# The Theory of Group Operating Characteristic Analysis in Discrimination Tasks

by

Vit Drga

A thesis submitted
in fulfilment of the requirements for the degree of
Doctor of Philosophy
in Psychology

Victoria University of Wellington
New Zealand

1999

# Abstract

Inconsistent decision making is a long-standing problem in psychophysics, where decisions based on the same stimulus often differ across replications of an experiment. Inconsistency is described statistically by the concept of unique noise, the effects of which are removed by averaging ratings across replications on a per-stimulus basis. A group operating characteristic (GOC) curve is a type of receiver operating characteristic (ROC) curve based on the mean rating per stimulus. GOC analysis is shown to improve task performance dramatically compared to ROC analysis, and can recover theoretical ROC curves from noisy data. This thesis presents a theory of GOC analysis showing why the procedure works. It also develops transform-average GOC analysis, transfer function analysis, and shows how to estimate unique-noise-free performance from a finite, unique-noise-affected data set.

Transform-averaging of ratings (for example, by using geometric or harmonic means) extends GOC analysis to include strictly monotonic increasing (s.m.i.) transformations of rating scale data. Although s.m.i. transforms do not alter ROC curves on any single replication, it is shown that they do alter GOC curves because of unique noise. Nevertheless, GOC analysis may be transform-invariant, apart from residual unique noise effects. Empirical evidence is given showing how GOC performance improves towards theoretical performance regardless of the particular rating scale that is involved.

A psychophysical transfer function is an s.m.i. mapping from a decision axis onto a rating scale. Transfer functions underlie theoretical interpretation of empirical ROC analysis, and it is shown how they can be estimated from empirical data. The theory of GOC analysis incorporates both transfer functions and transform-average GOC analysis under the same framework. The theory shows that GOC analysis will work under arbitrary (and possibly unknown) transfer functions, and under arbitrary ordinal scalings of a rating scale, but only when a family of unique-noise-affected evidence distributions are stochastically ordered on the decision axis. If stochastic ordering does not hold, unique-noise-free GOC performance changes according to the scaling of a rating scale. When that is the case, empirical results and subsequent theoretical interpretation become somewhat arbitrary. This finding about unique-noise-affected rating scales also extends to theoretical models that incorporate unique noise. Without stochastic ordering on a decision axis, the theoretical unique-noise-free ROC curve *can change* following an s.m.i. transform of the decision axis.

GOC performance improves as a function of replications added (FORA). Stable empirical FORAs result from all combinations analysis (ACA), where average performance is calculated over all possible GOC curves for a given number of replications. The logarithm of FORA increments is generally a linear function of the logarithm of the number of replications, typically with $r^2 > 0.995$. This pattern implies a three-parameter data model

that provided an excellent description of FORAs from six different experimental projects. These projects involved different aural discrimination tasks, experimental paradigms, decision methodologies, individual observers, levels of performance, stimulus parameters, and measures of sensitivity. Dozens of different FORAs followed the same mathematical form—only the three parameters of the data model changed.

Extrapolation of a FORA to an infinite number of replications makes it possible to estimate asymptotic unique-noise-free performance and its sample statistics based on a finite data set. Empirical FORA analysis showed that the observer with the best (unique-noise-affected) ROC performance was often not the observer with the best unique-noise-free performance. This shows that unique noise can generate deceptive results in psychophysics, but that its effects can be removed by using GOC analysis.

# Contents

# List of Figures

# List of Tables

# Notation

Table 1: Notation, acronyms, and abbreviations.

| Symbol | Meaning |
|---|---|
| $\overset{st}{<}$ | Stochastically strictly less than |
| $\overset{st}{\leqslant}$ | Stochastically less than or equal to |
| $\overset{st}{\not<}$ | Stochastically not less than |
| $\overset{st}{\not\leqslant}$ | Stochastically not less than or equal to |
| $\oplus$ | Unique and common noise mixing process |
| $\zeta$ | The Riemann zeta function |
| $\eta$ | Observer number, or division index |
| $\varepsilon$ | Stimulus index number |
| $\xi$ | Combination-size |
| $\Lambda$ | Quantising function from $R$ to $Q$ |
| $\Theta$ | Weighted sum of random variables |
| $\theta$ | Specific value of $\Theta$ |
| $A_1$ | Initial performance value in FORA regression |
| $A_\infty$ | Asymptotic performance value in FORA regression |
| $\mathcal{A}$ | Area under the ROC curve |
| $\mathcal{A}_{\text{SIFC}}$ | Area under the SIFC ROC curve |
| $\mathcal{A}_{\text{2IFC}}$ | Area under the 2IFC ROC curve |
| ACA | All combinations analysis |
| c.d.f. | Cumulative distribution function |
| $d'$ | Measure of sensitivity d–prime |
| $d_e$ | Alternative notation for $d_s$ |
| $d_s$ | Measure of sensitivity for a Gaussian unequal variance model |
| $d_z$ | Measure of sensitivity for a Gaussian unequal variance model |
| $\mathcal{D}_2$ | Scurfield's discriminability measure for two events (bits) |
| $\mathcal{D}_6$ | Scurfield's discriminability measure for six events (bits) |
| $\mathcal{D}_n$ | Scurfield's discriminability measure for $n$ events (bits) |
| DAC | Digital–to–Analog converter |
| ESO | Equivalent statistical observer |
| $f$ | Probability density function |
| $F$ | Cumulative distribution function |

(continued ... )

Table 1: Notation, acronyms, and abbreviations continued ...

| Symbol | Meaning |
|---|---|
| FAR | False alarm rate |
| FFT | Fast Fourier transform |
| FORA | Function of replications added |
| $G(t)$ | Difference between c.d.f.'s |
| $G^*(s)$ | Difference between c.d.f.'s |
| GOC | Group operating characteristic |
| $h$ | Transfer function from $Y$ to $R$ (or from $X$ to $R$) |
| $\mathcal{H}_2$ | Average uncertainty about the ordering of two events |
| HR | Hit rate |
| IFFT | Inverse fast Fourier transform |
| $\kappa$ | Scalar parameter in FORA regression |
| $k$ | Ratio of unique-to-common noise variances |
| $L(X)$ | Likelihood ratio decision axis |
| $L(x)$ | Specific value of likelihood ratio |
| $m$ | Number of observers, or<br>Number of replications |
| $\mu$ | Exponent parameter in FORA regression, or<br>Mean of a random variable |
| mean ROC | Mean receiver operating characteristic |
| $N$ | Noise–alone |
| $n_r$ | Number of resamplings of ACA sets |
| $P(\mathrm{C})_{\mathrm{SIFC}}$ | Proportion correct in the SIFC task |
| $P(\mathrm{C})_{\mathrm{2IFC}}$ | Proportion correct in the 2IFC task |
| $q$ | Size of discrete rating scale |
| $Q$ | Discrete rating scale |
| $Q_j$ | Random variable on $Q$ for the $j^{th}$ stimulus |
| $q_{ji}$ | Sample value of $Q_j$ on the $i^{th}$ replication |
| $\mathbb{R}$ | Real number line |
| $R$ | Continuous rating scale |
| $R_j$ | Random variable on $R$ for the $j^{th}$ stimulus |
| $r_{ji}$ | Sample value of $R_j$ on the $i^{th}$ replication |
| $r$ | Rating value, or<br>Pearson's product-moment correlation coefficient |
| $r^2$ | The square of the correlation coefficient |
| ROC | Receiver operating characteristic |
| $\sigma$ | Standard deviation |
| $\sigma^2$ | Variance |
| $\sigma_{\mathrm{N}}$ | Standard deviation for the $N$ event |
| $\sigma^2_{\mathrm{N}}$ | Variance for the $N$ event |
| $\sigma_{\mathrm{SN}}$ | Standard deviation for the $SN$ event |
| $\sigma^2_{\mathrm{SN}}$ | Variance for the $SN$ event |

(continued ... )

Table 1: Notation, acronyms, and abbreviations continued ...

| Symbol | Meaning |
|---|---|
| $\sigma_{\mathrm{c}}^2$ | Variance of common noise |
| $\sigma_{\mathrm{u}}^2$ | Variance of unique noise |
| SHIN | Signal hidden in noise |
| SIFC | Single–interval forced–choice task |
| s.m.i. | Strictly monotonic increasing |
| *SN* | Signal–plus–noise |
| SNR | Signal–to–noise ratio |
| sorting-key | The numerical value upon which a stimulus set may be ordered |
| SPL | Sound pressure level |
| 2IFC | Two–interval forced–choice task |
| TSD | Theory of signal detectability |
| $X$ | Decision axis, or |
| | Common noise decision axis |
| $X_{\mathrm{N}}$ | Random variable on $X$ conditional on the $N$ event |
| $X_{\mathrm{SN}}$ | Random variable on $X$ conditional on the $SN$ event |
| $x_c$ | Observer's criterion |
| $x_j$ | Sample value of $X$ for the $j^{th}$ stimulus |
| $Y$ | Decision axis incorporating unique noise |
| $Y_j$ | Random variable on $Y$ for the $j^{th}$ stimulus |
| $y_{ji}$ | Sample value of $Y_j$ on the $i^{th}$ replication |

## Acknowledgements

*"Wonder en is gheen wonder."*
(Nothing is the miracle it appears to be.)

Simon Stevin

I would like to express my gratitude to the people and institutions that made this thesis possible.

The thesis would not have seen daylight without much encouragement and support from John Whitmore, Judi Lapsley Miller, Linton Miller, Brian Scurfield, Alan Taylor, and Sue Galvin. The ideas and topics in this thesis, and other projects, developed through many fruitful and enjoyable discussions with them. Thanks to these people, and also John Podd , for allowing me to put their experimental data sets through the wringer. The findings in here, theoretical as well as empirical, ultimately stem from their data.

I am grateful to the School of Psychology at Victoria University of Wellington, for financial, material, administrative, and technical assistance, and particularly to Maree Hunt, Frank Walkey, John McDowall, Ngaire Lavery, John Bowden, Doug Drysdale, Keith Riach, and Linton Miller.

This research was supported by a U.G.C. Postgraduate Scholarship, J.L. Stewart Scholarship, the New Zealand Student Allowance scheme, and Victoria University of Wellington Grant No. IGC–950.

Special thanks go to John Whitmore, Judi Lapsley Miller, and Linton Miller, for all of their friendship, effort and support over the years.

To John, for having me as a research assistant, and teaching assistant (where I learnt as much as the students), for generous access to his personal computing equipment and scientific literature, for administrative help, and for volunteering many hours (!) in the sound chamber. John was once had on for being a natural philosopher, but it's a pity we don't have more like him.

To Judi, for assistance-*plus* with LaTeX, graphics, and e-things, for acting as on-line help, and for letting me tag along on her experimentation. All of the theory of GOC graphs, and other niggly ones, were hand-crafted by Judi. Thanks also to Judi for making her thesis available. Chapter 6 is neat.

To Linton, for his tremendous on-line, off-line and inline support and help, for being a LaTeX Wizard, and a guru on the innards of systems that have a life of their own.

John, Judi, and Linton helped enormously by reading the many successive approximations to this document, and generously volunteered their time, and red and green and purple ink. Their suggestions greatly improved the manuscript. Any errors, omissions, and chaff that remain are wholly my own. Proof reading is partly a process of error detection, and together they engaged in a practical application of GOC analysis, on GOC analysis.

On a personal note, I made an informal resolution, a while ago, to let my hair grow until the thesis was done—and grow it did. In the very last stages of the write-up, the comb I'd been using for years finally gave up the ghost and snapped in two.

I guess it's time for a haircut ...

# Preface

A psychophysical discrimination task is where an *observer*[1] has to make a decision about objectively defined events in the world. How well an observer performs in a discrimination task depends on the specifics of the task, the types of stimuli involved, the conditions under which they are presented, the motivation of the observer, and physiological and physical limitations of the observer. All of these factors, and more, contribute to *observer inconsistency* in decision making, which leads to extra error in the task. Observer inconsistency refers to the fact that observers make different decisions over repeated observations of the *same* stimulus. This is not a trivial influence on performance, because experimental evidence suggests there is as much decision variability due to inconsistency as there is due to uncertainty regarding the stimuli presented. Human beings are not perfect discriminators. The consequences of wrong decisions may be dire for some tasks, yet trivial for others. Investigating observers' performance characteristics and limitations is a key topic of modern psychophysics.

The effects of observer inconsistency can be reduced substantially by *group operating characteristic (GOC) analysis*. GOC analysis addresses the question of whether a group could perform better than an individual, that is, are many ears (or eyes) better than one? If the group is better than the individual observer, the potential benefits are great in situations where detection or discrimination is critical (such as cancer detection, search and rescue, jury verdicts, sonar or radar detection, or fault detection).[2] GOC methodology may also be used to improve and assess individual performance in discrimination tasks, by having the same observer make decisions about identical stimuli presented on different occasions.

Statistically, performance may be partitioned into *unique noise* and *common noise*, where unique noise differs across replications and common noise remains the same. GOC analysis removes unique noise effects from experimental data. Performance tends toward the level associated with common noise, as unique noise is removed. Reduction of observer inconsistency and the estimation of asymptotic, unique-noise-free performance are the main topics of this thesis.

All of the experimental data sets that are analysed in this thesis involve aural discrimination tasks in which an observer is presented with a sound and then makes a decision about what type of sound was presented. Some of these experiments used frequency discrimination tasks designed so that theoretical results were known *a priori*. Results from

[1] The term *observer* is used here in preference to *subject*, following psychophysical convention (Green & Swets, 1974, p. 11). The former encompasses theoretical and simulated decision-making devices, as well as human and non-human experimental subjects.

[2] Psychophysical methodologies and analyses find application in a wide range of topic areas (Swets, 1964; Green & Swets, 1974; Swets, 1986). Hutchinson (1981) describes a host of unusual applications, and is highly recommended reading for anyone with an interest in psychophysics.

such experiments unambiguously demonstrate the effects of observer inconsistency and show how these effects can be removed. Other experiments investigated aural amplitude discrimination. These were more substantive experiments for which theoretical performance was not known. Although various theories of performance in such tasks have been proposed, there is no model that is generally agreed upon. Theories of discrimination tasks have mostly been evaluated using psychophysical methods that incorporate the effects of inconsistent decision making. The methods and analyses that are developed here can be used to evaluate theories and performance, both with and without the effects of observer inconsistency.

The thesis is divided into two parts. Part I is about GOC analysis, what can be accomplished by using it, and how it works, in theory. Part II is about what is called FORA regression, which allows the estimation of asymptotic unique-noise-free performance from a finite, unique-noise-affected data set.

Chapter 1 describes aspects of the Theory of Signal Detectability (TSD) which form the basis for theoretical and experimental developments in later chapters. The main departures from conventional TSD in this thesis (which are not original) are: (1) an ideal observer is not necessarily an optimal observer, (2) receiver operating characteristic analysis and rating scales are just as applicable to two-interval forced-choice tasks as they are to single-interval forced-choice (or yes-no) tasks, and (3) there is no essential difference between binary-decision and rating scale tasks.

Chapter 2 describes observer inconsistency, how it can be modelled, and computational analyses for dealing with its effects. In particular, *mean receiver operating characteristic (mean ROC) analysis* and GOC analysis are each described in turn, and are applied to the same unique-noise-affected data set. GOC analysis is shown to remove unique noise effects, whereas mean ROC analysis does not.

Chapter 3 extends GOC analysis to incorporate general *transform-average* mean ratings, and shows how this extension formally relates to arbitrary ordinal rescalings of rating scale data. Transform-average GOC analysis is applied to an experimental data set, and the results are shown to be generally similar across a variety of transforms.

Chapter 4 describes psychophysical transfer functions, which relate a decision axis to a rating scale. Given an assumed decision axis, it is possible to estimate a transfer function from experimental data.

Both transfer functions and ordinal transforms are incorporated within a general model of a unique-noise-affected observer, presented in Chapter 5. A theory of GOC analysis follows from the model, which describes the statistical properties that must hold in order for GOC analysis to work in general. In the theory, the removal of unique noise from rating data parallels the removal of unique noise on a decision axis, and may do so under any transfer function and any arbitrary ordinal scaling of a rating scale.

In Part II, Chapter 6 introduces the function of replications added (FORA). Performance generally improves as more replications are combined in GOC analysis. Very smoothly increasing FORAs result from all combinations analysis (ACA), which involves calculating a GOC curve for each possible subset of a data set and averaging performance for all subsets of the same size. FORA increments and numbers of replications are generally related by straight lines in double-logarithmic coordinates. A summed power-law FORA regression function that follows from this provides an excellent description of data. The FORA function can be extrapolated to an infinite number of replications to provide an estimate of asymptotic unique noise-free performance fwrom a finite data set. FORAs

are found to be stable for measures based on an entire ROC curve, such as area under the curve, but are not as stable for performance measures based on only a single ROC point, such as percent correct.

Different sets of replications provide different estimates of an asymptote. Using a very large data set, Chapter 7 shows how it is possible to estimate sample statistics of asymptotes and to obtain an indication of how many, or few, replications are needed to obtain a stable estimate of asymptotic performance. Solutions are given to problems that arise when applying ACA to very large data sets.

Chapter 8 provides FORA results from a variety of experiments in which FORA regression was found to be extremely robust. Asymptotic estimation was possible for different experiments, tasks, rating scales, observers, types of stimuli, performance levels, and measures of sensitivity.

The experimental system that was used in the experiment in Section 8.3 was developed and tested by Judi Lapsley Miller and Linton Miller, for Lapsley Miller's (1999) doctoral thesis (Section 8.4). Lapsley Miller (1999) contains a full description of the experimental system. The experiment in Section 8.3 occurred after the work in Section 8.4, although they are presented in reverse chronological order to aid the development of Chapter 8. The core FFT code used to generate the stimulus sets for the experiments in Sections 8.3 and 8.4 was coded in assembler by Linton Miller, who also programmed all combinations analysis of $\mathcal{D}_6$ for Lapsley Miller's project.

**Stylistic notes.** Several stylistic points should be noted in this thesis. Long lists of references sometimes appear as bracketed footnotes, especially when such lists would have appeared in the middle of a paragraph. Numerical ranges are sometimes expressed using brackets, where square brackets denote inclusion, and round brackets denote exclusion, for example, $x \in [1,3)$ means that $1 \leqslant x < 3$. The term "ordering" is used as a noun more frequently than as an active verb. For example, "stochastic ordering" is not an active process, but rather describes an ordered pattern. In Part II, although *FORA* stands for "Function Of Replications Added", all graphs of FORAs have abscissas labelled *Replications Combined*, as a reminder that FORAs resulting from ACA are derived by averaging over combinations of replications, rather than from the addition of a single replication to an existing subset.

There is one historical note that does not quite fit in elsewhere. The linear log-log plot, which forms the intuitive basis for FORA regression, was discovered by accident during exploratory data analysis of a data set. A graph of the logarithm of FORA increment plotted against replications added (like Figure 6.2(c)) was displayed on a computer screen, when John Whitmore suggested to try applying the logarithm of the abscissa as well as the ordinate. A straight line resulted immediately. John got one of the axes, and I got the other.

V.D.

# Part I

# Group operating characteristic analysis

*Part I is concerned with the effects of observer inconsistency in discrimination tasks, and how the effects can be removed within the context of the Theory of Signal Detectability (TSD). Chapter 1 gives an overview of TSD, with details of methodologies and measures of performance that are used in later chapters. Chapter 2 describes models of observer inconsistency, and also mean receiver operating characteristic analysis and group operating characteristic (GOC) analysis as means for removing variability due to inconsistency. Chapter 3 describes transform-average GOC analysis, which is a generalisation of GOC analysis that encompasses generalised mean ratings and arbitrary ordinal scaling of a rating scale. Chapter 4 introduces the transfer function, which relates values on a decision axis to values on a rating scale and shows how it can be estimated from data. Chapter 5 provides a theory of GOC analysis that incorporates the developments of previous chapters within a single framework. Stochastic ordering is shown to be the key statistical property needed in order for GOC analysis to remove the effects of inconsistency from experimental data. If stochastic ordering holds, then GOC analysis works for arbitrary transfer functions and arbitrary scalings of a rating scale.*

# Chapter 1

# Elements of the Theory of Signal Detectability

**The fundamental detection problem.** The simplest, non-trivial discrimination task is the *fundamental detection problem*, where one of two possible events occurs in an observer's enviroment, and the task of the observer is to state which event occurred (Swets, Tanner, & Birdsall, 1961; Egan, 1975). In aural detection experiments, the events are labelled *SN* and *N*, for *Signal-plus-Noise* and *Noise-alone* respectively. The *N* event would typically be the presentation of a background masking noise during a trial, and the *SN* event would be the presentation of an extra signal waveform added to the masking noise. A trial formally consists of an observation interval, a decision interval, and an optional payoff interval, in that order. One of the events occurs during the observation interval, decisions are made during the decision interval, and possible consequences of the decision (if any) occur during the payoff interval. The two events are mutually exclusive, and are assumed to be independent over a series of trials.

| | | Decision | |
|---|---|---|---|
| | | "yes" | "no" |
| Event | **SN** | Hit | Miss |
| | **N** | False Alarm | Correct Rejection |

TABLE 1.1: Event-decision matrix for the fundamental detection problem.

