.01	0.00	-0.04	-0.05	0.03	0.02	0.04	0.00	0.03	-0.05 - 0.04

Appendix H

Response Distributions and Error Plots

.

Events d' = 1

Table H.1: *Distribution table for Observer 2.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
High	✓						2	8	19	24	26	29	17	7	3	1
	×				1	3	7	10	11	15	13	1	1	2		
Low	✓		1	3	9	18	26	30	25	14	6	3				
	×						4	9	14	15	12	6	3	1		

Table H.2: *Distribution table for Observer 38.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
High	✓						2	9	15	28	28	27	15	7	2	1
	×				1	3	7	9	15	11	11	3	3	2	1	
Low	✓		1	3	9	16	27	28	25	10	3	2	1			
	×					2	3	11	14	20	15	7	2	1		

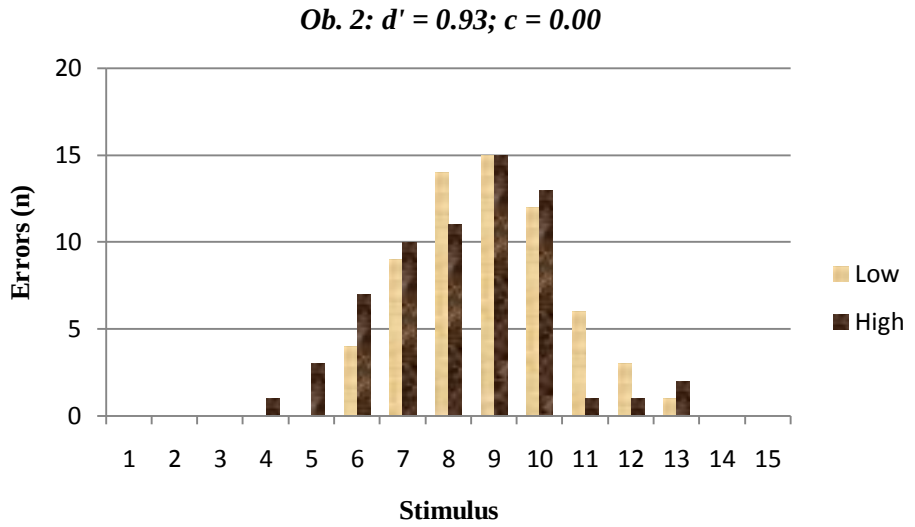


Figure H.1: *Error distributions for Observer 2, with relative statistics describing accuracy (d' and criterion location (c).*

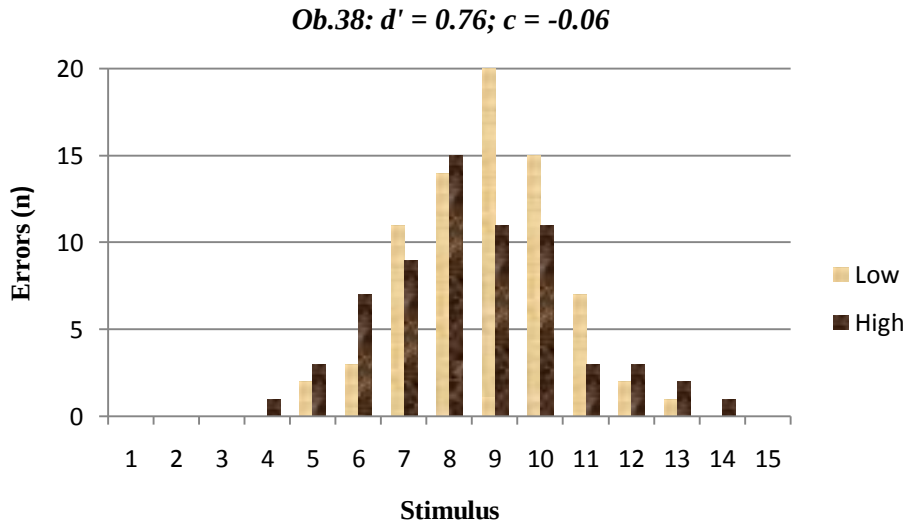


Figure H.2: *Error distributions for Observer 38, with relative statistics describing accuracy (d' and criterion location (c).*

Ideal d' = 1

Table H.3: *Distribution table for Observer 27.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
High	✓								8	34	38	29	18	9	3	1
	×				1	3	9	18	22	5	1	1				
Low	✓		1	3	9	18	30	38	31	5	1					
	×								7	25	17	9	3	1		

Table H.4: *Distribution table for Observer 24.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
High	✓						1	3	12	25	35	28	17	9	3	1
	×				1	3	8	15	18	14	4	2	1			
Low	✓		1	3	9	17	28	35	32	12	2					
	×					1	2	4	7	18	16	9	3	1		

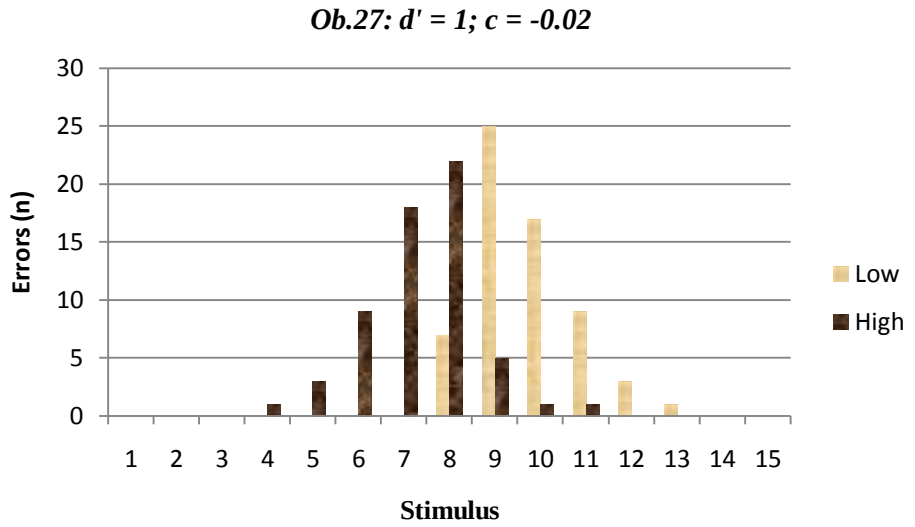


Figure H.3: *Error distributions for Observer 27, with relative statistics describing accuracy (d') and criterion location (c).*

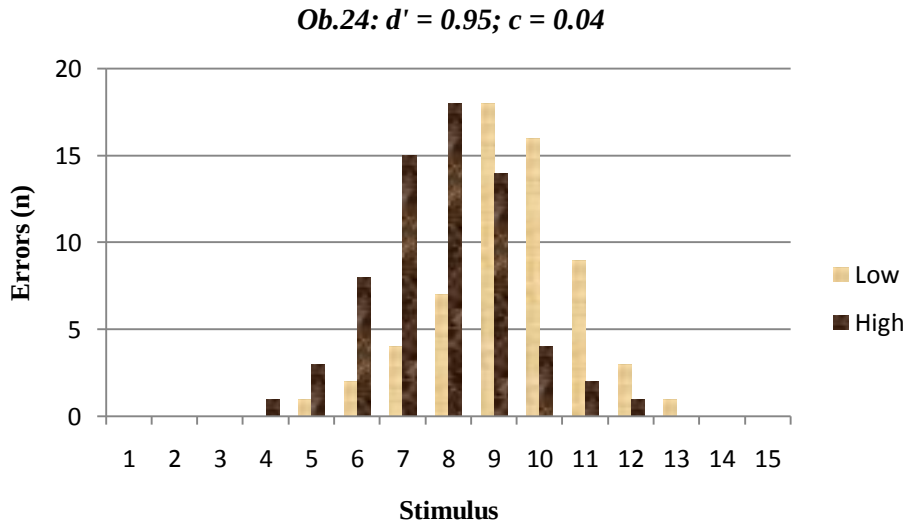


Figure H.4: *Error distributions for Observer 24, with relative statistics describing accuracy (d') and criterion location (c).*