The observer's task in the fundamental detection problem may be posed as the question "Did the *SN* event occur?", and the possible decisions are *"yes"* and *"no"*. There are four possible event-decision conjunctions, as seen in Table 1.1. These are labelled a *hit*, *miss*, *false alarm* and *correct rejection*. Over a series of trials, the proportion of

times these conjunctions occur define the *hit rate*, *false alarm rate*, *miss rate*, and *correct rejection rate*.[1] Generally, observers do not perform perfectly in discrimination tasks, and consequently hit rates and correct rejection rates do not typically equal one. The higher the hit rate and correct rejection rate, the better an observer performed in the task. These proportions are used as the basis of performance measures in the task (some of which are described in Section 1.2).

## 1.1 The Theory of Signal Detectability

**The Theory of Signal Detectability.** The Theory of Signal Detectability (TSD) is an amalgamation of two theories, the theory of ideal observers and statistical decision theory. The theory of ideal observers (Green & Swets, 1974; Tanner & Sorkin, 1970) is a theoretical framework which describes how stimuli may be processed by an observer. In a psychophysical context, statistical decision theory (Green & Swets, 1974; Egan, 1975) describes how the results of stimulus processing, which is typically a random variate, may form the basis of decision making in a discrimination task. Together, the theory of ideal observers and statistical decision theory provide a broad theoretical framework for describing observers in discrimination tasks. Each of these components of TSD is described in turn.[2]

**Ideal Observers.** In psychophysical discrimination tasks, it is assumed that an observer's decisions are made based on a particular characteristic (or characteristics) of the stimuli presented. There are many possible characteristics that may suffice, for example in aural tasks the characteristic may be the amplitude, energy, frequency, phase, bandwidth or duration of the sounds presented. An *ideal observer* is viewed here as a model, theory, or simulation of an observer. The model is able to process or analyse stimuli, and provide some quantity which differs across events, or has different statistical properties for each event. The quantity may then be used as the basis for decisions in a task.

In a strict sense, an ideal observer is an input-output system that does not make decisions. Its input is a stimulus (a sound pressure wave for example, or a representation of such a wave) and its output is a measurable quantity or numerical value. In a broader

[1] These terms were borrowed in the 1950s from the field of engineering. In medical diagnostics, which also uses discrimination tasks, the terms *sensitivity* or *true positive rate* are used instead of *hit rate*, the term *specificity* is used of *correct rejection rate*, and *false positive rate* is used instead of *false alarm rate* (Hanley, 1988; Hsieh & Turnbull, 1996). A potential confusion exists in the use of the term *sensitivity*, which is synonymous with *hit rate* in a medical diagnostic task, and in a classical psychophysical context, but is synonymous with overall performance in a psychophysical discrimination task. Throughout this thesis, *sensitivity* is used in the modern psychophysical sense, and is never used to refer to *hit rate* in any context.

[2] The terminology in this area of psychophysics differs across authors, and sometimes the distinctions among TSD, statistical decision theory, and the theory of ideal observers are blurred. Egan (1975, p. 3), for example, equated the theory of signal detectability with the theory of ideal observers.

sense which is often used (including here), an ideal observer makes decisions if a decision rule (part of statistical decision theory) is incorporated with the observer.

Many psychophysical experiments involve sampling stimuli from random processes, a case in point being the detection of a signal in the presence of background noise. Given a random process input to an ideal observer, the output values are distributed according to some random variable, $X$, which is called the *decision axis*. $X$ has also been called the *evidence*, or *evidence variable*, or *decision variable* (McNicol, 1972; Egan, 1975).

The form of $X$ depends on the stimuli and type of ideal observer. Gaussian random variables are widely used in TSD, largely because of the widespread applicability of the Gaussian form to empirical data (Hanley, 1988), and also because of theoretical derivations that lead to Gaussian statistics (Elliot, 1964; Green & Swets, 1974). TSD is seen by some as a theory *restricted* to Gaussian random variables only (Simpson and Fitter, 1973; Eijkman, 1992, cited in Scurfield, 1995), but this is false (Scurfield, 1995; Lapsley Miller et al., 1998). Ideal observers based on other types of distributions have been derived in psychophysics (for example, chi, chi-squared and $F$ distributions; Jeffress, 1964; Green & Swets, 1974; Green & McGill, 1970). It has also been stated that TSD assumes that the decision axis $X$ is continuous (Gilkey, 1981, p. 3). This may have been true in the original formulation of TSD (Peterson, Birdsall and Fox, 1954, cited in Gilkey, 1981), but is not true in general. For example, auditory signal detection tasks have been modelled using ideal observers that are based on poisson counting (McGill, 1967; Schacknow & Raab, 1976; McGill & Teich, 1991). In the broadest form of TSD, there are no constraints on the type of evidence distributions that may be generated by an ideal observer.

**Events, evidence and stimuli.** Three key concepts in TSD are *events*, *evidence* and *stimuli*. In TSD, the task is to discriminate between environmental *events*, which is done on the basis of *evidence* derived from *stimuli*. The most difficult concept to define is that of the *stimulus*. (Gibson, 1960, describes several very different uses of the term.). As it is used here, a *stimulus* refers to a pattern which occurs in a given portion of an observer's immediate physical environment, for example a particular acoustical waveform. In an experimental context, *stimulus* may also refer to the pattern only, rather than the medium or location in which it occurs (e.g. waveforms stored on computer may be referred to as *stimuli*). An *event* is a state of nature, which may or may not be part of an observer's immediate environment. In some circumstances, an *event* could be synonymous with a particular set of stimuli, or a class or type of stimulus (e.g. a set of waveforms with particular spectral characteristics). *Evidence* is taken to be the output of an ideal observer, although the term has been used in a broader sense, by referring to physical evidence (i.e. , a stimulus) in a discrimination task (Egan & Clarke, 1966).

**Statistical decision theory.** In TSD, the type of stimulus characteristic used to derive $X$ is the same for each event, *SN* and *N*, but the properties or parameters of *SN* stimuli and *N* stimuli are different (e.g. if the characteristic was acoustical power, the average signal-plus-noise power is typically greater than the average noise-alone power). The form of $X$, including parameters of the distribution of $X$, typically depend on the event that occurs. There are *two* probability distributions that need to be taken into account. The $x$-values of the *SN* stimuli are distributed, say, as $X_{\mathrm{SN}}$, and the $x$-values of the *N* stimuli are distributed as $X_{\mathrm{N}}$. By itself, the random variable $X$ refers to the distribution of $x$-values over the entire stimulus set, whereas $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ are its event-conditional forms.

Knowing the evidence value on a trial does not by itself result in a decision. What is also needed is a *decision rule*, which is an algorithm for producing a decision, given an evidence value, $x$. The application of decision rules to evidence values is a topic of statistical decision theory (Green & Swets, 1974; Egan, 1975). There are many possible decision rules, some deterministic, some probabilistic, some optimal, and others sub-optimal. A simple, yet effective, and widely used rule is a *criterion-based decision rule*, in which the evidence, $x$, is compared to a criterion value, $x_c$. In answer to the question "Did the *SN* event occur?", the decision rule would conventionally be expressed as

> If $x \geqslant x_c$ then say "yes",
> if $x < x_c$ then say "no".

Given the distributions of $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ for an ideal observer, the value of the criterion determines the hit rate, false alarm rate, miss rate and correct rejection rate.

**Ideal observers and optimal observers.** It is well known that performance can be optimised, with respect to a variety of decision goals, by a criterion-based decision rule applied to the likelihood ratio, $L(X)$, which is derived from event-conditional probability density functions (or probability mass functions) on $X$ (Green, 1960b; Green & Swets, 1974; Egan, 1975). If $X$ is a strictly monotonic increasing transform of $L(X)$, then any criterion on $X$ has an equivalent criterion on $L(X)$, resulting in the same decision for specific related values $x$ and $L(x)$ (Egan, 1975, Chapter 2). For a given decision axis $X$, and with respect to a given decision goal, optimal performance is achieved by setting an appropriate criterion on $L(X)$.

According to Green and Swets, "the adjective 'ideal' [in the term *ideal observer*] refers to the best possible performance in detecting signals under specified conditions," (Green & Swets, 1974, p. 151). For them, an ideal observer is synonymous with an optimal observer. That equivalence is *not* used in this thesis because there is no objective definition of an optimal observer which is independent of the observer itself. The "best possible performance" is dependent on the detector or ideal observer, and also on the model that is used to describe stimuli (Green & Swets, 1974, Section 6.6). In a very general sense,

the only optimal observer is one that always performs a task perfectly (i.e. makes no error in the task). Under that definition, most observers, including ideal observers, are sub-optimal.

Finding or inventing an optimal observer is different from optimising performance *given* an ideal observer or decision axis. Statistical decision theory shows that it is possible to achieve optimal performance for any given decision axis, $X_1$, by converting to likelihood ratio, $L(X_1)$, but there is no guarantee that $L(X_1)$ *must* result in the best possible performance in the task.[3] For the same discrimination task, some other detector could result in a different decision axis, $X_2$, whose likelihood ratio, $L(X_2)$, provides better performance than $L(X_1)$. Which decision axis provides better performance depends on how $X_1$ and $X_2$ are derived from the stimuli, or most importantly in a theoretical context, how they are derived from a model, theory, or description of the stimuli (Green & Swets, 1974, Section 6.6).

An ideal observer represents a stimulus transduction and measurement device, or a theory of such a device. A demonstration that such a device or theory is in some sense optimal is not required, and the decision axis $X$ from the ideal observer is not necessarily strictly monotonic increasing with $L(X)$. The benefit of this is twofold: (1) real observers often do not perform optimally, and non-optimal ideal observers may provide a better model than optimal observers, and (2) it is sometimes easier to construct an ideal observer than it is to demonstrate that it is optimal.

**Bias.** There are two ways of being correct and two ways of being incorrect in the fundamental detection problem. Hits and correct rejections are correct decisions, while misses and false alarms are incorrect decisions. There is redundancy in Table 1.1, because the hit and miss rates sum to one, and the false alarm and correct rejection rates sum to one. If one value from each pair of rates is known, then performance in the task is completely specified. Conventionally, the hit rate ($HR$) and false alarm rate ($FAR$) are presented (Green & Swets, 1974; Egan, 1975), although other pairings have also been used (Bamber, 1975; Scurfield, 1995).

Hit and false alarm rates are not independent, since they both depend on an observer's willingness to decide "yes". This willingness is called the *decision bias* (or just *bias*). The bias is determined, in TSD, by the criterion, $x_c$. If $x_c$ is high (relative to the evidence distributions), then an observer is biased towards deciding "no" and *both* the hit and false alarm rates are low. If $x_c$ is low, then an observer is biased towards deciding "yes", and *both* the hit and false alarm rates are high. Factors that can affect bias include an observer's motivation, instructions in the task, consequences of possible decisions (payoffs), and the prior probabilities of each event (Green & Swets, 1974). Bias influences an observer's apparent ability by affecting task performance, without changing an observer's inherent

[3] Lapsley Miller, 1999, personal communication.

ability to discriminate between events, which is determined by the nature of the stimuli and the observer. A key contribution of TSD to psychology is to formalise the distinction between sensitivity, or discriminability, and bias.

**Receiver Operating Characteristic analysis.** Given a criterion, $x_c$, the hit rate is

$$HR\,(x_c) \quad = \quad P(X \geqslant x_c|SN)$$

and the false alarm rate is

$$FAR\,(x_c) \quad = \quad P(X \geqslant x_c|N).$$

These are conditional probabilities which systematically increase as $x_c$ decreases. The plot of *HR* versus *FAR* taken over *all* values of $x_c$ is called the *receiver operating characteristic curve*, or *ROC curve*. The ROC curve is plotted in the *ROC space*, in which the false alarm rate defines the abscissa and the hit rate defines the ordinate. Chance performance implies that $HR = FAR$ for all values of the criterion, and so the positive diagonal, where $HR = FAR$, is called the *chance line*. Another feature of the ROC space is the *negative diagonal*, where $HR = 1 - FAR$, which is used in Section 1.2 to calculate measures of performance. The use of ROC curves is called *ROC analysis*.

**Ideal observers and ROC analysis.** The use of ideal observers puts the results of real observers into a theoretical context. If an ideal observer is used to explain the results of a real observer, it is *assumed* that the real observer uses a decision axis and decision rule analogous that of the ideal observer. It is not possible to work backwards from an ROC curve to derive the decision axis underlying it, although it has been claimed otherwise McNicol (1972). Egan (1975, Appendix B) shows that the form of the underlying distributions cannot be inferred from a given ROC curve. Any decision axis which is a *strictly monotonic increasing* transform of an ideal observer's decision axis will result in the same theoretical ROC curve. This shows that there are an unlimited number of possible decision axes and ideal observers that result in the same ROC curve. Even if a theoretical ROC curve matches an empirical ROC curve exactly, the strict monotonicity result implies that no single ideal observer can be claimed to provide *the* explanation of the real observer's decision making. Without other knowledge, it can only be claimed that the ideal observer's decisions are consistent with those of the real observer.

### 1.1.1 The single-interval forced-choice (SIFC) task

The simplest discrimination task is the *single-interval forced-choice* (SIFC) task. An SIFC trial consists of a single observation interval, decision interval and payoff interval. The simplest example of an SIFC task is the fundamental detection problem. The SIFC task extends beyond two-event, binary-decision methodology to incorporate multiple-point rating scales, and multiple-event tasks. Formal extensions to multiple-event, single-interval tasks are described in Scurfield (1995, 1996, 1998), and analyses by Lapsley Miller (1999), based on such extensions, are summarised in Chapter 8. Apart from Lapsley Miller's (1999) experiments, all of the other SIFC experiments presented in later chapters involve two-event tasks.

### 1.1.2 The two-interval forced-choice (2IFC) task

Another widely-used task is the *two-interval forced-choice* (2IFC) task. A 2IFC trial consists of two observation intervals, followed by a decision interval and an optional payoff interval. During each trial, one of the *SN* or *N* events occurs in the first observation interval, while the other event occurs in the second observation interval. The order of events is not known to the observer, and the task is to decide which order of events occurred.[4] Care must be taken in using the term *"event"* when dealing with 2IFC tasks, because it may refer to either the *SN* or *N* event within one of the observation intervals, or it could refer to either event-ordering across observation intervals. The two possible event-orderings define two *2IFC events*. These are labelled $\langle \mathrm{SN}, \mathrm{N} \rangle$ when *SN* occurs in the first interval, and $\langle \mathrm{N}, \mathrm{SN} \rangle$ when *SN* occurs in the second interval. Although both the *SN* and *N* events occur during each 2IFC trial, only one of the two 2IFC events (event-orderings) occurs. Which of the 2IFC events occurs in a trial is not known, and hence there is a parallel between the 2IFC task and the fundamental detection problem.

The Theory of Signal Detectability is much more complicated for 2IFC tasks than for SIFC tasks, due to the need to compare stimuli from two observation intervals. The requirement of a 2IFC decision rule introduces an extra layer of abstraction, and extra assumptions, in psychophysical theories of a 2IFC tasks compared to theories of SIFC tasks. Roughly speaking, SIFC tasks test stimulus processing ability, while 2IFC tasks test stimulus processing *and* stimulus comparison ability. The confounding of stimulus comparison with stimulus processing in 2IFC tasks is not widely recognised.

There are at least two approaches to modelling an ideal observer in the 2IFC task: (1) by manipulation and comparison of the two stimuli prior to deriving an evidence value, or (2) by the separate calculation of an evidence value for each stimulus, and a

[4] The 2IFC task could be formulated in several different, but equivalent ways. It is usually formulated as the observer having to state in which interval the *SN* event occurred (Green & Swets, 1974). The formulation in terms of orderings follows Scurfield's extension of TSD to multiple-event, multiple-interval tasks (Scurfield, 1995, 1996, 1998), in which the order of events is important.

comparison of the two evidence values. These two ways may produce the same result, or different results, depending on the specifics of the model. The second approach is much more common, possibly because the derivations involved may be more straightforward.

An example of the first approach would be to sample the stimulus waveforms in each observation interval, subtract the two sampled waveforms, and then derive a 2IFC decision axis from the difference waveforms (Tanner & Birdsall, 1958). The second approach requires an ideal observer to separately process the stimulus from each observation interval, resulting in one SIFC evidence value per 2IFC observation interval. Statistical decision theory applies in the form of a 2IFC decision rule that combines the two SIFC evidence values across intervals in order to arrive at a decision. Possible decision rules include those based on differences of SIFC evidence values (Tanner & Birdsall, 1958; Green, 1960a; Robinson & Watson, 1970; McNicol, 1972; Simpson & Fitter, 1973; Green & Swets, 1974; Egan, 1975; Siegel, 1979; Lapsley Miller et al., 1998), and those based on ratios of SIFC evidence values (Green & McGill, 1970) or likelihood ratios (Marill, 1956; Green & Swets, 1974). Whatever type of decision rule is used, the result is a 2IFC decision axis, which is analogous to the SIFC decision axis, but which has evidence distributions that are different from those on the SIFC axis for any specific experiment.

However a 2IFC decision axis is derived, a criterion-based decision rule can be applied to it in the same way as for an SIFC task, and systematic manipulation of the 2IFC criterion yields a 2IFC ROC curve. (It is customary in 2IFC ROC analysis to associate the 2IFC hit rate with the $\langle \mathrm{SN}, \mathrm{N} \rangle$ event, and the 2IFC false alarm rate with the $\langle \mathrm{N}, \mathrm{SN} \rangle$ event.) When the prior probabilities of each 2IFC event are equal, which is typically the case in 2IFC studies, ROC performance of an unbiased 2IFC observer falls on the negative diagonal of the ROC space.

**Prior probabilities.** The prior probabilities of each event play a role in optimal decision making for any given decision axis, since observers should adjust their criteria for deciding *yes* or *no*, according to which event is more likely to occur (Green & Swets, 1974; Egan, 1975). The priors are denoted as $P(SN)$ and $P(N)$ in the SIFC task and $P(\langle \mathrm{SN}, \mathrm{N} \rangle)$ and $P(\langle \mathrm{N}, \mathrm{SN} \rangle)$ in the 2IFC task. One way of deriving experimental ROC curves is by manipulation of the priors, and while this is sometimes done in SIFC experiments (e.g. Emmerich, 1968b; Nachmias, 1968; Schulman & Greenberg, 1970), it is *rarely* done in 2IFC experiments, although a notable exception is found in a study by Friedman and Carterette (1964). Usually, both priors in 2IFC experiments are set to 0.5 to aid the achievement of unbiased performance.

### The 2IFC task in the context of TSD

The 2IFC task is similar to the fundamental detection problem because there are two mutually exclusive events and two mutually exclusive decisions. Consequently, results from 2IFC experiments can be analysed in the same way as SIFC experiments, although the interpretation may differ due to the differing experimental designs.

ROC analysis of 2IFC tasks has appeared only rarely in the psychoacoustic literature. The majority of studies employing 2IFC ROC analysis are Friedman and Carterette (1964), Schulman and Mitchell (1966), Markowitz and Swets (1967), Leshowitz (1969), Watson, Kellogg, Kawanishi, and Lucas (1973), and Lapsley Miller et al. (1998). Theoretical descriptions of 2IFC ROC curves can be found in these studies, and also in McNicol (1972), Simpson and Fitter (1973), Green and Swets (1974), and Egan (1975).

As Luce (1997) points out, there is a common *misconception* that the 2IFC task is bias-free, whereas the influence of bias in the SIFC task is well known. For example, Mackworth (1970) states that "no change in criterion can occur [in the 2IFC task]" (Mackworth, 1970, p. 35), which is false. In both the SIFC and 2IFC tasks, TSD formalises the influence of bias in terms of decision criteria, and ROC analysis provides a means of taking such bias into account. Much of the misconception about the 2IFC task stems from the apparent symmetry of the 2IFC procedure, and from the theoretical result that the proportion of correct decisions in the 2IFC task, $P(\mathrm{C})_{\mathrm{2IFC}}$, is equal to the area under the SIFC ROC curve, $\mathcal{A}_{\mathrm{SIFC}}$ (Green & Swets, 1974).