Events d' = 3

Table H.5: *Distribution table for Observer 29.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
High	✓										4	14	29	39	37	30	18	9	3	1	
	×								1	3	5	4	1								
Low	✓		1	3	9	18	30	39	38	27	14	4									
	×								1	3	4	5	3	1	2						

Table H.6: *Distribution table for Observer 44*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
High	✓									1	7	9	23	30	34	29	19	9	3	1	
	×								1	2	2	9	7	9	5	1					
Low	✓		1	3	9	18	30	37	32	20	9	4		1							
	×							2	7	10	9	5	3								

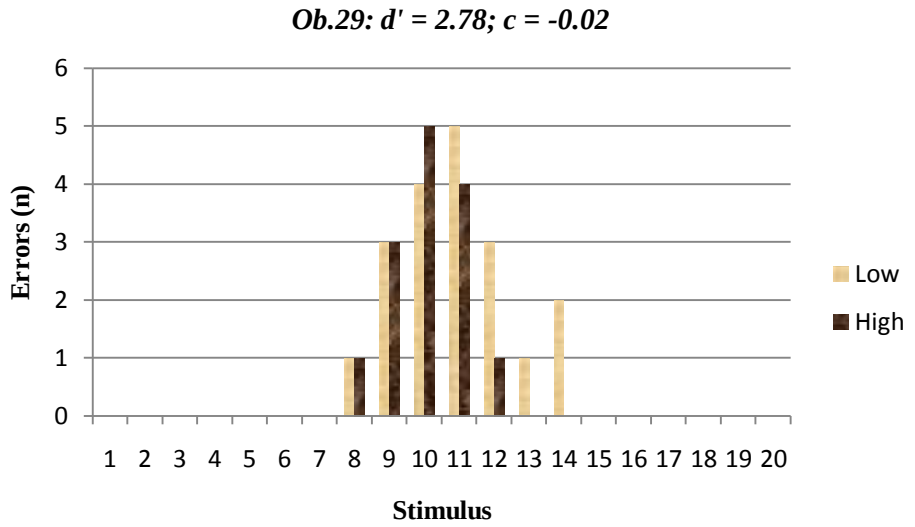


Figure H.5: *Error distributions for Observer 29, with relative statistics describing accuracy (d') and criterion location (c).*

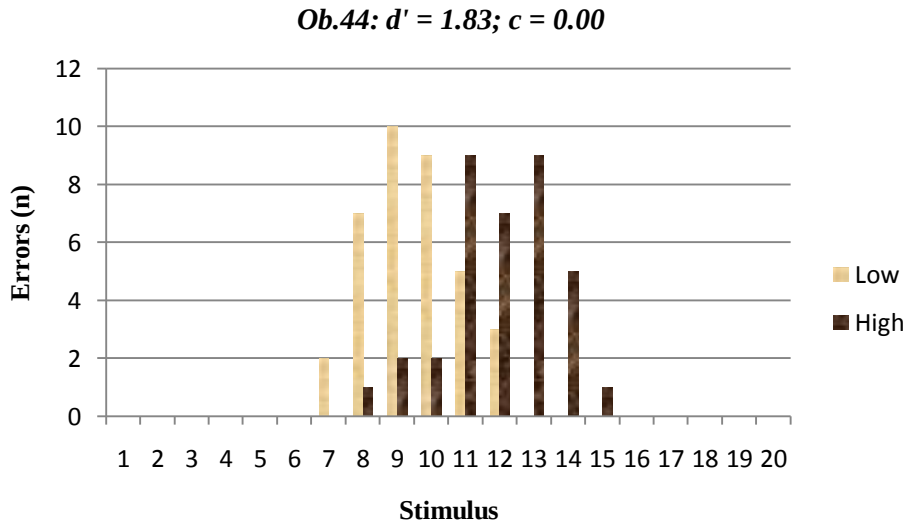


Figure H.6: *Error distributions for Observer 44, with relative statistics describing accuracy (d') and criterion location (c).*

Ideal d' = 3

Table H.7: *Distribution table for Observer 23.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
High	✓										3	12	29	38	39	30	18	9	3	1	
	×								1	3	6	6	1	1							
Low	✓		1	3	9	17	30	39	36	26	14	2									
	×					1			3	4	4	7	3	1							

Table H.8: *Distribution table for Observer 2.*

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
High	✓										4	13	21	32	34	29	18	8	2	1	
	×								1	3	4	5	9	7	5	1		1	1		
Low	✓		1	3	9	18	28	35	30	20	10	4									
	×						1	4	9	10	8	5	3	1							

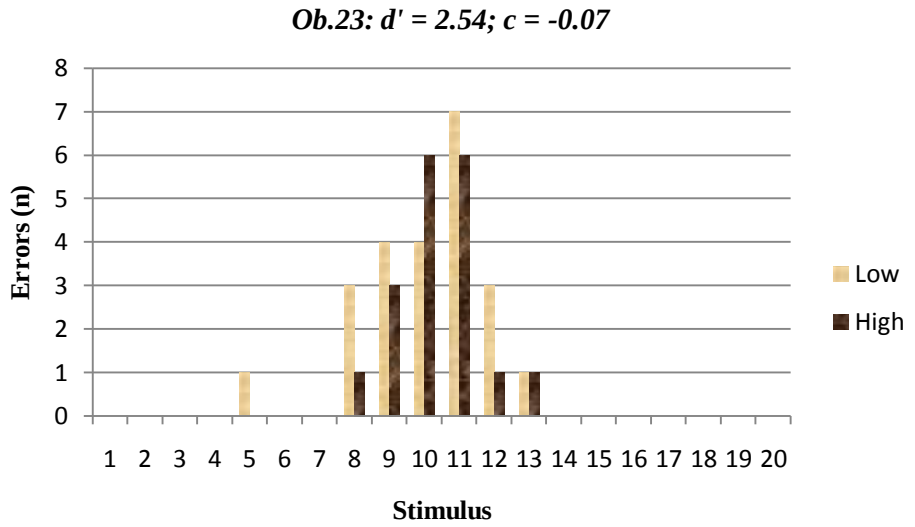


Figure H.7: *Error distributions for Observer 23, with relative statistics describing accuracy (d') and criterion location (c).*

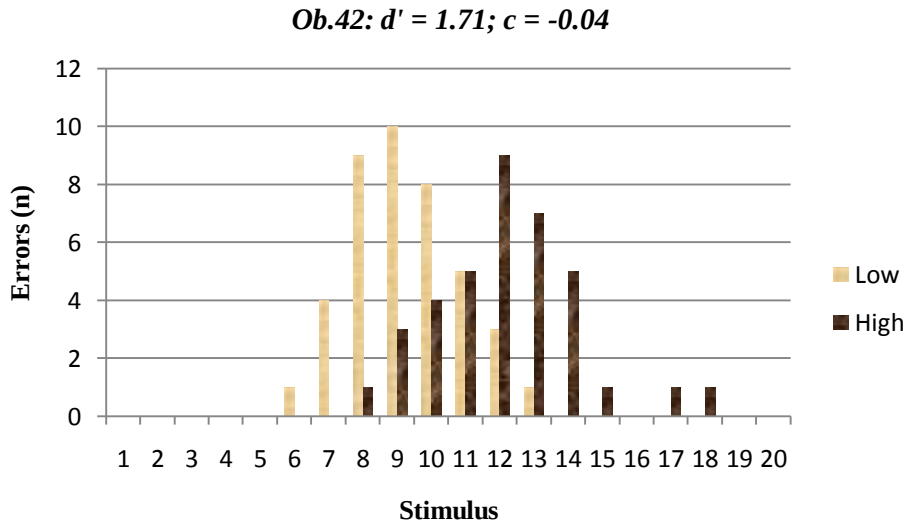


Figure H.8: *Error distributions for Observer 42, with relative statistics describing accuracy (d') and criterion location (c).*