The symmetry of 2IFC tasks is two-fold, and relates to decision making and to evidence statistics. In terms of decision making, there appears to be no inherent reason for preferring one 2IFC decision over the other. This is largely a consequence of experimental design. 2IFC experiments typically use equal 2IFC prior probabilities, neutral payoffs and symmetrical instructions to observers (ones that do not favour one decision over another). In SIFC experiments, systematic manipulations of priors, payoffs and instructions have been shown to change the operating characteristics of observers (Egan, Schulman, & Greenberg, 1959; Schulman & Greenberg, 1970; Green & Swets, 1974; Emmerich, 1968b). Similar manipulations in 2IFC experiments should do the same, but there are few such experiments in the psychoacoustical literature.

The symmetry of evidence statistics results from the *assumption* that there is no temporal interference across observation intervals; specifically, that the statistical properties of $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ do not change according to the observation interval, and that $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ are independent across observation intervals (Egan, 1975; Green & Swets, 1974, Section 3.2.4). If this is the case, then a 2IFC decision rule based on differences of $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ results in mirror-symmetric distributions the 2IFC decision axis, which in turn results in a symmetric 2IFC ROC curve (Green & Swets, 1974; Egan, 1975). A 2IFC decision rule based on ratios would do the same (where the logarithm of the ratio results in mirror-symmetric

distributions). If the assumptions above do not hold, for example if there are time-order effects due to memory, or to temporal differences in masking (Jeffress, 1970; Lakey, 1976), then there may be asymmetry on a 2IFC decision axis, and in the resulting 2IFC ROC curve.

## 1.2 Measures of performance

There are a plethora of measures of task performance (or measures of sensitivity) in TSD (Simpson & Fitter, 1973; Green & Swets, 1974; Egan, 1975). Three widely used measures are the area under the ROC curve (denoted here as $\mathcal{A}$), the Gaussian-related measure, $d'$, and the proportion of correct decisions, $P(C)$. Two other measures, $\mathcal{D}_2$ and $\mathcal{D}_6$, are also used later. These are both special cases of a new multiple-event measure, $\mathcal{D}_n$ (Scurfield, 1995, 1996, 1998). $\mathcal{A}$ is the primary measure used in this thesis, and both $d'$ and $\mathcal{D}_2$ are calculated using transforms of $\mathcal{A}$.

The terms *measures of performance* and *measures of sensitivity* are used interchangably throughout the thesis, depending on the context. The former term applies to discrimination tasks both within psychology and outside of psychology, whereas the latter term reflects usage in psychophysics. The terms *detectability* and *discriminability* could also be used instead of *sensitivity*. *Discriminability* refers to the ability to discriminate between events, regardless of whether the events are correctly labelled, whereas *performance* also includes the ability to correctly label events. For example, an observer could say "no" for the $SN$ event and "yes" for the $N$ event, in which case, discriminability may be high, but performance is low. Performance and discriminability are synonymous when situations where ROC curves lie above the chance line, which is mostly the case in this thesis.

Two categorisations of measures are useful. A measure is either *criterion-free* or *criterion-dependent*, and is either *distribution-free* or *distribution-dependent* (Robinson & Watson, 1970). A criterion-free measure is independent of any specific decision criterion, where the criterion is a parameter or parameters in a decision rule, such as $x_c$ in a theoretical context, or a cutoff on a rating scale in an empirical context (Section 1.3). A distribution-free measure is free of assumptions that the evidence variables take on particular distributional forms.

**The area under the ROC curve, $\mathcal{A}$.** A well known measure in TSD is the area under the ROC curve, denoted here as $\mathcal{A}$ (Green & Swets, 1974; Egan, 1975). It is both criterion-free and distribution-free. An ROC curve is the locus of all possible hit and false alarm rate pairings, taken over all possible criteria. If $X$ is continuous, then the $\mathcal{A}$ is calculated by integration of a continuous ROC curve. If $X$ is discrete, then the ROC curve consists of discrete points in the ROC space. Conventially, these points (along with the points (0,0) and (1,1)) are joined together by line segments to make up a continuous

curve. ROC line segments result when the decision rule given earlier is modified for the discrete case, to take into account the possibility that $X$ may equal the criterion $x_c$. A detailed justification is given by Egan (1975, p. 34), and is discussed by Lapsley Miller et al. (1998). In the discrete case, $\mathcal{A}$ is defined as the area under the curve defined by the joined line segments, and is typically calculated using a trapezoidal rule (McNicol, 1972, pp. 113-116, provides a worked example). Experimental ROC curves are always discrete rather than continuous, which relates to the fact that only a finite number of stimuli can be used in practice.

The experimental analyses in later chapters are primarily based on $\mathcal{A}$, calculated using the trapezoidal rule. There are valid concerns that empirical values of $\mathcal{A}$ obtained in this manner can underestimate theoretical values of $\mathcal{A}$ (Bamber, 1975), particularly when the discrete rating scale ROC curve (including line segments) is a poor approximation to the theoretical curve. These concerns are alleviated in practice when using continuous rating scale methodology (Watson, Rilling, & Bourbon, 1964), which is described in Section 1.3.1.

For both continuous and discrete $X$, the area under the theoretical ROC curve (in an SIFC task) is equal to

$$\mathcal{A} \quad = \quad P(X_{\text{SN}} > X_{\text{N}}) + \frac{1}{2}P(X_{\text{SN}} = X_{\text{N}}) \tag{1.1}$$

(Bamber, 1975; Lapsley Miller et al., 1998). If an observer can correctly discriminate between the two events, then the ROC curve would lie above the chance line and the area under the curve would be greater than 0.5. The better the performance, the higher the ROC curve is above the chance line and the greater the value of $\mathcal{A}$, up to the possible maximum value of one. The measure $\mathcal{A}$ is equal to a scaled Mann Whitney U statistic (Bamber, 1975; Hanley & McNeil, 1982).

**The sensitivity measure $d'$.** The distribution-dependent measure, $d'$, is widely-used throughout psychophysics. It was originally based on the assumption that $X_{\text{SN}}$ and $X_{\text{N}}$ are Gaussian with equal variance and different means (McNicol, 1972; Green & Swets, 1974). The use of $d'$ also extends to any strictly monotonic increasing transform of the decision axis, since the ROC curve is identical to that based on the original decision axis. This extension is referred to as the binormal assumption (Hanley, 1988; Somoza & Mossman, 1991). In practice, $d'$ may be calculated from any point lying on an ROC curve. If the ROC curve is binormal and symmetrical (where symmetry is consistent with Gaussian $X_{\text{SN}}$ and $X_{\text{N}}$ *equal* variance), then $d'$ is criterion-free. This is because all criteria, and all points on the binormal ROC curve, yield the same value of $d'$. If the ROC curve is not binormal and symmetrical, the assumption is not met, and $d'$, is criterion-dependent.

**The relationship between $d'$ and $\mathcal{A}$.** When the assumptions underlying $d'$ are met, then

$$d' = \sqrt{2}\,\Phi^{-1}(\mathcal{A}) \tag{1.2}$$

and

$$\mathcal{A} = \Phi\left(\frac{d'}{\sqrt{2}}\right),$$

where $\Phi$ is the cumulative distribution function of the standard N(0,1) Gaussian random variable, and $\Phi^{-1}$ is its inverse (Wilcox, 1968; Simpson & Fitter, 1973).

Values of $d'$ reported in later chapters are always calculated using Equation 1.2, where $\mathcal{A}$ is calculated by using the trapezoidal rule. The measure $d'$ is not interpreted in terms of a Gaussian decision axis, or any strictly monotonic increasing transform of it. Rather, $d'$ is viewed as a scaled $z$-score, without requiring assumptions about the decision axis (Wilcox, 1968). When used as a scaled $z$-score, $d'$ is criterion-free and distribution-free because $\mathcal{A}$ is criterion-free and distribution-free.

**Measures based on Gaussian evidence distributions: $d'$, $d_z$, and $d_s$.** As well as $d'$, two other measures that are based on the assumption that $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ are Gaussian random variables are $d_z$, and $d_s$. Let $\mu_{\mathrm{N}}$ and $\sigma_{\mathrm{N}}$ be the mean and standard deviation, respectively, of the $N$ distribution, and let $\mu_{\mathrm{SN}}$ and $\sigma_{\mathrm{SN}}$ be the mean and standard deviation, respectively, of the $SN$ distribution. If $\sigma_{\mathrm{N}} = \sigma_{\mathrm{SN}}$, then

$$d' = \frac{\mu_{\mathrm{SN}} - \mu_{\mathrm{N}}}{\sigma_{\mathrm{N}}}. \tag{1.3}$$

In theory, $d'$ can only be calculated from parameters of the Gaussian distributions if the variances, or standard deviations, are equal. In the more general case where the variances may or may not be equal,

$$d_z = \frac{\mu_{\mathrm{SN}} - \mu_{\mathrm{N}}}{\left(\frac{\sigma_{\mathrm{SN}}^2 + \sigma_{\mathrm{N}}^2}{2}\right)^{\frac{1}{2}}}, \tag{1.4}$$

(Jeffress, 1967), and

$$d_s = \frac{\mu_{\mathrm{SN}} - \mu_{\mathrm{N}}}{\left(\frac{\sigma_{\mathrm{SN}} + \sigma_{\mathrm{N}}}{2}\right)}, \tag{1.5}$$

(Simpson & Fitter, 1973).[5] If $\sigma_{\mathrm{N}} = \sigma_{\mathrm{SN}}$, then $d_z$ and $d_s$ are both equal to $d'$ in Equation 1.3.

[5] The measure $d_z$ has also been labelled as $d'$ (de Boer, 1966), and also as $d_a$ (Simpson & Fitter, 1973; Bamber, 1975). The measure $d_s$ has also been labelled $d_e$ (Green & Swets, 1974).

**Proportion correct.** The proportion of correct decisions, or $P(C)$, is a distribution-free, criterion-dependent measure, that can be applied equally well to either the SIFC task or the 2IFC task. Given a hit rate and a false alarm rate in an SIFC task ($HR_I$ and $FAR_I$),

$$P(\mathrm{C})_{\mathrm{SIFC}} \quad = \quad HR_I \times P(SN) + (1 - FAR_I) \times P(N). \tag{1.6}$$

Similarly, for a hit rate and a false alarm rate in a 2IFC task ($HR_{II}$ and $FAR_{II}$),

$$P(\mathrm{C})_{\mathrm{2IFC}} \quad = \quad HR_{II} \times P(\langle \mathrm{SN}, \mathrm{N} \rangle) + (1 - FAR_{II}) \times P(\langle \mathrm{N}, \mathrm{SN} \rangle).$$

$P(\mathrm{C})_{\mathrm{2IFC}}$ is widely used as a measure of sensitivity, and is generally seen as a bias-free measure. In either task, however, $P(C)$ varies as a function of the criterion, because both $HR$ and $FAR$ are functions of the criterion. This makes it uncertain whether a change in $P(C)$ reflects a change in sensitivity or of bias, a problem affecting both the SIFC task and the 2IFC task.

Proportion correct is used in Chapters 6 and 8, where it is calculated from a ROC curve by using the hit and false alarm rates at the point where the ROC curve crosses the negative diagonal. For any point on the negative diagonal, $HR = 1 - FAR$, so $P(C) = HR$, from Equation 1.6.[6] The ROC point generated by an unbiased observer in a symmetric binary-decision 2IFC task lies on the negative diagonal.

**The relationship $\mathcal{A}_{\mathrm{SIFC}} = P(C)_{\mathrm{2IFC}}$.** A key relationship in theoretical ROC analysis is that the proportion of correct decisions in the 2IFC task is equal to the area under the SIFC ROC curve (Green & Swets, 1974; Egan, 1975; Bamber, 1975; Lapsley Miller et al., 1998). This result, originally derived by Green, is distribution-free. It is claimed to hold experimentally (Green & Moses, 1966; Emmerich, 1968a), but the variability in performance found across observers (Emmerich, 1968a) and across multiple experimental replications (Emmerich, 1968a; Lapsley Miller et al., 1998) suggest that experimental claims that the measures are equal must be taken with caution.

**$\mathcal{A}_{\mathrm{SIFC}}$, $P(C)_{\mathrm{2IFC}}$, and bias.** Whereas $\mathcal{A}_{\mathrm{SIFC}}$ is a criterion-free measure of performance in the SIFC task, $P(\mathrm{C})_{\mathrm{2IFC}}$ is *not* a criterion-free measure of performance in the 2IFC task. Corrections for bias in the 2IFC task have been suggested because, "we should like an index of sensitivity uncontaminated by response bias," (Green & Swets, 1974, p. 409). If one of the main reasons for using the 2IFC task and $P(\mathrm{C})_{\mathrm{2IFC}}$ is because 2IFC performance *should* be unbiased, then it is better to use 2IFC ROC analysis, which systematically takes 2IFC bias into account, and to calculate $\mathcal{A}_{\mathrm{2IFC}}$ (a criterion-free measure), than it is

[6] If there is no ROC point exactly on the negative diagonal, then $P(C)$ is taken from the intersection of the negative diagonal with the line segment between the two ROC points that straddle the diagonal.

to calculate $P(\mathrm{C})_{\mathrm{2IFC}}$ (even when corrected for bias). If necessary, $P(\mathrm{C})_{\mathrm{2IFC}}$ may also be calculated from the 2IFC ROC curve. In an early study of 2IFC ROC analysis, Schulman and Mitchell (1966) state that

> The use of confidence ratings circumvents the problem of bias since an entire operating characteristic is obtained. At the same time, existing listener bias, given nearly free rein, may be measured by those interested in such phenomena. (Schulman & Mitchell, 1966, p. 474)

In short, 2IFC ROC analysis provides an entire ROC curve, and is much more informative than the single ROC point on which $P(\mathrm{C})_{\mathrm{2IFC}}$ is based.

**Scurfield's measure of sensitivity, $\mathcal{D}_2$**

The measure of sensitivity $\mathcal{D}_2$ was recently developed by Scurfield (Scurfield, 1995, 1996). It can be derived from $\mathcal{A}$, and is both criterion-free and distribution-free. $\mathcal{D}_2$ follows from an information-theory analysis of the fundamental detection problem, and is expressed in units of information (e.g. bits).[7] Scurfield showed that $\mathcal{D}_2$ measures the amount of information that is contained in an observer's decisions about event-orderings. The following explanation is based on the development of $\mathcal{D}_2$ in an SIFC task, showing how $\mathcal{D}_2$ is derived from $\mathcal{A}$ (specifically, $\mathcal{A}_{\mathrm{SIFC}}$). The same development, under different notation, also applies to $\mathcal{D}_2$ calculated in a 2IFC task from $\mathcal{A}_{\mathrm{2IFC}}$. How $\mathcal{D}_2$ is used is made clear from the context in which it appears.

Given two values, $x_{\mathrm{SN}}$ and $x_{\mathrm{N}}$, sampled from continuous random variables, $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$, the information gained from knowing that $x_{\mathrm{SN}} > x_{\mathrm{N}}$ is a function of the probability $P(X_{\mathrm{SN}} > X_{\mathrm{N}})$, namely

$$I(X_{\mathrm{SN}} > X_{\mathrm{N}}) \;=\; -\log\left(P(X_{\mathrm{SN}} > X_{\mathrm{N}})\right),$$

and the information gained from knowing that $x_{\mathrm{SN}} < x_{\mathrm{N}}$ is

$$I(X_{\mathrm{SN}} < X_{\mathrm{N}}) \;=\; -\log\left(P(X_{\mathrm{SN}} < X_{\mathrm{N}})\right),$$

These are both non-negative, finite values, assuming the probabilities do not equal either zero or one. Prior to sampling $x_{\mathrm{SN}}$ and $x_{\mathrm{N}}$, the average uncertainty about whether $\{x_{\mathrm{SN}} > x_{\mathrm{N}}\}$ or $\{x_{\mathrm{SN}} < x_{\mathrm{N}}\}$ will occur is

$$\begin{aligned}\mathcal{H}_2 &= -P(X_{\mathrm{SN}} < X_{\mathrm{N}})\log\left(P(X_{\mathrm{SN}} < X_{\mathrm{N}})\right) - P(X_{\mathrm{SN}} > X_{\mathrm{N}})\log\left(P(X_{\mathrm{SN}} > X_{\mathrm{N}})\right)\\ &= -(1-\mathcal{A})\log(1-\mathcal{A}) - \mathcal{A}\log(\mathcal{A}),\end{aligned}$$

[7] Scurfield showed how an observer in a two-event discrimination task can be viewed as a binary-symmetric information channel. The derivation of $\mathcal{D}_2$ given here is directly from Scurfield (1995). Details are also available in Scurfield (1996, 1998), Lapsley Miller et al. (1998) and Lapsley Miller (1999). $\mathcal{D}_2$ is unrelated to Sakitt's (1973) D, Schulman and Mitchell's (1966) $D_{YN}$, or Green and Swets's (1974) $D(\Delta m, s)$.

where $P(X_{\text{SN}} > X_{\text{N}}) = \mathcal{A}$ is the area below the ROC curve for continuous $X$, by Equation 1.1, and $P(X_{\text{SN}} < X_{\text{N}}) = 1 - \mathcal{A}$ is the area above the ROC curve in the ROC space. If information is measured in bits then the logarithms are taken to the base 2, and $\mathcal{H}_2$ can vary from a minimum of zero bits (if $\mathcal{A} = 0$ or $\mathcal{A} = 1$) to a maximum of one bit (if $\mathcal{A} = 1 - \mathcal{A} = 0.5$). Scurfield defined his measure of discriminability, $\mathcal{D}_2$, as the "difference between the maximum [possible] value of $\mathcal{H}_2$ and the obtained value of $\mathcal{H}_2$," (Scurfield, 1995, p. 24), which is

$$\mathcal{D}_2 \quad = \quad \log 2 - \mathcal{H}_2.$$

Using base 2 logarithms,

$$\mathcal{D}_2 \quad = \quad 1 + (1 - \mathcal{A}) \log_2(1 - \mathcal{A}) + \mathcal{A} \log_2(\mathcal{A}) \tag{1.7}$$

as measured in bits.[8]

Equation 1.7 is for continuous $X$. Although modifications are possible when $X$ is discrete, because $P(X_{\text{SN}} = X_{\text{N}})$ may be non-zero, Scurfield (1995) recommends using Equation 1.7 for both the discrete and continuous cases, where $\mathcal{A}$ is calculated using Equation 1.1 in both cases.

$\mathcal{D}_2$ is strictly monotonic increasing with both $d'$ and $\mathcal{A}$ for $d' > 0$ and $\mathcal{A} > 0.5$. Unlike the other measures though, $\mathcal{D}_2$ does not distinguish between ROC performance below the chance line and ROC performance above the chance line. From Equation 1.7, $\mathcal{D}_2$ would be the same for $\mathcal{A} = 0.2$, for example, as for $\mathcal{A} = 0.8$. Because of this, $\mathcal{D}_2$ is a primarily a measure of discriminability (the ability to discriminate between events), rather than purely a measure of performance (the ability to discriminate *and* correctly label events). ROC performance below the chance line rarely appears in this thesis, and so for practical purposes, $\mathcal{D}_2$ is treated as a measure of performance.

$\mathcal{D}_2$ fits into a more general framework of multiple-event and multiple-interval tasks (Scurfield, 1995, 1996, 1998). $\mathcal{D}_2$ is a specific example of a more general measure, $\mathcal{D}_n$, which describes overall performance in an $n$-event discrimination task. Most of the results that appear in later chapters are from two-event experiments, so $\mathcal{D}_2$ is used. Data analysis from Lapsley Miller (1999), which is based on the six-event measure $\mathcal{D}_6$, is described in Section 8.4.2.

## 1.3 Rating scale experiments

The SIFC and 2IFC tasks were described as *binary-decision tasks* because of the number of possible decisions. A binary-decision task can be posed as a question such as "Did the *SN* event occur?", two possible answers to which are "yes" and "no". It is also reasonable

[8] The subscript "2" in $\mathcal{D}_2$ and $\mathcal{H}_2$ refers to a two-event task, and not to the base of the logarithm.

| Event | Rating: 1 | 2 | 3 | 4 | $\cdots$ | $q-1$ | $q$ | No. Stimuli |
|---|---|---|---|---|---|---|---|---|
| **SN** | | | | | $\cdots$ | | | $n_{\mathrm{SN}}$ |
| **N** | | | | | $\cdots$ | | | $n_{\mathrm{N}}$ |

TABLE 1.2: Event-decision matrix for a $q$-point rating scale experiment, where $n_{\mathrm{SN}}$ and $n_{\mathrm{N}}$ are the number of stimuli per event.

to allow other answers also, such as "maybe", "possibly so", "probably not", and so on (McNicol, 1972; Watson et al., 1973). A graded decision scheme can apply to any two-event discrimination task, and is called an *ordinal rating scale*, or just a *rating scale*. The implementation and benefits of using rating scales in psychophysical discrimination tasks have been widely described (Egan et al., 1959; Watson et al., 1964; Schulman & Mitchell, 1966; Emmerich, 1968b; Robinson & Watson, 1970; McNicol, 1972; Watson et al., 1973; Green & Swets, 1974).

Let $q$ be the number of possible ratings that defines a *discrete* $q$-point rating scale, where $q$ is greater than or equal to two. The results of a rating scale experiment can be summarised in a $2 \times q$ event-decision matrix such as in Table 1.2. It is analagous to the $2 \times 2$ table which results from a binary-decision task (Table 1.1). Table 1.2 applies to an SIFC task, but could apply equally well to a 2IFC task, given only minor modification to the labelling of events (Schulman & Mitchell, 1966). If $q = 2$, then a rating scale task reduces to a binary-decision task, and Table 1.2 reduces to Table 1.1, so the binary-decision task is seen as a limiting case of a rating scale task.

Initial entries in Table 1.2 are counts of the number of times each event-decision conjunction occurred. These are then converted into relative proportions by dividing by the number of stimuli per event. An empirical ROC point results from setting a cutoff on the rating scale and applying a criterion-based decision rule to the rating scale (rather than to a decision axis, as before). Cumulating the proportion of trials greater than or equal to the cutoff, conditional on *SN* and on *N*, results in a hit rate and a false alarm rate, respectively. The steps of cumulation and conversion into proportions may be done in either order (a fact that is relevant to the equivalence between pooled-ROC curves and mean ROC curves, described in Section 2.3). The set of hit and false alarm rate pairings, taken over all possible rating cutoffs, defines the rating scale ROC curve.

A crucial characteristic of a rating scale is that it is an *ordered* scale with respect to the discrimination task that is demanded of an observer. The graded decisions may be interpreted as reflecting the *confidence* of an observer in the occurrence of a particular event, say the *SN* event, on a given trial (Green & Swets, 1974), so that rating value and confidence are linked in an increasing manner. For example, the decision *"1"* reflects

the lowest confidence that *SN* occurred, whereas the decision *"q"* in Table 1.2 reflects the highest confidence that the *SN* event occurred. The way in which rating categories between the ends are used is determined by observers. They may be instructed or trained to use all categories equally often. This results in a set of approximately equally-spaced points in the ROC space, and a well-defined rating scale ROC curve, given a large enough rating scale.

Successive criteria on a rating scale have been interpreted as reflecting successive criteria defined by an observer on a decision axis (McNicol, 1972; Green & Swets, 1974). An equivalent interpretation is that rating scales have an order-preserving relationship with the decision axis (Green & Swets, 1974; Metz & Shen, 1992). This property holds if there is a monotonic increasing function that maps the decision axis onto the rating scale. Such functions, and how they may be estimated, are the topic of Chapter 4. The theoretical implications of such transforms are explored in Chapter 5.

In TSD, it is conventional that successive ratings on a rating scale are assigned successive positive integers, namely $\{1,2, \ldots q\}$. Although conventional, this is not necessarily required. Numerical values are not even required, as long as there is an implied order to the categories (semantically-labelled scales are used, for example). Developments in the following chapters require that real-valued numbers are assigned to each rating category. The numbers *do not need to be integers*, although they often are integers in practice, because of convenience, convention and simplicity.

### 1.3.1 Continuous rating scales

There are two broad types of rating scales: discrete and continuous. In a discrete rating scale experiment, observers may indicate their confidence that a particular event had occurred by pushing one of a row of buttons, for example. In a continuous rating scale experiment, they may indicate their confidence by positioning a slider on a continuum, for example. Watson et al. (1964) were the first researchers to apply a continuous rating scale to ROC analysis. In their experiment, observers indicated confidence by adjusting the position of a 14-inch mechanical slider. One end of the slider was used to indicate high confidence in the *N* event, while the other end was used to indicate high confidence in the *SN* event. Observers were instructed that they should "use the whole rating scale in indicating their certainty that the signal was presented, and be *very* sure when using positions close to the ends" (Watson et al., 1964, p. 285, original emphasis). Watson et al. manually measured the slider position on each trial and partitioned the continuum into an ordered set of $q = 36$ intervals, each of which was interpreted as a separate rating category. The data was analysed in terms of a 36-point rating scale, which provided very well-defined, 37-point ROC curves (including the points (0,0) and (1,1)).

There are different ways of implementing continuous rating scales, including via mechanical sliders, icons on a computer screen, and by reporting posterior probabilities. Con-

tinuous rating scales have primarily been implemented using mechanical sliders.[9] They also may be implemented by using a computer mouse to move a pointer icon on a computer screen (Friedman & Massaro, 1998). Furthermore, observers could be asked to estimate the posterior probability that a specific event had occurred during a trial (Swets et al., 1961; Rockette, Gur, & Metz, 1992), which provides a continuous scale, in principle.[10]

Continuous rating scales may seem continuous from the point of view of an observer, but they are measured and analysed as discrete scales in practice, typically by using a few dozen categories or less. Even if the number of categories was much larger, there are practical limitations on the resolution of the measurement device that effectively result in a discrete scale (e.g. when using a computer mouse). Even estimated posterior probabilities are discrete in practice, because human observers are unlikely to report values to more than 2 or 3 significant figures. The essential distinction between discrete and continuous scales is the flexibility available by using the latter. Continuous scales provide a much finer resolution and ability to grade decisions than do discrete scales.

The number of rating categories used to analyse a continuous rating scale is up to the experimenter. Watson et al. (1964) showed that the amount of information contained in a multiple-category partition of a continuous scale increases as more categories are used, and that most of that increase is achieved within 20 categories. In practice, as many categories as possible should be used if there is no extra cost involved in doing so.[11]

[9] (Watson et al., 1964; Emmerich, 1968b, 1968a; Leshowitz, 1969; Robinson & Watson, 1970; Taylor et al., 1991; Whitmore et al., 1993; Lapsley Miller et al., 1998; Galvin, Podd, Drga, & Whitmore, 1998; Lapsley Miller, 1999)

[10] In an important article on the role of TSD in psychology, Swets et al. (1961) asked observers to estimate the posterior probability of the *SN* event, but had observers categorise each probability estimate *themselves* into one of six numerical ranges, and then report which range they had used. In practice, this was a discrete six-point rating scale and not a continuous rating scale.

[11] The following number of rating categories have been used in practice: 12 (Leshowitz, 1969); 20 (Emmerich, 1968a, 1968b); 36 (Watson et al., 1964; Robinson & Watson, 1970; Taylor et al., 1991; Galvin et al., 1998); 64 (Lapsley Miller et al., 1998); 600 (in Section 8.3); and approximately 2000 (Lapsley Miller, 1999).

# Chapter 2

# Observer inconsistency in discrimination tasks

The major factor influencing performance in a discrimination task is the nature of the stimuli. For example, the detection of tonal signals in the presence of masking noise depends on the duration and frequency of the tone, as well as the signal-to-noise ratio (Green, Birdsall, & Tanner, 1957; Green & Swets, 1974). Two other factors that have a substantial bearing on performance are observer bias and observer inconsistency. As discussed in the preceding chapter, the effect of bias is well recognised, and has been systematically incorporated into modern psychophysics. The effect of observer inconsistency is less well recognised.

In a discrimination task, observer inconsistency occurs when different decisions are made for the *same* stimulus. This applies across individuals in a group, and also applies within individuals, if the stimulus is reproducible and can be presented multiple times.[1] The effect of observer inconsistency is to increase the amount of *error* in the task. It is not a trivial influence, because experimental evidence suggests that there is as much decision variability due to observer inconsistency as there is due to the variation of stimuli in a stimulus set.[2] If the variability and error due to observer inconsistency was removed, task performance could increase substantially.

Inconsistent decision making based on any particular stimulus has consequences for performance based on an entire stimulus set. For example, if an experiment is replicated multiple times using identical stimuli, and an ROC curve is calculated for each replication,

[1] (Smith & Wilson, 1953; Watson, 1963; Green, 1964; Pfafflin & Mathews, 1966; Ahumada, 1967; Pfafflin, 1968; Thijssen & Vendrik, 1968; Yerushalmy, 1969; Bell & Nixon, 1971; Ahumada & Lovell, 1971; Ahumada, Marken, & Sandusky, 1975; Boven, 1976; McAulay, 1978; Siegel, 1979; Gilkey, 1981; Spiegel & Green, 1981; Taylor, 1984; Gilkey, Robinson, & Hanna, 1985; Siegel & Colburn, 1989; Isabelle & Colburn, 1991; Taylor et al., 1991; Metz & Shen, 1992; Whitmore et al., 1993; Galvin et al., 1998; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

[2] (Swets, Shipley, McKey, & Green, 1959; Watson, 1963; Green, 1964; Ahumada et al., 1975; Spiegel & Green, 1981; Gilkey, 1981; Siegel & Colburn, 1989)

then ROC curves and measures of performance tend to differ across replications. Overall performance is often much less than what it could be without the extra error.[3]

Apparent discrepancies between psychophysical theory and data may be because of error due to observer inconsistency, or because of an inappropriate theory, or both. Since the effects of observer inconsistency are confounded with those due to stimuli, these two effects are not immediately separable in an experimental data set. The situation is reminiscent of the confounding of sensitivity and bias in earlier psychophysical paradigms, in that there are two factors that influence performance, and a change in one could be mistaken for a change in the other. The solution to the confounding of bias and sensitivity was a change in experimental design and data analysis. Similarly, changes in experimental design and data analysis are also needed to account for *and remove* the effects of observer inconsistency. Such changes have been available for some time (e.g. Swets et al., 1959; Watson, 1963), but they have not been widely implemented.

### Identical stimuli

In order to assess observer inconsistency and to remove its effects, it is necessary to run a multiple-observation experiment, in which the same stimulus set is presented multiple times to the same observer, or is presented one or more times to each of a group of observers. The stimulus set needs to be the same for each set of observations, otherwise decision variability across observations or observers could be attributed solely to sampling variability of the stimuli,[4] rather than to inconsistency on the part of an observer.

Identical stimuli refer to stimuli that can be faithfully reproduced, or can be simultaneously presented to multiple observers. Reproducible stimuli are possible if waveforms are stored on magnetic tape (Swets et al., 1959; Green, 1964) or as digital waveforms in a computer (Pfafflin, 1968; Siegel, 1979). Simultaneously identical stimuli could be achieved by running multiple observers in a free-field environment, or by splitting a single electronic channel into multiple channels, each of which leads to a different pair of headphones (Smith & Wilson, 1953; Watson, 1963; Watson, Franks, & Hood, 1972). Reproducible visual stimuli have been used in medical diagnostic tasks, for example, by re-presenting the same set of X-ray films to different observers (Yerushalmy, 1969; Metz & Shen, 1992).

[3](Smith & Wilson, 1953; Watson, 1963; Ahumada, 1967; Yerushalmy, 1969; Boven, 1976; McAulay, 1978; Taylor, 1984; Taylor et al., 1991; Metz & Shen, 1992; Whitmore et al., 1993; Galvin et al., 1998; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

[4]Since performance is known to vary across masker waveform samples (Pfafflin, 1968; Gilkey, 1981; Gilkey et al., 1985; Isabelle & Colburn, 1991).

### Multiple-observation experiments

Two broad types of multiple-observation experiments that have been run are *multiple-presentation* experiments, and single-presentation *multiple-replication* experiments. In both cases, an observer's task is to discriminate between predetermined experimental events, such as the $SN$ and $N$ events in a signal-detection task. In a multiple-presentation experiment, stimuli are presented more than once per trial, before a decision is made (Swets et al., 1959; Berg, 1987, 1989, 1990; McKinley & Weber, 1994), whereas in a single-presentation experiment, stimuli are presented only once per trial, prior to the decision. A multiple-replication experiment involves repeating a single-presentation experiment multiple times, using an identical stimulus set each time.[5] Multiple replications can be achieved within observers, or across observers, or both. If stimuli are reproducible, a group of observers can be run sequentially rather than simultaneously, and multiple replications are possible from the same observer, either running within a group, or running individually.[6]

The benefit of using either type of multiple-observation experiment is that performance generally improves as an observer is allowed more observations of the same stimulus. Both multiple-replication experiments and multiple-presentation experiments share theoretical features in common, although the assumptions underlying each are different. The main difference is that multiple-presentation experiments rely on an observer averaging out error internally, prior to the decision on each trial, whereas in multiple-replication experiments, variability in decisions is averaged out externally by the experimenter, once a data set is collected. Whether or not the two types of averaging are the same, or have the same effect, is unclear.

Much of experimental psychophysics is based on single-presentation experiments that are run only once. Although observer inconsistency can affect any single replication, its effects can only be assessed and removed once multiple replications have been run. This thesis is primarily concerned with multiple-replication experiments, and the removal of error due to inconsistency. Some of the results in later chapters may also apply to multiple-presentation experiments.

### Overview of chapter

This chapter deals with descriptions of observer inconsistency, and what can be done to remove it. Section 2.1 deals with classification of error in discrimination tasks, and attempts to clarify what is intended in psychophysical models of error. Section 2.2 introduces a multiple-replication experiment, originally from Taylor et al. (1991), which graphically

[5] (Smith & Wilson, 1953; Watson, 1963; Ahumada, 1967; Yerushalmy, 1969; Bell & Nixon, 1971; Boven, 1976; McAulay, 1978; Taylor, 1984; Taylor et al., 1991; Metz & Shen, 1992; Whitmore et al., 1993; Galvin et al., 1998; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

[6] (Yerushalmy, 1969; Bell & Nixon, 1971; Ahumada & Lovell, 1971; Ahumada et al., 1975; Taylor, 1984; Taylor et al., 1991; Metz & Shen, 1992; Whitmore et al., 1993; Galvin et al., 1998; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

shows the results of observer inconsistency, and the detrimental effect it can have on performance. Section 2.3 deals with mean ROC analysis, which is one way of removing variability from a multiple-replication data set. When applied to Taylor et al.'s data set, it is seen that although mean ROC analysis removes variability, it does not remove error from the data—mean ROC performance remains relatively poor. Section 2.4 develops group operating characteristic (GOC) analysis, which is a different way of analysing the results of a multiple-replication experiment. GOC analysis is also applied to Taylor et al.'s data set. It provides a great improvement over mean ROC analysis, both in terms of performance and in the recovery of the theoretical ROC curve. Section 2.5 provides a historical development of GOC analysis, which can be traced at least as far back as Smith and Wilson's (1953) study of multiple observer psychophysics.

## 2.1 Classification of error

Error in multiple-observation experiments has been classified according to its *sources* as either *external noise* or *internal noise*, that is, external or internal with respect to an observer (Swets et al., 1959; Green, 1964). Error has also been classified according to its *effects*, depending on whether the error is correlated across observations, or is uncorrelated (Taylor et al., 1991; Metz & Shen, 1992; Sorkin & Dai, 1994). Classification of error according to its effects involves the concepts of *unique noise* and *common noise* (Taylor et al., 1991).

### 2.1.1 Internal noise and external noise

Researchers in psychophysics are interested in the error due to the types of stimuli that are used, because it is indicative of the capabilities of the observer under investigation. Such *external noise* has also been called *stimulus-dependent noise* (Gilkey et al., 1985). A major source of external noise would be masking noise that is presented in conjunction with a signal.

The effects of external noise on discriminability are compounded by the effects of other sources of noise that are internal to an observer. These extra effects come under the general label of *internal noise*, which is seen as "internal fluctuation" or "random perturbation of the sensory processes" (Green, 1964; Spiegel & Green, 1981). Internal noise sources are many and varied, and include things such as muscle tension, heartbeat and breathing (Soderquist & Lindsey, 1972; Lindsey & Soderquist, 1972), physiological aural noise in the ear canal (Shaw & Piercy, 1962; Anderson & Whittle, 1971), neural noise and membrane noise (Fatt & Katz, 1950), memory noise (Durlach & Braida, 1969; Jesteadt & Sims, 1975), sequential effects (Speeth & Mathews, 1960; Green, 1964; Spiegel & Green, 1981; Triesman & Faulkner, 1984), fatigue (Smith & Wilson, 1953), and criterion variability (Tanner, 1961; Wickelgren, 1968; Triesman & Faulkner, 1984).

On the topic of *internal noise*, Green (1960b) said

> If the concept is to have any importance, it must be made specific. This implies that we have to (1) state exactly what this noise is, i.e. that we have to characterize it mathematically, (2) specify in what way it interacts with the detection or discrimination process, and (3) evaluate specifically what effect it will have on performance. (Green, 1960b, p. 1202)

It is clear from this description that Green views internal noise as a theoretical construct which is part of a model or theory of an observer. Together, Green's first two points approximately equate to a noise-affected decision axis and decision rule. Green's third point relates to effects on experimental data, and may be addressed at two different levels, *molecular* and *molar* (Green, 1964; Gilkey, 1981; Gilkey et al., 1985; Gilkey & Robinson, 1986). Molecular psychophysics describes performance on the basis of an individual stimulus, whereas molar psychophysics describes performance on the basis of an entire stimulus set. The fundamental effect of internal noise is molecular, and is apparent as observer inconsistency in decisions that are made based on repeated presentations of a given stimulus. The global effect of internal noise is molar, and is apparent in the variability of ROC curves and measures of performance such as $\mathcal{A}$ or $d'$. Task performance is typically depressed compared to what it could otherwise be without internal noise. In a later study on observer inconsistency, Green (1964) said

> On an operational level, internal noise is equivalent to the observation that the same physical stimulus may elicit different responses. (Green, 1964, p. 397)

Similarly, Richards and Zhu (1994) said

> The term internal noise is intended to capture the fact that identical stimuli do not always generate identical responses. (Richards & Zhu, 1994, p. 428)

These statements describe observer inconsistency. In comparison to the earlier list of internal noise sources, the statements by Green (1964) and by (Richards & Zhu, 1994) show that there are, in fact, two different but related concepts that are both called *internal noise*. One concept is described in terms of the *sources* of noise, whereas the other is described in terms of the *effects* of noise. Using the same term to describe these two different concepts obscures the difference in emphasis. Furthermore, the double-meaning allows measurements of one concept to be confused with measurements of the other. Although the two concepts are related, they are not synonymous, and can be inconsistent with each other under some circumstances. From hereon, the term *internal noise* is used with reference to sources of noise that are internal to an observer, rather than to the effects of noise.

### 2.1.2 Unique noise and common noise

**Taylor, Boven and Whitmore's formulation of unique noise and common noise.** A break with the original conception, emphasis, and terminology associated with internal noise was taken by Boven (1976), Taylor (1984) and Taylor et al. (1991). Instead of the internal-versus-external noise dichotomy, they distinguish between noise that is correlated across replications, which they label *common noise*, and noise that is uncorrelated across replications, which they label *unique noise*. According to Taylor et al. (1991),

> Unique noise is a statistical concept that refers to the component of the total noise variance of an observer that is uncorrelated with other noise. It may include internal noise. The complement of unique noise is common noise. ... The two most important points are that *unique noise is a statistical concept* [emphasis added] and that, unlike internal noise, it is not identified with any particular source. (Taylor et al., 1991, p. 133)

Although unique and common noise are defined in terms of correlation, it is unclear as to what is correlated or uncorrelated. It also needs to be stated what the "total noise variance of an observer" refers to. Generally, the variability and variance of decisions on a rating scale, for example, are different from the variability and variance of evidence values on a decision axis. Total noise variance may depend on the scale that is being considered, whether it is a rating scale, or a decision axis, or some strictly monotonic increasing transform of a given decision axis.

Whitmore (1999, personal communication) said that the concepts of unique and common noise refer to a model of observer performance based on correlated and uncorrelated noise processes (e.g. time series), akin to Licklider and Dzendolet's two-source model of correlated noise (Licklider & Dzendolet, 1948; Jeffress & Robinson, 1962; McFadden, 1968). Within this context, unique and common noise are viewed as different channels within a multichannel system. Unique noise and common noise sources in Taylor et al.'s (1991) model are akin to random process generators, and the correlation of noise in an experiment is viewed in terms of a mixture of independent unique and common noise processes. In general, and outside of the multichannel context, unique and common noise do not need to be *defined* in terms of correlation, although they could be specified or quantified in terms of correlation.

The concepts of unique and common noise are a useful way of viewing performance in a multiple-replication, discrimination task experiment. Under Taylor et al.'s (1991) model, unique noise results in observer inconsistency, and it is possible to remove the effects of unique noise, essentially by averaging it out (Watson, 1963; Boven, 1976; Taylor, 1984; Taylor et al., 1991; Metz & Shen, 1992; Lapsley Miller, 1999). Once all of the unique noise is removed, whatever data pattern that remains is due to common noise. The removal of unique noise can improve performance substantially and better indicate an observer's

abilities and limitations, once performance is unhampered by inconsistent decision making. Boven summarised this by saying that

> The sources of unique noise are an interesting research problem in themselves, but need not hold up the search for models of hearing based on the common component of noise which will remain if the unique noise can be removed from the data. (Boven, 1976, p. 6)

### Reformulation of unique noise and common noise

In order for the concepts of unique and common noise to be generally applicable, their definitions should be independent of any particular observer, either real or ideal, and of any particular task or experiment. If the definitions depended on the particular observer or experiment, then the concepts would not be broadly applicable.

The aim of the following is to clarify what is intended by *unique noise* and *common noise*, and by the model proposed by Boven (1976), Taylor (1984) and Taylor et al. (1991). The concepts of unique and common noise are based on a broad statistical model of an inconsistent observer in a multiple-replication experiment. The model is called an *equivalent statistical observer (ESO)*.[7] The ESO is set up so that, in principle, the statistics of the decisions of the ESO are identical to the statistics of the decisions of either a real or an ideal observer. The statistics would be the same at a molecular level for any single stimulus across replications, and at a molar level for an entire stimulus set, both across stimuli and across replications.

For the purpose of defining unique and common noise, the ESO is not an ideal observer, in the sense of a stimulus-processor (Section 1.1). The inputs to the ESO are not stimuli, but are instead samples from separate unique and common noise processes. The processes are analogous to the output of a noise generator, or of a random number generator, but are not necessarily defined in time, or space. They are a set of numbers, and in the simplest formulation (used in Chapter 5), there is only one number per source, per stimulus, and per replication. For the purpose of *defining* unique and common noise, it does not matter whether these processes provide a single value each, or a set of values. What is important is what the ESO does with the unique noise and common noise samples, which is to make decisions that have statistical properties that match those of a given observer. The concept of the ESO is only used here to help in the definition of unique and common noise. It is not suggested as a replacement for the concept of an ideal observer.

Figure 2.1 shows a general outline of an ESO, that illustrates the contribution of both unique and common noise to decisions that are made on different replications. For a given stimulus, the decision that is made on each replication is based on two inputs,[8] one input

[7] Boven (1976), Taylor (1984) and Taylor et al. (1991) did not use this term.

[8] Taylor et al. (1991) modelled unique noise as an input, although it can affect all parts of the decision process including the output process. Modelling it as an input is just a simplification (Whitmore, 1999, personal communication).

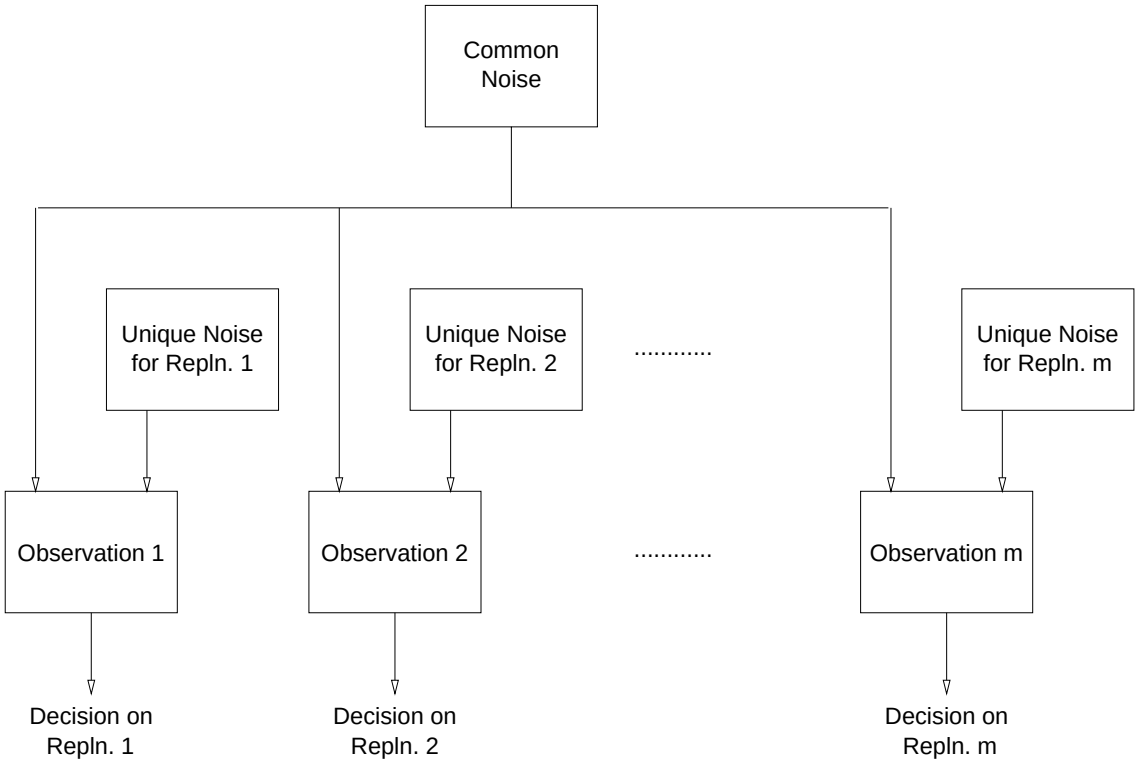


FIGURE 2.1: Block diagram of an equivalent statistical observer, showing how unique and common noise contribute to decisions based on a given stimulus over the course of $m$ replications. Each source outputs a number, or set of numbers, that contribute to the observation (or evidence) on each replication. The common noise is the same across replications, whereas the unique noise differs across replications (after Taylor et al., 1991, Figure 2(b)).

sampled from each of the common and unique noise sources. One type of input is common or identical across replications, and the other type of input is unique to each replication. The decision of the ESO is based on an amalgamation of the two types of input, and is a function of a decision rule. The ESO is like an ideal observer, except applied to outputs from the unique and common noise sources, rather than stimuli. *The unique and common noise sources are not necessarily associated with any particular sources of internal or external noise.*

The ESO is a model of any observer, including any unique-noise-affected ideal observer. Many ideal observers model unique noise in straightforward ways, for example, as an internal Gaussian random variable that is either added or subtracted from Gaussian external noise distributions on the decision axis (Swets et al., 1959; Tanner, 1961; Wickelgren, 1968; Metz & Shen, 1992). In such cases, the proposed ideal observer *is* an ESO, for the simple reason that the partitioning of variability into common and unique components is obvious. Such an ideal observer is based on the assumption that all unique noise can be characterised by a single statistical distribution or process. More complicated

ideal observers can have multiple internal noise sources at various stages of processing (e.g. Taylor, 1984; Durlach, Braida, & Ito, 1986) and non-linear transforms between stages. For example, the initial input is affected by internal noise factors such as physiological noise and extraneous environmental noise. Processing of the input has its own, less tangible, unique noise such as that from inattention, criterion variability and random neural firings. The process of expressing decisions (the output stage) is also affected, for example by the trial-by-trial motor coordination of subjects, as well as measurement factors such as equipment limitations and the size of the rating scale. As more complex ideal observers and models are proposed (e.g. Sorkin & Dai, 1994), the partitioning of their resulting statistics into unique and common components will not necessarily resemble the statistics of the contributing internal and external noise sources. Hence a unique-noise-affected ideal observer is not necessarily an ESO, although in simpler cases, it could be.

Common noise takes on event-conditional forms, and the ESO model treats reproducible signals in the same way as reproducible noise. Decision variability associated with signal-plus-noise trials can be partitioned into common and unique components, as can variability associated with noise-alone trials. Generally, unique noise is modelled as being independent of experimental event, implying that event-conditional forms of a unique-and-common-noise mix are due to event-conditional forms of common noise only. Unless otherwise stated, the term *common noise* will generally refer to common components of variability, regardless of event.

This description of common noise allows for unusual situations in which it is possible to have noisy decisions and still achieve perfect performance (e.g. $\mathcal{A} = 1$ or $d' \to \infty$). This could occur, for example, if there is no unique noise, and if the event-conditional common noise distributions on a decision axis do not overlap. Although there is no variability in decisions across replications (because there is no unique noise), there is still variability in decisions across stimuli (because the $SN$ and $N$ distributions are not identical). Historically, "noise" is associated with error in the task, for example in the masking of a signal, but that is not the case here. The term "noise" is used here as a label to identify a component of the model of an observer in Figure 2.1, and otherwise should not be taken too literally as something that only interferes with detection of a signal.[9]

### The relationship between the internal-external noise dichotomy and the unique-common noise dichotomy

Both common and unique noise are statistical in nature, and are not tied to any particular sources of noise or error. Common and unique *are not synonymous* with the original conception of external and internal noise. Both external and internal noise sources could contribute to either common noise, or to unique noise, or to both. An example of each of

[9] In a similar vein, what constitutes noise, and what constitutes signal, is largely arbitrary.

the four possible combinations across the two dichotomies is given.

- One example where external noise contributes to common noise is where external signals and maskers are identical across replications.[10] The signals and maskers presumably contribute to a component of the decision statistics that is the same across replications, that is, common noise.

- An example where external noise contributes to unique noise is where maskers are different for each replication.[11] The signals and maskers presumably contribute to a statistical component that is different across replications, that is, unique noise.

- An example of internal noise contributing to unique noise is random physiological noise that is not controlled experimentally, such as physiological aural noise in the ear canal (Shaw & Piercy, 1962; Anderson & Whittle, 1971) and neural noise (Fatt & Katz, 1950). The internal noise presumably contributes to the unique noise, because it has a similar effect effect to non-reproducible external masking noise, except that the masker source is internal to the observer.

- An example where internal noise contributes to common noise is where stimuli are presented in phase with either heartbeat or breathing cycles (Soderquist & Lindsey, 1972; Lindsey & Soderquist, 1972). In that case, the masking effects of heartbeat or breathing are reproducible to some degree, and so contribute to the common noise. This type of experiment may seem odd, but it has potential benefits. For example, Soderquist and Lindsey (1972) measured the effects of heartbeat phase on detection and found an improvement in $d'$ of up to 0.6 for 100 Hz tonal signals, depending on heartbeat phase.

The point of these examples is that unique noise is not equivalent to internal noise, and common noise is not equivalent to external noise. Hence, models of observers defined in terms of real sources of decision variability are not the same as models defined in terms of the statistical consequences of decision variability.

In the examples given, each source of internal or external noise does not necessarily contribute exclusively to unique noise, or exclusively to common noise—each source may contribute to both common noise and unique noise. Other known sources of observer inconsistency are difficult to classify as wholly internal noise or as wholly external noise. For example, sequential effects depend on both internal and external circumstances, namely the observer's sequential propensities in decision making (which are internal to the observer), and the particular experimental trial sequence (which is external to the observer).

[10] (Smith & Wilson, 1953; Watson, 1963; Pfafflin, 1968; Gilkey, 1981; Gilkey et al., 1985; Whitmore et al., 1993; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

[11] (Boven, 1976; Taylor, 1984; Taylor et al., 1991; Whitmore et al., 1993; Galvin et al., 1998; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

### The use of experimental design and procedure in determining unique noise and common noise

Experimental design and procedure often, but not always, determines whether a particular source of error contributes to unique noise or to common noise. A particularly important source of error is the set of experimental maskers. As noted in the examples above, whether the maskers mainly contribute to unique noise or to common noise depends upon whether they are different or the same across replications. A combination of both identical and unique maskers can also be used, for example, when reproducible masker transients are presented along with a continuous background masker (e.g. Whitmore et al., 1993; Lapsley Miller et al., 1998; Lapsley Miller, 1999).

The contribution of sequential dependencies to unique noise and to common noise is also a question of experimental design (Boven, 1976; Taylor, 1984; Taylor et al., 1991; Lapsley Miller et al., 1998; Lapsley Miller, 1999). Sequential dependencies depend on past decisions as well as on past stimuli. Assume that identical stimuli are used in a multiple-replication experiment, and that sequential dependencies exist. If the stimulus order is different on each replication, perhaps in a random or haphazard order, then sequential dependencies contribute mainly (if not entirely) to unique noise, because their effect is randomised across replications. If the stimuli in the experimental stimulus set are presented in the same order on each replication, then sequential dependencies could contribute towards common noise (because the stimulus sequence is the same), as well as to unique noise (if the decision sequence is different across replications, due to other sources of unique noise). If it is desirable to average out the effects of sequential dependencies, then an experiment should use a different haphazard order on each replication. This helps ensure sequential dependencies contribute to random error rather than constant error, because random error can then be removed using GOC analysis.

The contribution of physiological noise to unique noise, or to common noise, may be influenced by experimental design. As noted previously, internal noise associated with the heartbeat and breathing cycles can be made mainly common across replications, if desired, by timing the presentation of stimuli with particular phases of the heartbeat or breathing cycles. This type of internal noise would usually be associated with unique noise, because it is not usually controlled for by the experimental procedure. Soderquist and Lindsey (1972) showed how this internal noise could be controlled for, at least to some degree.

If identical stimuli are used across replications, then their effects may contribute mainly to common noise. However, if the stimulus set is different on each replication, then the stimuli would contribute to unique noise. This follows even if the stimulus sets have been generated by the same mechanism (e.g. sampled from the same noise generator), and have the same parameters and statistics, but are non-identical. This is why it is important to use identical stimuli when the aim of a multiple-replication experiment is to understand performance without confounding by unique noise.

Using identical stimuli does not in itself guarantee the presence of common noise, however. Identical stimuli could contribute partly or even entirely to unique noise. For example, in aural discrimination experiments involving replications of identical long-duration stimuli, an observer may only attend to, or sample, a portion of a reproducible waveform on one replication and another portion on another replication (Siegel & Colburn, 1983; Taylor et al., 1991). One possible mechanism for this effect would be if the integration time of the ear was small relative to the duration of the signal. Such sampling would effectively decrease the amount of common noise and increase the amount of unique noise, because the evidence that is *used*—as opposed to *presented*—changes to some extent from replication to replication. A similar sampling effect could occur in the frequency domain for wideband signals if the effective bandwidth of the ear was small relative to the bandwidth of the signal. The effect could also occur in visual discrimination tasks, if observers were able to scrutinise a complex scene in detail on each replication, where the detail attended to differs on each replication.

### Statistical aspects of existing models

Figure 2.1 is not specific about the statistical details of decision making, since the diagram is only intended to illustrate unique and common noise in general. An ESO that is amenable to detailed statistical modelling is given later in Section 5.2 (particularly Figure 5.2). For any given experiment, details would be be filled in, specifying the nature of the unique and common noise, how they interact, the types of distributions involved, assumptions about correlation and independence, and how all of these factors translate into decisions. Many models of unique-noise-affected ideal observers often rely on the assumption that unique and common noise sources are either uncorrelated (Siegel, 1979; Yost, 1988; Metz & Shen, 1992), or are independent, implying zero correlation.[12]

Along with a common noise random variable, $X$ (or its event-conditional forms, $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$), it is often assumed that unique noise can also be characterised by a single random variable, $U$, and that unique and common noise are *additive* (or subtractive), especially on a decision axis. The mathematical benefit of assuming additivity of uncorrelated noise sources is two-fold: (1) unique-noise-affected event-conditional random variables are simply sums of two random variables, either $X_{\mathrm{SN}} + U$ or $X_{\mathrm{N}} + U$; and (2) that the variance of the sum of unique and common noise is equal to the sum of the variances (the "total noise variance" referred to by Taylor et al., 1991, p. 133).

Under the assumptions of zero correlation, additivity, and having a single unique noise distribution, $U$, then unique and common noise can be characterised in terms of variances.[13] If $\sigma_{\mathrm{u}}^2$ is the unique noise variance, and $\sigma_{\mathrm{c}}^2$ is the common noise variance, then the

[12](Swets et al., 1959; Watson, 1963; Wickelgren, 1968; Wilcox, 1968; Green & Swets, 1974; Boven, 1976; Taylor, 1984; Taylor et al., 1991; Richards & Zhu, 1994)

[13](Swets et al., 1959; Watson, 1963; Wickelgren, 1968; Green & Swets, 1974; Boven, 1976; Taylor, 1984; Metz & Shen, 1992)

total noise variance is $\sigma_c^2 + \sigma_u^2$. The ratio of unique-to-common noise variances, $k = \sigma_u^2/\sigma_c^2$, is one way of characterising the relative contribution of unique noise towards overall performance. (This ratio is usually stated in the literature in terms of internal and external noise.) If $k = 0$, then there is no unique noise, and all of the error in the task is due to common noise; if $k \to \infty$, then unique noise entirely dominates common noise; and if $k = 1$, then unique and common noise contribute equally to the error in the task.

Different ways of estimating $k$ have been proposed, including derivations based on multiple-observation experiments that compare performance when unique noise is fully present with when it has been partly removed (Swets et al., 1959; Watson, 1963; Taylor, 1984), and modified 2IFC experiments that compare performance using identical and non-identical noise samples in each observation interval (Green, 1964; Pfafflin & Mathews, 1966; Siegel, 1979; Spiegel & Green, 1981). Empirical estimates of $k$ vary across observers, experimental parameters and methodologies, and range from extremes of 0.2, Swets et al. (1959) to infinity (Spiegel & Green, 1981). Earlier estimates lie in the vicinity of $k = 1$ (Swets et al., 1959; Watson, 1963; Green, 1964), but later estimates lie in the range from $k = 1$ to about $k = 10$ (Siegel, 1979; Spiegel & Green, 1981; Taylor, 1984; Siegel & Colburn, 1989).

The concept of $k$ is well-established as an indicator of the relative influence of unique noise on performance in a discrimination task. The estimation of $k$ is not a major topic in this thesis, although some of the analyses developed here could contribute towards its estimation. A new way of estimating $k$ is developed in Chapter 4, and follows as a consequence of the main material in that chapter.

Since $k$ represents a ratio of variances of random variables on a decision axis, some type of distributional form must be assumed in order to assign values to $k$ based on experimental data. The type of distribution usually reflects statistical details of a model of unique noise, and the most common assumption is that unique noise is Gaussian (Swets et al., 1959; McNicol, 1972; Siegel, 1979; Metz & Shen, 1992; Sorkin & Dai, 1994). Common noise is also often assumed to be Gaussian, although it need not be (Boven, 1976; Taylor, 1984; Berg, 1987; Richards & Zhu, 1994; Lapsley Miller et al., 1998). Unlike the concept of sensitivity, for which there are various non-parametric measures, there is not, as yet, a general non-parametric analog of $k$ that quantifies the level of unique noise.

## 2.2 Taylor, Boven, and Whitmore's (1991) continuous rating scale experiment

Taylor et al. (1991) ran a series of multiple-replication SIFC frequency discrimination experiments, to demonstrate group operating characteristic (GOC) analysis as a method for reducing the effects of observer inconsistency. The data set for their continuous rating scale experiment was obtained through the courtesy of the authors. The data is re-analysed here to illustrate the effects of observer inconsistency, and to illustrate the steps required and results obtained for three possible types of analysis applicable to a multiple-replication experiment. These analyses are: (1) *single-replication ROC analysis*, (2) *mean ROC analysis*, and (3) *GOC analysis*. Experimental methodology is described below, followed by ROC analysis of single replication data. Mean ROC analysis of the data is presented in Section 2.3, and GOC analysis is presented in Section 2.4.

The effectiveness of GOC analysis can only be demonstrated unambiguously under three conditions: (1) if the theoretical ROC curve is known, (2) if there is enough unique noise to appreciably affect performance, and (3) if enough replications are run to remove unique noise effects. To achieve the first requirement, all of Taylor et al.'s frequency discrimination experiments were unusual, in that the distributions of tonal frequency were completely specified by the experimenters beforehand. The experiments were like aural equivalents of the dice game (Swets et al., 1961; Green & Swets, 1974), in order that the theoretical ROC curves could be known *a priori*. The second and third requirements were achieved largely through experimental design and procedure.

### Method[14]

Three observers ran eight replications each of an SIFC, aural frequency discrimination experiment. A tonal transient was presented on each trial during the observation interval. The transient could have come from a set of high-frequency tones or a set of low-frequency tones.[15] The task of an observer was to provide a confidence rating as to whether or not the higher-frequency event had occurred.

**Observers.** The observers were three adults, two of whom had not previously acted as observers in an experiment. All three of them knew the distributions of tonal frequency. The observers' training was brief, but enough for them to demonstrate proficiency in the task.

[14] Experimental details are taken from Taylor et al. (1991), and also from the original experimental documentation.

[15] "Frequency" refers to tonal frequency in Hertz, and not frequency of occurrence.

**Stimuli.** Reproducible tonal transients were generated by computer as digital code sequences. The absolute duration of each transient was 250 ms. Tones were gated on and off over the first and last 18 ms of the transient and had a constant amplitude over the central 214 ms. The gating used a Kaiser data window with a shape parameter of 11. Digital codes were converted to voltage transients using a 12-bit digital-to-analog converter (DAC) clocked at 7 kHz. The output of the DAC was smoothed using a passive 1.25 kHz low-pass filter, and passed through a passive attenuator, additive mixer, and a headset amplifier which drove a pair of headphones. An analog Gaussian noise generator produced a continuous masker process, which was low-pass filtered at 4 kHz, attenuated, and passed to the mixer. The masker ran continuously throughout the experimental session.

Stimuli were presented diotically to observers using TDH-39 headphones mounted in Rudmose Tracor RA-125 headsets with MX-41/AR cushions. Observers sat in a booth in a sound-attenuated chamber, which attenuated airborne sound by 50 dB at 500 Hz and by 65 dB at 2000 Hz (Taylor, 1984). The sinusoidal transients were presented to observers at 62 dB SPL, while the Gaussian noise masker had a spectrum level of 39 dB SPL, resulting in a signal-to-masker ratio of 23 dB.

The frequencies of the tonal transients followed the discrete, overlapping uniform distributions shown in Figure 2.2. The label *SN* refers to the set of high-frequency tones and *N* refers to the set of low-frequency tones. There were 19 separate frequencies altogether, defined in 5 Hz steps over the range 595 Hz to 685 Hz inclusive. The *N* event was associated with the bottom 13 frequencies (595-655 Hz), the *SN* event with the top 13 frequencies (625-685 Hz), and the overlap consisted of the middle 7 frequencies (625-655 Hz).

**Procedure.** Each replication consisted of 416 trials, involving the presentation of 16 identical stimuli per frequency per event, or $16 \times 13 = 208$ stimuli per event. On each replication, stimuli were randomly sampled without replacement, resulting in a haphazard sequence of events and frequencies. A different sequence was used for each replication and for each observer. On each replication, the empirical proportions of tonal frequency for each event exactly matched the theoretical uniform distributions.

Each trial consisted of a warning interval of 100 ms, an observation interval of 250 ms, a decision interval of 1000 ms, and a reset interval of 750 ms. The reset interval was a minimum duration. The next trial could not begin until the slider had been reset to its extreme left. A set of LED lights on the decision panel were switched on and off to mark the trial intervals. No trial-by-trial knowledge of results was given, but observers could later view their single-replication ROC curves at the conclusion of each replication. Only one observer ran at a time, and there was only one session per replication, which took about 14 minutes to complete. All 24 replications were completed over the course of one week.

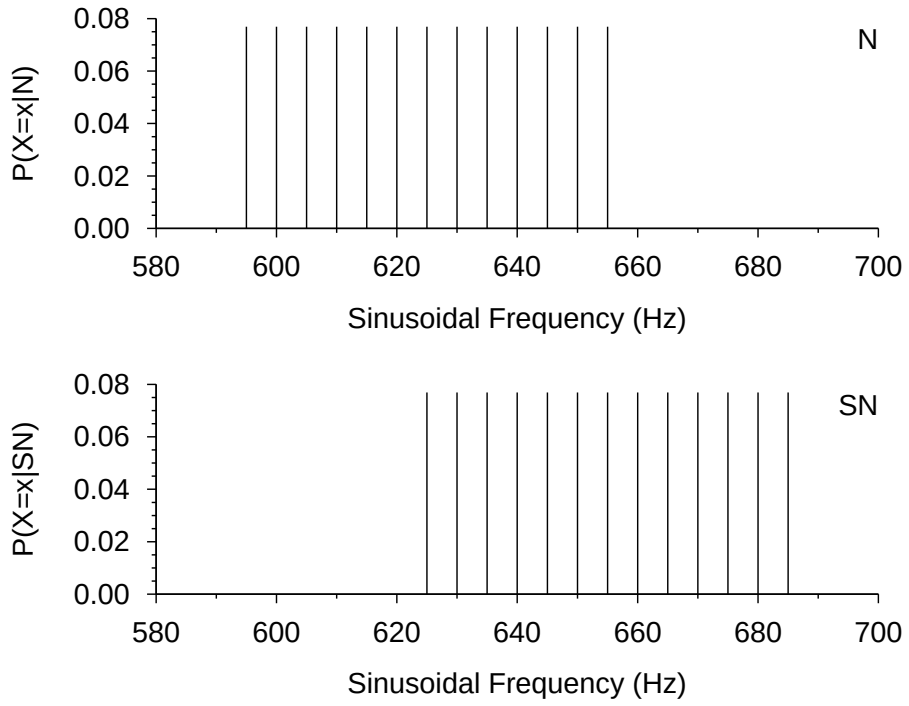


FIGURE 2.2: Probability mass functions of sinusoidal frequency used for the $N$ and $SN$ events (after Taylor et al., 1991, Figure 2(b)).

Observers rated their certainty that the $SN$ event had occurred by using a 12 cm horizontal slider as a continuous rating scale. According to Taylor et al., "The instructions suggested that observers use tonal frequency as the decision variable, and that they map their ratings onto this variable," (Taylor et al., 1991, p. 140). Observers were asked to use all portions of the slider length equally often. The slider was a continuously variable resistor, and the voltage across it was a linear function of slider position. The position was measured electronically at the end of each decision interval, and the slider continuum was partitioned evenly into 36 rating categories.

Unique noise was deliberately introduced into this experiment, *in addition to* any unique noise which observers contributed. Experimentally-introduced unique noise primarily came from the continuous background masker, and possibly from sequential dependencies (since the haphazard trial sequence was different for each replication). The extent of the unique noise was not measured by Taylor et al., since this was not the purpose of their study.

### Results

The distributions of tonal frequencies in Figure 2.2 result in a discrete theoretical ROC curve consisting of points lying along two axes of the ROC space, and along a line that is parallel to and above the chance line. Figure 2.3 shows the theoretical ROC curve and all 24 single-replication ROC curves.

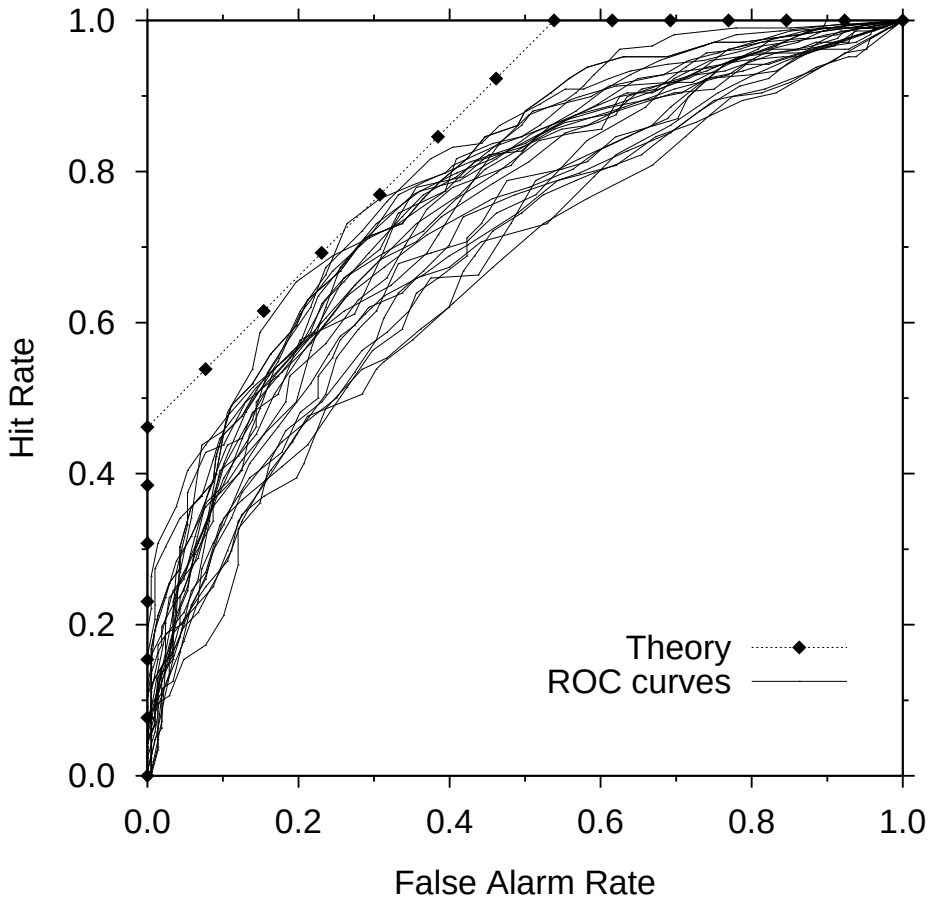


FIGURE 2.3: All 24 single-replication, 36-point rating scale ROC curves, and the theoretical ROC curve.

A large amount of variability was evident across replications, both from within and across observers. (If each observer was entirely consistent, but performance differed across observers, then there would be only three ROC curves evident in Figure 2.3.) The ROC curves were generally irregular, and none of them followed the theoretical curve, including the best ROC curve out of eight from each observer (Taylor et al., 1991, Figure 6). Given any *single* replication, the theoretical interpretation of results is likely to be wrong, because common noise was confounded by unique noise. If the theoretical ROC curve was unknown, it would not be recovered, or even approximated, based on ROC results.

All single-replication values of $\mathcal{A}$ fell short of the theoretical value of 0.8550. The best single-replication value of $\mathcal{A}$ was 0.7901 for Observer 3's eighth replication, and the worst was 0.6612 for Observer 2's eighth replication. Each observer showed a different level of unique noise, which is known to be the case *because* the theoretical performance was known to be the same across observers. The mean and standard deviation of $\mathcal{A}$ were, respectively, 0.7566 and 0.0125 for Observer 1; 0.6893 and 0.0149 for Observer 2; 0.7723 and 0.0150 for Observer 3; and 0.7394 and 0.0392 for all 3 observers together. Although the standard deviations are very similar across observers, Observer 2 demonstrated the largest amount of unique noise, because his performance was the worst compared to the theoretical curve.

Since the theoretical performance was the same across observers, no distinction is made from hereon between replications from different observers for this experiment. Instead, the data is treated as a single set of 24 replications.

## 2.3 Mean ROC curves

One solution to the problem of variablity across replications is to calculate some form of average ROC curve (Macmillan & Kaplan, 1985). There are at least four variants of mean ROC curve, namely the *pooled ROC curve*, the *arithmetic mean ROC curve*, the *z-averaged mean ROC curve*, and the *arcsine-averaged* mean ROC curve. All of these analyses indicate typical or expected single-replication ROC performance, and all reduce ROC variability across replications.

It is shown here that under certain conditions, the *pooled ROC curve* and the *arithmetic mean ROC curve* are identical. It is also shown that there are problems with the arithmetic mean (or pooled) ROC curve because it is subject to floor and ceiling effects, which result in biased estimation of average performance. These problems may be circumvented by calculating the *z-averaged mean ROC curve*, but new problems are introduced, because probabilities of zero or one result in $z$-scores that are infinite. The *arcsine-averaged mean ROC curve* is a compromise between floor and ceiling effects, and setting finite transform values for probabilities of zero or one. Each of these analyses are described in turn, and

the arithmetic mean ROC curve and arcsine-averaged mean ROC curve are applied to Taylor et al.'s continuous rating scale experiment.

Apart from the mean ROC curves shown in this section, *all other mean ROC curves shown in subsequent chapters are arcsin-averaged mean ROC curves.*

### 2.3.1 Pooled ROC curves and arithmetic mean ROC curves

**Pooled ROC curves.** Let $q$ be the number of categories on the rating scale, and let $m$ be the number of replications. Each replication's $q$-point ROC curve derives from its own $2 \times q$ event-decision matrix, like the one in Table 1.2. The initial entries in Table 1.2 for each replication are tallies showing the number of times each event-decision conjunction occurred for that replication. A *pooled ROC curve* is derived by summing tallies across $m$ replications for each event-decision conjunction (i.e. across matrices on a cell-by-cell basis), resulting in a pooled event-decision matrix. If the numbers of stimuli per event are $n_{\mathrm{N}}$ and $n_{\mathrm{SN}}$, then the total numbers of tallies per event in the pooled matrix are $m \times n_{\mathrm{N}}$ and $m \times n_{\mathrm{SN}}$ for the $N$ and $SN$ events respectively. The pooled tallies are converted into proportions by dividing by $m \times n_{\mathrm{N}}$ and $m \times n_{\mathrm{SN}}$, which may then be cumulated in the same way as an ordinary rating scale data set to produce the pooled ROC curve. The pooled ROC curve indicates what a single-replication ROC curve would be like if $m \times n_{\mathrm{N}}$ and $m \times n_{\mathrm{SN}}$ stimuli were used instead of just $n_{\mathrm{N}}$ and $n_{\mathrm{SN}}$ stimuli.

**Arithmetic mean ROC curves.** Section 1.3 described how each replication's ROC curve derives from a $2 \times q$ event-decision matrix. The entries are initially tallies, which are converted into hit rates and false alarm rates by the converting tallies into proportions and cumulating appropriately. The *arithmetic mean ROC curve* is defined by the arithmetic mean hit rate paired with the arithmetic mean false alarm rate, where the averaging occurs across $m$ replications for each given rating category. In terms of the ROC space, there are $m$ ROC points associated with the $k^{th}$ rating category. The arithmetic mean ROC point associated with the $k^{th}$ category is the centroid of contributing ROC points (Macmillan & Kaplan, 1985). Hence, the arithmetic mean ROC curve must, by definition, lie in the middle of the set of single-replication ROC curves.

**The relationship between the pooled ROC curve and the arithmetic mean ROC curve.** The pooled ROC curve is identical to the mean ROC curve based on the arithmetic mean hit and false alarm rates, in a multiple-replication experiment.[16] To show this, assume that a multiple-replication data set has been tallied into event-decision matrices such as Table 1.2, with one matrix per replication. Starting with the tallied data,

[16] Pooled and arithmetic mean ROC curves can also be calculated without running replications based on identical stimuli. The general conditions that need to be satisfied in order for the pooled ROC curve to equal the arithmetic mean ROC curve are that the number of stimuli per event, and the rating scale that is used, are the same for each contributing ROC curve. These conditions are automatically satisfied by a multiple-replication experiment.

the four steps involved in deriving the arithmetic mean ROC curve are: (1) conversion into proportions by dividing by either $n_{\mathrm{N}}$ or $n_{\mathrm{SN}}$, (as appropriate), (2) cumulation of all values greater than or equal to a given rating cutoff, for each event, (3) summing (cell-by-cell) across replications, and (4) dividing the results by the number of replications, $m$. Steps (1) and (2) are first applied, in either order, resulting in a set of single-replication ROC curves. Next, steps (3) and (4) are applied in either order to average the single-replication ROC matrices and obtain the arithmetic mean ROC curve. To derive a pooled ROC curve, step (3) is applied to the raw tallies first, and then steps (1), (2) and (4) could be applied in any order, although (1) and (4) generally go together for the pooled ROC curve. Mean ROC curves could be based on any type of mean, and not just the arithmetic mean. If the type of mean is something other than the arithmetic mean, then the pooled and mean ROC curves are generally not identical.

### 2.3.2 Mean ROC curves based on transform-averaging

Strictly speaking, probabilities and proportions should not simply be averaged using the arithmetic mean, because these quantities are subject to floor and ceiling effects when the values being averaged are close to either zero or one. In such circumstances, samples which lie away from the extremes have disproportionate effects on the mean, and the distribution of probability values is skewed away from zero or from one.[17]

The effect of averaging hit and false alarm rates is to pull the mean ROC curve away from the axes of the ROC space, and consequently to depress related measures of performance. In the context of a Gaussian decision axis, McNicol (1972, p. 111-113) gives a worked example showing how $d'$ based on average hit and false alarm rates can seriously underestimate the average $d'$ value when performance levels are high.

One solution to the problem of floor and ceiling effects is to transform the hit and false alarm rates into $z$-scores of the standard N(0,1) Gaussian random variable, and find the arithmetic mean of the transformed values (Macmillan & Kaplan, 1985). To retrieve the ROC curve in the linear ROC space, the mean $z$-score must be transformed back (from zero to one) onto the original scale. Calculating the $z$-transformed average has well known problems, though, because the $z$-score is undefined whenever a hit or false alarm rate equals either zero or one. At least four solutions have been proposed to this problem (mostly in the context of calculating values of $d'$). These are: (1) setting the $z$-score to equal $\frac{1}{2n}$ or $1 - \frac{1}{2n}$ for probabilities of zero or one, respectively, where $n$ is the number of stimuli per event (Macmillan & Kaplan, 1985; Hautas, 1995); (2) adding a small, non-zero value to each and every category of Table 1.2 for each replication, and adjusting

[17] For example, if a set of false alarm rates (FARs) from 9 replications consists of nine values of 0.0, the mean is 0.0. If one additional replication has a FAR of 0.1, the mean FAR becomes 0.01. If 10 further replications all had FARs of 0.0, then the mean becomes 0.005. This example illustrates that the effect of one value displaced away from zero has a larger effect on the mean than the 10 additional values which are all bounded below at zero. Similar effects apply at the upper bound of one.

the nominal number of stimuli per event appropriately (Hautas, 1995); (3) arbitrarily setting the largest possible absolute $z$-score to a finite value (Isabelle & Colburn, 1991, used $z = 5$, for example); and (4) for each rating criterion, only averaging over ROC points whose hit *and* false alarm rates do not equal either zero or one. These four solutions are noted here, but $z$-averaged mean ROC curves will not be used in subsequent data analyses. The problem of undefined $z$-scores are circumvented by using a sigmoidal transform that is well-defined at probabilities equal to zero or one.

The purpose of using the normal cumulative distribution function to transform probability values is to provide an average probability value which is unbiased by floor and ceiling effects. Stuart Slater (1987, personal communication) suggested this could also be done using a transform based on a strictly monotonically increasing section of the sine function. The transform is sigmoidal, like the normal cumulative distribution function, but there are finite values (at $\pm\frac{\pi}{2}$ radians) for when the proportions (hit or false alarm rates) equal either zero or one. If $p_i$ is the $i^{th}$ proportion from a set of $m$ proportions, then the *arcsin-averaged mean* proportion is

$$\overline{p} \;=\; \frac{1}{2}\left[1 + \sin\left(\frac{1}{m}\sum_{i=1}^{m}\arcsin\left(2p_i - 1\right)\right)\right]. \qquad (2.1)$$

Arcsin-averaging is described in detail in Appendix A, and Equation 2.1 is the amalgamation of Equations A.6 and A.7 in the appendix. There is another alternative to the $z$-transformed average, which is the transform-average based on the function $\arcsin(\sqrt{p_i})$ and its inverse (McNicol, 1972; Macmillan & Kaplan, 1985). It is shown in Appendix A that the transform-average mean value based on $\arcsin(\sqrt{p_i})$ is identical to the mean value given by Equation 2.1.

### 2.3.3 Mean ROC curves for Taylor et al.'s (1991) experiment

The arcsin-averaged mean ROC curve, the arithmetic mean (pooled) ROC curve and the theoretical ROC curve are given[18] in Figure 2.4. It is clear that the average ROC curves lie well below the known theoretical ROC curve for this experiment, and are not even of the same form as the theoretical curve. Expected or average ROC performance is distinctly different from and appreciably worse than theoretical performance because of unique noise. The implications of this result are described in Section 2.4.1, in comparison with the results of GOC analysis.

The arcsin-averaged mean ROC curve is almost the same as the arithmetic mean ROC curve for this data set. The former is consistently higher than the latter, albeit by a very small amount. The areas under the curves are 0.7400 and 0.7349 respectively.

[18] A $z$-averaged mean ROC curve was also calculated, based only on ROC points that lay off the borders of the ROC space. This was the fourth $z$-averaging option. The $z$-averaged mean ROC curve lay *between* the two curves shown in Figure 2.4.

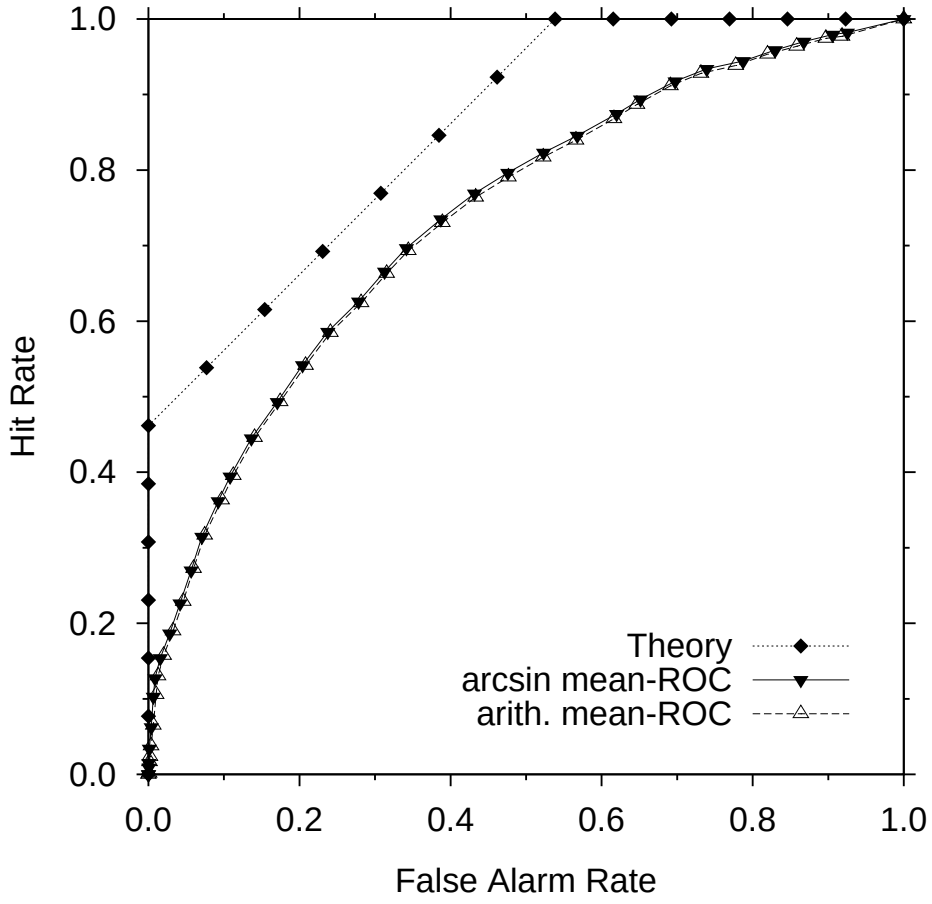


FIGURE 2.4: The 24-replication arcsin-averaged mean ROC curve (upper curve), the 24-replication arithmetic mean ROC curve (lower curve) and the theoretical ROC curve (dashed line). The areas under the curves are 0.7400, 0.7349, and 0.8550, respectively. The arithmetic mean ROC curve is equivalent to the pooled ROC curve.

(In comparison, $\mathcal{A}$ for the theoretical curve was 0.8550, and the average value of $\mathcal{A}$ across single-replication ROC curves was 0.7394.) Arcsine-averaging does provide an improvement in performance compared to arithmetic averaging, but only minutely. The large discrepancy between the mean ROC curves and the theoretical ROC curve is not due to floor or ceiling effects. Transform-averaging is more important if the performance level is much higher (e.g. $\mathcal{A} > 0.99$), because in that case, the contributing single-replication ROC curves would tend to lie mainly along the axes of the ROC space, but not in this case where the curves lie well off the axes.

An ROC curve based on a Gaussian unequal variance model was fitted (by eye) to the arcsin-averaged mean ROC curve in Figure 2.4. The model assumed Gaussian event-conditional distributions with respective means and standard deviations of $\mu_{\mathrm{N}} = 0.0$, $\sigma_{\mathrm{N}} = 1.0$, $\mu_{\mathrm{SN}} = 0.94$, and $\sigma_{\mathrm{SN}} = 1.07$, resulting in a sensitivity measure of $d_z = 0.9077$.

The Gaussian curve provided a very good fit to the data. If plotted on linear normal-deviate scales, the ROC curve is close to a straight line with slope $1.0/1.07 = 0.93$ and intercept $-0.88$ (from Egan, 1975, Equation 3.46).

The various mean ROC curves, including the pooled ROC curve, estimate expected single-replication performance. Although variability is decreased by averaging ROC curves, *unique noise effects are not removed* by averaging ROC curves. If unique noise was removed by mean ROC analysis, then the mean ROC curve should follow the theoretical ROC curve, which it clearly does not.

### Summary of mean ROC curves

The results from a unique-noise-affected multiple-replication experiment may be presented as a set of single-replication ROC curves. A mean ROC curve is a single ROC curve which is derived by averaging ROC points across replications based on the same rating category. Pooling data prior to calculating an ROC curve is also possible. It was shown that the pooled ROC curve and the arithmetic mean ROC curve are identical. Mean ROC curves can be calculated based on transform-averaged hit and false alarm rates. Averaging ROC curves based on Gaussian $z$-scores is desirable, in order to decrease bias due to floor and ceiling effects. However, $z$-averaging suffers from the limitation that the $z$-score is undefined whenever a hit or false alarm rate is either zero or one. A compromise solution to is to use the arcsine transform, which is well-defined when its argument is either zero or one.

Mean ROC curves were calculated for Taylor et al.'s (1991) continuous rating scale experiment. Mean ROC performance was well down from theoretical performance, and was of a different form to the theoretical ROC curve. Mean ROC performance was described in terms of a Gaussian unequal variance model, whereas unique-noise-free performance in the experiment was based on overlapping uniform distributions. Mean ROC analysis reduced variability across replications by the process of averaging ROC curves, but mean ROC analysis did not remove error due to unique noise.

There was minimal difference across different types of mean ROC curves. The arcsine-averaged mean ROC curve was consistently higher than the arithmetic-averaged mean ROC curve, but only slightly. Any floor and ceiling effects for this data set were minimal in comparison to the effects of unique noise.

## 2.4 Group operating characteristic (GOC) analysis

The main topic of this thesis is *group operating characteristic (GOC) analysis*, which is an empirical technique that removes the effects of observer inconsistency. GOC analysis can be applied to multiple-replication experiments, in which all of the replications are based on an identical set of stimuli. GOC analysis results in a *GOC curve*, which is a type of ROC curve calculated from group data. A GOC curve is based on the sum of ratings, or the average rating, taken across replications for the same stimulus. Any TSD measure that is applicable to an ROC curve is also applicable to a GOC curve.

GOC and mean ROC analyses can be performed on the same multiple-replication data set, but the calculations involved are different and the resulting curves in the ROC space are different. This can be seen in the next section, which compares mean ROC with GOC results for Taylor et al.'s (1991) experiment. Following that, Section 2.4.2 describes in detail how to calculate a GOC curve. A historical development of GOC analysis, given in Section 2.5, is best left until last.

### 2.4.1 GOC analysis of Taylor et al.'s (1991) experiment

Group operating characteristic analysis of Taylor et al.'s (1991) continuous rating scale experiment was done using the conventional GOC algorithm, based on the sum of ratings per stimulus, where the ratings were coded as integers from 1 to 36. Figure 2.5 shows the GOC curve based on all 24 replications, in comparison with the (arcsin-averaged) mean ROC curve for the same data (from Figure 2.4), and the theoretical ROC curve. The GOC curve based on 24 replications was in very good agreement with the theoretical curve, especially when compared with the mean ROC curve. Consequently, GOC performance was much better than mean ROC performance. The area under the GOC curve was 0.8483, compared to a theoretical value of 0.8550, whereas the area under the mean ROC curve was 0.7400. The GOC curve did not precisely match the theoretical ROC curve, which indicates there was a small amount of error still present. Adding further replications to the data set would help to remove the remaining error, and achieve an even better approximation to the theoretical curve.

The difference between average ROC performance and theoretical performance in this experiment is due to the effects of observer inconsistency. This is known precisely because the form of the common noise was experimentally controlled as much as possible. Both mean ROC analysis and GOC analysis average out variability across replications. To use Green's (1964) terminology, mean ROC analysis removes variability at a molar level, whereas GOC analysis does so at a molecular level. The removal of variability across replications is not the same as the removal of error in the task. Figure 2.5 demonstrates that although both analyses decrease variability, GOC analysis *removes* the effects of unique noise, whereas mean ROC analysis still *incorporates* these effects.

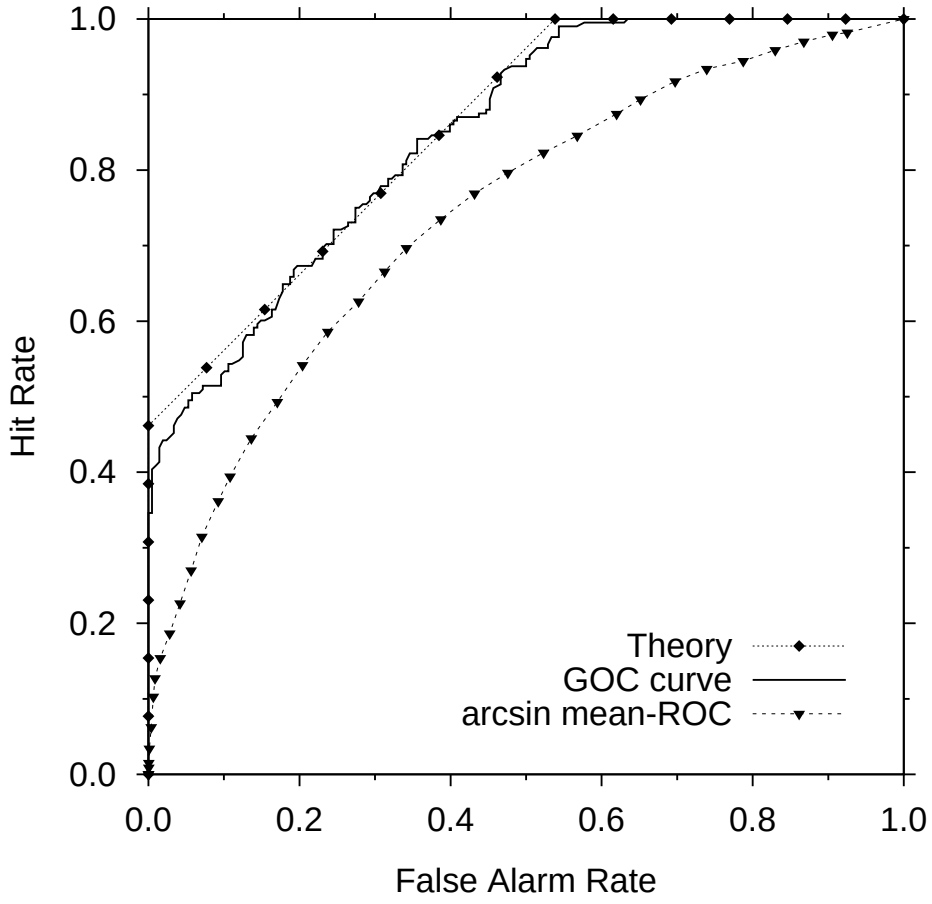


FIGURE 2.5: The theoretical ROC curve, 24-replication GOC curve and 24-replication arcsin-averaged mean ROC curve. The areas under the curves are 0.8550, 0.8483, and 0.7400, respectively (after Taylor et al., 1991, Figure 6(d)).

There are important implications for psychophysical theory that are illustrated by Figure 2.5. The mean ROC curve can be fitted by a theoretical ROC curve based on a Gaussian unequal variance model with $d_z = 0.9077$ (Section 2.3.3), whereas the GOC curve approximates the theoretical uniform model. The results show that *not only can unique noise depress performance, but it can also alter the form of the apparent model.* Much of experimental psychophysics involves single-replication single-presentation experiments, and performance in such experiments is often described by Gaussian models of either equal or unequal variance (e.g. Green & Swets, 1974; Egan et al., 1959; Hanley, 1988). If the experiments are not replicated, erroneous conclusions may be drawn based on ROC analysis. Unless multiple replications are run, the contribution of observer inconsistency to performance is unknown, and the form and level of unique-noise-free performance remains unknown.

### 2.4.2 GOC algorithms

There are two different ways of calculating a GOC curve from a multiple-replication data set, the *conventional algorithm* and the *generalised algorithm*. The conventional algorithm refers to how GOC analysis has been done until now,[19] whereas the generalised algorithm is relatively new to GOC analysis (it was used by Lapsley Miller, 1999, and presumably by Metz and Shen, 1992). Each algorithm reflects a different way of thinking about ROC analysis.

The conventional algorithm consists of calculating a GOC curve using sums-of-ratings, where ratings are summed across replications on a *per stimulus* basis. Once the tallying is complete over all stimuli in a stimulus set, the sum-of-ratings is treated as a new rating scale, and a GOC curve is derived from tallied sum-of-ratings data in the same way that a (single-replication) rating scale ROC curve is derived from tallied rating data. In contrast, the generalised algorithm is based on the sorting or ordering of a stimulus set according to a sorting-key. The *sorting-key* is either the sum of ratings per stimulus, or some type of average rating. If the sum of ratings is used as the sorting-key, then both algorithms result in the same GOC curve.

Ordering a stimulus set to calculate an ROC curve is not new. It would be used, for instance, to calculate an ROC curve for simulated ideal observers that produced continuously-distributed evidence values on a decision axis. A different procedure is generally used to calculate empirical ROC curves, however. Rating scale ROC curves are based on event-decision matrices, such as Table 1.2 in the preceding chapter. The process of calculating an event-decision matrix is in fact a type of ordering or sorting procedure, even though the calculations may not seem to be like that. Similarly in GOC analysis, the conventional algorithm may not be viewed as the ordering of a stimulus set, but that is what the algorithm does. Much emphasis is placed here on stimulus ordering, because the idea is crucial for the theory of GOC analysis given in Chapter 5.

The difference between the two GOC algorithms partly relates to the scaling of a rating scale. Since rating scales are ordinal scales, then the set of numbers that are assigned to the rating categories are, in principle, arbitrary but ordered. Rating scales in ROC analysis typically use successive integers (1,2,3 ... ), which are treated here as part of the conventional algorithm. Arbitrary-valued ordinal scales are treated as part of the generalised algorithm. The scaling of a rating scale is not the crucial distinction between algorithms, but it partly characterises the different approach that underlies each of the two algorithms.

[19](Watson, 1963; Ahumada, 1967; Boven, 1976; McAulay, 1978; Taylor, 1984; Taylor et al., 1991; Whitmore et al., 1993; Galvin et al., 1998; Lapsley Miller et al., 1998)

### The conventional GOC algorithm

Watson (1963) and Taylor et al. (1991) describe the conventional GOC algorithm as follows. Assume an experiment is run consisting of $m$ replications having the same $n$ stimuli presented in each replication. The results can be presented in an $n \times m$ table, like in the centre of Table 2.1 (in the example, $m = 7$ replications and $n = 150$ stimuli). The entries in the table beneath the heading *Replication* are ratings made for each stimulus on each replication. The $i^{th}$ column gives the ratings made on the $i^{th}$ replication, and the $i^{th}$ ROC curve is based on the data in this column. The $j^{th}$ row gives the ratings made for the $j^{th}$ stimulus. (Row order in Table 2.1 does not have to reflect the order of stimulus presentation.) A sum of ratings is calculated for each stimulus, and the event is also noted for each stimulus.

Unique noise is indicated by rating variability within rows and across columns. Common noise is indicated by variability across rows, although it cannot be properly assessed until unique noise has been removed. The two types of error are not taken here to be necessarily independent, although independence is usually assumed in GOC studies (Watson, 1963; Taylor et al., 1991).

The conventional GOC algorithm uses an event-decision matrix like in Table 2.2(a). The table entries are counts or tallies of the number of times that each event-*and*-sum-of-rating conjunction occurs in Table 2.1. For an $m$-replication GOC curve, the sum of ratings data set (second column from the right in Table 2.1) is collated in a $2\times(m(q-1)+1)$ event-decision matrix like in Table 2.2(a).[20]

Table entries in Table 2.2(a) are tallies of the number of times that each event-versus-sum-of-ratings configuration occurs (in Table 2.1). The total number of tallies per event equals the total number of stimuli per event. The entries are converted into relative proportions, in Table 2.2(b), by dividing the entries in each row of Table 2.2(a) by the appropriate number of stimuli per event ($n_{\mathrm{SN}}$ or $n_{\mathrm{N}}$, where $n_{\mathrm{SN}} = 100$ and $n_{\mathrm{N}} = 50$ in the example). A cutoff-based rule is applied to the sum-of-ratings axis and GOC hit and false alarm rates are calculated by successively cumulating relative proportions from the right, as displayed in Table 2.2(c). This is analogous to the calculation of a rating scale ROC curve. For each sum of ratings in Table 2.2(c), the hit rate is the upper fraction and the false alarm rate is the lower fraction. The false alarm and hit rates, paired according to sum-of-ratings value, define the points of the GOC curve. The GOC point (FAR,HR) $= (0, 0)$ (from proportions $\frac{0}{50}$ and $\frac{0}{100}$) is included by convention. (In a binary-decision task this is the equivalent of saying "no" all of the time.) The GOC curve for the example is given in Table 2.2(c), that is, hit rate (the *SN* row) as a function of false alarm rate (the *N* row), paired according to sum-of-rating value.

[20] When ratings are integers in the set {1,2, ... $q$}, then the sum of ratings across $m$ replications lies in the set {$m$,$m+1$, ... $mq$}, meaning there are $mq - m + 1 = m(q-1) + 1$ possible sums of ratings. In the example in Table 2.2(a), there are $63 - 7 + 1 = 57$ possible sums, but not all of them have occurred.

| Stimulus Number | Event | Replication 1 | 2 | ... | 7 | Sum of Ratings | Average Rating |
|---|---|---|---|---|---|---|---|
| 1 | *SN* | 3 | 2 | ... | 9 | 34 | 4.86 |
| 2 | *N* | 1 | 5 | ... | 3 | 43 | 6.14 |
| 3 | *SN* | 4 | 1 | ... | 6 | 42 | 6.00 |
| 4 | *SN* | 7 | 3 | ... | 8 | 52 | 7.43 |
| . | . | . | . | | . | . | . |
| . | . | . | . | | . | . | . |
| . | . | . | . | | . | . | . |
| 150 | *N* | 4 | 3 | ... | 2 | 17 | 2.43 |

TABLE 2.1: Example data table for a GOC experiment, showing the sum of ratings and arithmetic average rating per stimulus. In this example, 7 replications were run using a 9-point rating scale, and there were 150 stimuli in total across both events (after Taylor et al., 1991, Table 1).

(a) Sum of Ratings

| Event | 7 | 8 | 9 | ··· | 58 | 59 | 60 | 61 | 62 | 63 | No. Stimuli |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *SN* | 0 | 1 | 0 | ··· | 2 | 0 | 0 | 1 | 0 | 3 | 100 |
| *N* | 2 | 1 | 0 | ··· | 1 | 0 | 1 | 0 | 0 | 1 | 50 |

(b) Sum of Ratings

| Event | 7 | 8 | 9 | ··· | 58 | 59 | 60 | 61 | 62 | 63 |
|---|---|---|---|---|---|---|---|---|---|---|
| *SN* | $\frac{0}{100}$ | $\frac{1}{100}$ | $\frac{0}{100}$ | ··· | $\frac{2}{100}$ | $\frac{0}{100}$ | $\frac{0}{100}$ | $\frac{1}{100}$ | $\frac{0}{100}$ | $\frac{3}{100}$ |
| *N* | $\frac{2}{50}$ | $\frac{1}{50}$ | $\frac{0}{50}$ | ··· | $\frac{1}{50}$ | $\frac{0}{50}$ | $\frac{1}{50}$ | $\frac{0}{50}$ | $\frac{0}{50}$ | $\frac{1}{50}$ |

(c) Sum of Ratings

| Event | 7 | 8 | 9 | ··· | 58 | 59 | 60 | 61 | 62 | 63 |
|---|---|---|---|---|---|---|---|---|---|---|
| *SN* | $\frac{100}{100}$ | $\frac{100}{100}$ | $\frac{99}{100}$ | ··· | $\frac{6}{100}$ | $\frac{4}{100}$ | $\frac{4}{100}$ | $\frac{4}{100}$ | $\frac{3}{100}$ | $\frac{3}{100}$ |
| *N* | $\frac{50}{50}$ | $\frac{48}{50}$ | $\frac{47}{50}$ | ··· | $\frac{3}{50}$ | $\frac{2}{50}$ | $\frac{2}{50}$ | $\frac{1}{50}$ | $\frac{1}{50}$ | $\frac{1}{50}$ |

TABLE 2.2: Event-decision matrix calculations for GOC analysis of the example data in Table 2.1. In this example, 7 replications were run using a 9-point rating scale, with $n_{\mathrm{SN}} = 100$ *SN* stimuli and $n_{\mathrm{N}} = 50$ *N* stimuli. The matrices show (a) raw tallies, (b) relative proportions, and (c) cumulated proportions (hit and false alarm rates) (after Taylor et al., 1991, Table 1).

There are generally more points on a GOC curve than on any of the single-replication ROC curves, because the sum-of-ratings can take on more possible values than the original ratings do. The number of points on a GOC curve generally increases as more replications are combined. Some GOC points may occur more than once (e.g. for sums of ratings of 59 and 60, in Table 2.2(c)). This happens whenever there are possible sum-of-rating values that do not occur (such as 59, in Table 2.2(a)).

**The generalised GOC algorithm**

The generalised GOC algorithm derives GOC curves from a set of data based on the ordering of stimulus event-labels ($N$ or $SN$) according to some sorting-key that is derived from a data set of ratings. The sorting-key could be the sum of ratings per stimulus, or an average rating, or some other quantity derived from the data set (for example, the maximum rating, or median rating). The particular type of sorting-key is determined by the experimenter, depending on the purposes of the analysis. A wide variety of averages may be used as sorting-keys, as is shown in Chapter 3.

The first step in the generalised algorithm is to calculate the sorting-key value for each stimulus. The second step is to order the stimulus set according to sorting-key value. The next step is to derive the sequence of event-labels associated with the ordered stimulus set. If the sorting-key was the sum of ratings, for example, then the second column from the right in Table 2.1 would be extracted along with the second column from the left, with sorting-keys and event-labels paired together for each stimulus. The stimuli would then be ordered according to sum-of-rating value, and the event-labels would be rearranged according to this ordering. Assume the event-label sequence is written horizontally, like in Table 2.3, with sorting-key value increasing from left to right. A GOC curve is derived by setting cutoffs and counting the number of stimuli for each event that have sorting-keys with values greater than or equal to the cutoff. As the cutoff is systematically moved from the top of the sequence on the right-hand side, to the bottom of the sequence on the left-hand side, cumulative tallies are kept of how many $N$-stimuli and $SN$-stimuli occur to the right of, and including, the cutoff. The event-label sequence may have ties in it whenever two or more stimuli have the same sorting-key value. For each set of ties, the cumulative tally per event incorporates tallies for *all* of the tied stimuli of the same sorting-key value.

Cumulative proportions are obtained by dividing the two cumulative tallies by $n_{\mathrm{SN}}$ and $n_{\mathrm{N}}$ (Table 2.4). This gives the false alarm and hit rates that define the GOC curve. With the inclusion of the point $(0, 0)$, it can be seen that the GOC points in Table 2.4 are the same as those given in Table 2.2(c).

**GOC analysis based on the arithmetic mean rating.** Metz and Shen (1992) performed GOC analysis based on the arithmetic mean rating per stimulus. In order to do this they must have used something like the generalised algorithm to compute what they called *mean-rating ROC curves*. The curve based on the arithmetic mean rating is iden-

| sorting-key | 7 | 7 | 8 | 8 | ... | 58 | 58 | 58 | 60 | 61 | 63 | 63 | 63 | 63 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| event-label | *N* | *N* | *SN* | *N* | ... | *N* | *SN* | *SN* | *N* | *SN* | *SN* | *N* | *SN* | *SN* |

TABLE 2.3: Table for calculating a GOC curve using the generalised GOC algorithm, for the same example data in Tables 2.1 and 2.2. The sorting-key is the sum of ratings, which increases systematically from left to right. Entries have been grouped where there are ties.

| sorting-key | | 7 | 8 | ... | ... | 58 | 60 | 61 | 63 | — |
|---|---|---|---|---|---|---|---|---|---|---|
| cutoff index | | 45 | 44 | 43 | ... | 4 | 3 | 2 | 1 | 0 |
| event- | *SN* | $\frac{100}{100}$ | $\frac{100}{100}$ | $\frac{99}{100}$ | ... | $\frac{6}{100}$ | $\frac{4}{100}$ | $\frac{4}{100}$ | $\frac{3}{100}$ | $\frac{0}{100}$ |
| labels | *N* | $\frac{50}{50}$ | $\frac{48}{50}$ | $\frac{47}{50}$ | ... | $\frac{3}{50}$ | $\frac{2}{50}$ | $\frac{1}{50}$ | $\frac{1}{50}$ | $\frac{0}{50}$ |

TABLE 2.4: Cumulative proportions of stimuli for each event having a sorting-key greater than or equal to successive cutoff values, derived from Table 2.3. The cumulative proportions paired across events according to the same sorting-key cutoffs form the (FAR,HR) pairs that are the coordinates of the GOC curve. The proportions $\frac{0}{50}$ and $\frac{0}{100}$ are included by convention. There were only 45 different sorting-keys values in this example.

tical to that based on the sum of ratings, because the arithmetic mean rating and the sum of ratings are strictly monotonic increasing transforms of each other. This does not change the event-label sequence. If the arithmetic mean rating is used as the sorting-key, then each of the sum of ratings in Tables 2.3 and 2.4 would be divided by the number of replications (which is seven in the example), but the rest of Tables 2.3 and 2.4 remain unchanged.

### A comparison of the conventional and generalised GOC algorithms

There are clear similarities in the two GOC algorithms. If sums of integer-valued ratings are used as the sorting-key in the generalised algorithm, like in Tables 2.3 and 2.4, then the two algorithms produce the same GOC curve. Integer-valued rating scales are convenient, but are not mandatory. The generalised algorithm is necessary in order to perform GOC analysis based on other types of scalings (e.g. square-roots of integers), examples of which are given in Chapter 3.

The generalised algorithm is more efficient than the conventional algorithm whenever Table 2.2 in the conventional algorithmis a sparse matrix. This can occur whenever rating values are irregularly spaced on a rating scale. It can also occur whenever the number of replications or the number of rating categories are large compared to the number of stimuli per replication.

There is no fundamental difference between the two GOC algorithms. Either of can be used to calculate a GOC curve or, for that matter, a single-replication ROC curve. The reason why a distinction has been made between them here is that they reflect different approaches to ROC analysis. Rating scales (especially two-point scales) may not always be seen as the basis for ordering a stimulus set. However, ordering (with ties) is exactly what occurs when an event-decision matrix such as Table 1.2 is collated.

### Summary of GOC analysis

GOC analysis is a form of empirical ROC analysis that combines data across replications on a per stimulus basis. Conventionally, this involves calculating the sum of ratings per stimulus, and collating sums across stimuli. A generalisation of the GOC algorithm was described, in which a GOC curve is based on the ordering a stimulus set according to a sorting-key, such as a mean rating or a sum of ratings. Using a data set from Taylor et al. (1991), it was shown that GOC analysis can remove unique noise, and improve performance in a discrimination task, even when an experimental data set is heavily affected by unique noise. GOC analysis almost recovered the known theoretical ROC curve, whereas mean ROC analysis did not. A comparison of the mean ROC curve and the GOC curve showed that unique noise effects not only depress performance, but can also alter the form of an inferred theoretical curve. Much of experimental psychophysics uses single-replication analyses, which are very likely to incorporate unique noise effects. In spite of its potential, GOC analysis has been underutilised in psychophysics for several decades now.

## 2.5 Historical development of GOC analysis

In a seminal monograph in aural psychophysics, Smith and Wilson (1953) ran several SIFC, tone-in-noise detection experiments. They investigated the effects of signal-to-noise ratio, criterion and size of rating scale on detectability in a study that involved over 100 observers. Smith and Wilson investigated how group performance from multiple observers could be used to improve performance over that of individual observers. Groups of five observers ran simultaneously, and the number of observers ($n$) that said *"yes"* they had heard the tone was recorded on each trial. Smith and Wilson presented a family of classical psychometric functions, but based on group performance, with $n$ held as a parameter. The functions showed the proportion of trials in which $n$ out of five observers had said *yes*, as a function of signal-to-noise ratio. On any trial, a group decision could be taken as a *yes* decision if $n$-or-more observers individually reported *yes*. Smith and Wilson's results showed how, at each given signal-to-noise ratio, the lower the number $n$, the more likely it was that the group would decide *yes*. Conversely, the higher the number $n$, the less likely it was that the group would decide *yes*.[21] Reinterpreted in the context of GOC analysis,

[21] For example, if $n = 1$, only one observer would need to decide *yes* and the group decision would be *yes*. If $n = 5$, all five observers would have to say *yes* in order for the group decision to be *yes*.

each observer used a binary-decision scale, where a *yes* and *no* decisions may be coded as one and zero respectively. The number of observers saying *yes* then becomes a sum of ratings, and the value $n$ is a cutoff on the sum-of-ratings decision axis. Framed in modern terms, Smith and Wilson's results showed how, at each given signal-to-noise ratio, GOC hit rate increased as the sum-of-ratings cutoff ($n$) decreased. Smith and Wilson also ran catch-trials (noise-alone trials), and it is possible that some of their data points could be re-analysed and converted into GOC curves. Smith and Wilson (1953) anticipated TSD with their experimental and theoretical analysis. They also anticipated GOC analysis.

Swets et al. (1959) presented the first major theory of multiple observations within the context of TSD. The theory was based on additive, Gaussian internal and external noise, where the effect of the internal noise was a decrease in detectability, measured in $d'$. They ran several multiple-presentation experiments in order to test their theory, and showed that performance increased substantially with repeated stimulus observation. Although Swets et al.'s methodology was not the same as for GOC analysis,[22] their theory is relevant to GOC analysis, because the mechanisms for the removal of extra noise are presumably similar. Swets et al. assumed that internal noise is removed internally as the observer averages over multiple observations, each of which is affected by internal noise. In GOC analysis, the averaging of ratings occurs externally to any observer. Swets et al.'s (1959) study set the scene for subsequent experimentation and theories relating to internal noise and unique noise. (Watson, 1963; Green & Swets, 1974; Boven, 1976; Siegel, 1979; Taylor, 1984; Berg, 1987, 1989, 1990; Taylor et al., 1991).

The first publication describing GOC analysis within the context of TSD was Watson (1963), who used it in his doctoral thesis as a way of deriving a detection statistic for a group of observers. (He also coined the term *group operating characteristic analysis.*) In combining four-point rating scale data from three observers, Watson found that the $d_s$ value from the GOC curve he presented was 1.51, whereas the mean value of $d_s$, averaged over the three individual observers, was only 1.23. This was further evidence that sums of ratings, and GOC analysis in particular, could improve the detectability of a group of observers compared to that for just an average single observer. He also derived a model of an observer based on unique noise that was additive with common noise on a decision axis. Although the result was the same as in Swets et al. (1959), the derivation was more general. Watson derived his model by considering the effect of additive noise in terms of variances of noise components. The theory was derived without assuming any particular distributional form, although some form was needed to tie the theory to data. Gaussian distributions were assumed, since $d_s$ was used as a sensitivity measure. Watson (1963) presented a dice game analogy to GOC analysis, which is described further in Chapter 5.

[22] In a multiple-presentation experiment (p. 22), a stimulus is presented multiple times per trial, prior to a single decision, whereas in a multiple-replication experiment, used for GOC analysis, a stimulus is presented over multiple trials, with a separate decision per trial.

The second study to use GOC analysis was by Ahumada (1967). In his doctoral thesis, he derived GOC curves based on SIFC binary-decision data in order to estimate the critical bandwidth. Four observers ran 5 replications each of a tone-in-noise detection experiment. He found that an individual's GOC performance was better than their pooled, or average, single-replication performance for all observers. GOC performance based on all 20 replications (across both observers and replications) was much better than the group's single-replication performance; $d_s = 2.32$ for the GOC curve compared to only 1.26 for pooled, single-replication data. He does not say where the idea of using sums of ratings in experimental ROC analysis came from, but it is quite likely to have come from Watson (1963), because Ahumada followed up Watson's (1963) work on energy detectors.

Ahumada and Lovell (1971) and Ahumada et al. (1975) also ran SIFC tone-in-noise detection experiments. Both studies involved multiple-replication, 4-point rating scale experiments with reproducible stimuli. They used the sums of ratings in order to reduce unique noise, and compared the sum per stimulus waveform with the results of a filter-bank model based on waveform measurements. The purpose of these studies was to investigate the frequency characteristics of human hearing, rather than GOC analysis, and no GOC curves were presented. Summation of ratings was done only across replications within observers, but not across observers. Ahumada et al. (1975) reported $d'$ values that were equivalent to results based on the pooled (or arithmetic mean) ROC curve, and on the 8-replication GOC curve for each observer. The rating scale was partitioned into two halves (for *yes* and *no* decisions) in order to calculate $d'$ values. By going from 1 to 8 replications, $d'$ values increased appreciably for all observers—the average $d'$ across observers improved from 1.95 for ROC performance to 2.56 for average 8-replication GOC performance.

Bell and Nixon (1971) used arithmetic mean ratings in an attempt to decrease observer inconsistency in a multiple-replication tone-in-noise signal detection experiment. Their use of average rating stemmed from Ahumada's (1967) comparison of averaged (summed) ratings for each observer with the output of an electronic detector. Bell and Nixon were mainly concerned with inter-rater and intra-rater reliability, however, and did not present any GOC curves.

After Bell and Nixon (1971), Ahumada and Lovell (1971) and Ahumada et al. (1975), the use GOC analysis in psychoacoustics essentially stopped in the United States. There was a series of quasi-molecular studies[23] that developed independently of GOC analysis itself, but which shared some influences such as Swets et al. (1959) and Watson (1963), and also shared some aspects of experimental design. GOC analysis was developed independently in the medical literature, stemming from practical problems in diagnostics (Yerushalmy, 1969; Metz & Shen, 1992). GOC analysis also continued in psychoacoustics in New Zealand.[24] Each topic area is described separately.

[23](Green, 1964; Pfafflin & Mathews, 1966; Pfafflin, 1968; Siegel, 1979; Gilkey, 1981; Gilkey et al., 1985; Siegel & Colburn, 1989; Isabelle & Colburn, 1991)

[24](Boven, 1976; McAulay, 1978; Taylor, 1984; Galvin et al., 1998; Lapsley Miller et al., 1998; Lapsley Miller, 1999)

### Quasi-molecular studies

The idea of collating decisions over multiple presentations of the same individual stimulus was discussed in a study on observer inconsistency by Green (1964). Green acknowledged Watson's (1963) thesis and GOC analysis, but did not develop the topic.[25] Green's (1964) focus was on the problem of trial-by-trial prediction of an observer's decisions, based on the characteristics of individual stimuli. He suggested a type of experiment in which each stimulus from a small set of reproducible stimuli is repeatedly presented over the course of many trials. Some stimuli have a reproducible signal added, and observers are asked to perform in a conventional discrimination task. Decisions vary from trial-to-trial for identical stimuli, presumably due to internal noise, so the "most-common" response per stimulus could be calculated as a means of reducing the effects of inconsistency. Green (1964) suggested that the "most-common" response for each stimulus could be related back to the characteristics of individual stimuli. Green (1964) described this approach as *quasi-molecular*, to incorporate the idea that analysis would be done on a per-stimulus basis, but that multiple trials were still required.

A number of studies have followed up on Green's (1964) quasi-molecular approach (Siegel, 1979; Gilkey, 1981; Gilkey et al., 1985; Siegel & Colburn, 1989; Isabelle & Colburn, 1991). These studies involved (mostly SIFC) detection task experiments in which sets of 10 to 25 reproducible maskers were each presented to an observer from 50 to 200 times each, both with and without a signal added. Stimuli were intermixed within experimental sessions to minimise learning of particular waveforms. Results from these experiments may be shown as points in the ROC space, where each point represents the hit and false alarm rate pair associated with an individual masker (with and without a signal added). The hit rate, or false alarm rate, is the proportion of *yes* decisions per stimulus. If each decision is coded as ratings, with zero for *no*, and one for *yes*, then the proportion of *yes* decisions represents the average rating, per stimulus. Under this coding scheme, each stimulus-pair ROC point is formally a GOC curve, albeit based on a stimulus set consisting of a single stimulus per event. If the handful of stimuli used in these experiments were analysed as one larger stimulus set, the result would be conventional GOC analysis.

Single stimulus-pair ROC points mostly lie above the chance line, which reflects the contribution of the signal to detectability. The ROC points are usually spread throughout the triangle above the chance line in the ROC space (e.g. Gilkey et al., 1985, Figure 3). The variability across stimulus-pair ROC curves theoretically reflects individual maskers contributing different evidence values on a decision axis. The fact that the hit and false alarm rates are not either zero or one reflects the influence of unique noise. If there was no unique noise, then all decisions for a given stimulus would be the same—either always *yes*, or always *no*. Consequently, the proportion of *yes* decisions (i.e. hit or false alarm

[25] Probably because Watson's thesis was quite new when Green's paper was written.

rate) would either be zero or one. Single stimulus-pair ROC analysis is not used in this thesis, but the topic is revisited in Chapter 5, where quasi-molecular studies are discussed with respect to the theory of GOC analysis.

### GOC analysis in medical diagnostic tasks

Major growth in the development and application of TSD has occurred in the area of medical diagnostics. ROC analysis has been applied to practical problems such as detection of tuberculosis or tumours (Yerushalmy, 1969; Metz & Shen, 1992). Problems of observer inconsistency have also been encountered in diagnostic tasks, and similar solutions have been proposed.

Yerushalmy (1969) reported results from a number of studies in which multiple readers (observers), who were professional radiologists, attempted to detect pulmonary lesions based on x-ray films. In some of the studies, readers would view the same set of x-rays twice each. A high degree of inconsistency was found, both within and across readers. For example, one study found that

> In judging a pair of x-ray films for evidence of progression, regression, or stability of disease, two readers are likely to disagree with each other in one third of the cases, and a single reader is likely to disagree with himself in about one fifth of the film pairs. (Yerushalmy, 1969, p. 390)

Yerushalmy did not develop GOC analysis *per se*, but all of the elements were present in the analyses of the studies, including hit and false alarm rates based on decisions that were combined either across or within observers. Yerushalmy's (1969) study showed that the problem of observer inconsistency, found in sensory psychophysics, is also a substantial problem in medical diagnostics.

Similar problems lead to similar solutions. GOC analysis was developed in full by Metz and Shen (1992), semi-independently of its development in psychophysics.[26] They described "mean-rating ROC analysis" in multiple replication tasks, which is the same as GOC analysis based on the arithmetic mean rating per stimulus. Metz and Shen developed a theory of observer inconsistency based on additive Gaussian unique and common noise. The theory made a distinction between unique noise variability within readers and across readers. For given parameters of each type of unique noise, the theory predicted how the GOC curve would change, and how performance would improve, as a function of replications added.

Like Yerushalmy (1969), Metz and Shen (1992) were interested in assessing the improvement in performance that was possible in a practical situation. Metz and Shen analysed a multiple-replication data set for a mammographic diagnostic task, and found

[26] Metz and Shen (1992) were aware of the material on multiple-observations in Green and Swets (1974), which included details of Swets et al.'s (1959) study, but not Watson's (1963) development of GOC analysis.

that performance could be improved appreciably by combining data both within readers, and across readers. As part of their analysis, they calculated GOC curves and estimated theoretical parameters based on all possible pairs of replications from a set of six replications. This is a step towards *all combinations analysis*, an extension of GOC analysis which is fundamental to Part II of this thesis. Metz and Shen (1992, Figure 4) presented theoretical GOC curves based on their model, where the parameters were derived from the mammographic data set. The improvement appears relatively small in the ROC space compared to the gains from GOC analysis in psychoacoustical discrimination tasks. Nevertheless, for a fixed false alarm rate of 0.1, the hit rate improved in value from 0.63 to 0.74 just by combining data from four replications, one replication per reader. This shows the promise of GOC analysis in medical diagnostics, because an improvement in detection of even a few percentage points may have very significant consequences for the treatment of disease in individual cases.

### GOC analysis in New Zealand

GOC analysis has been developed extensively at the Psychophysics Laboratory at Victoria University of Wellington, in New Zealand. A series of theses (Boven, 1976; McAulay, 1978; Taylor, 1984; Lapsley Miller, 1999), and experimental projects (Galvin et al., 1998; Lapsley Miller et al., 1998) were supervised by John Whitmore. GOC analysis was also used in a number of postgraduate projects, which are not reported here. Whitmore was introduced to GOC analysis in 1966 at the University of Texas by Lloyd Jeffress and Sandy Gaston,[27] who ran a number of unpublished GOC experiments on aural amplitude discrimination. Jeffress and Gaston were originally introduced to GOC analysis through Charles Watson.[28]

GOC analysis was a major focus of Boven's (1976) and Taylor's (1984) theses, which together formed the basis of Taylor et al. (1991). The main aim of these studies was to evaluate and demonstrate the effectiveness of GOC analysis, which was done through a series of multiple-replication experiments involving aural frequency discrimination of tonal transients. Boven (1976) ran experiments using groups of human observers, whereas Taylor (1984) ran pigeons, and also applied GOC analysis to electronic and computer simulations of unique-noise-affected observers.

A common idea in these studies was the use of *known*, discrete, underlying distributions of tonal frequency that were invented by the experimenters. One example of this was Taylor et al.'s (1991) continuous rating scale experiment, described in Section 2.2. Using known distributions meant the theoretical ROC curve was known for each experiment.

[27] Gaston may have been the first to implement GOC analysis on a computer, which is an indispensible tool for any sizable GOC analysis, especially all combinations analysis.

[28] Whitmore, 1999, personal communication.

Generally, a background masker was used in the experiments with human observers in order to deliberately introduce extra unique noise, whereas criterion variability was introduced into simulations. Having inconsistent observers and known theoretical performance allowed unambiguous evaluation of the effectiveness of GOC analysis, because the theory represented unique-noise-free performance. Boven (1976), Taylor (1984), and Taylor et al. (1991) found that GOC curves generally showed better performance than single-replication ROC curves, and that GOC analysis was effective in not just improving performance, but also in *recovering* the theoretical ROC curve from data that was heavily affected by unique noise (like the results shown in Figure 2.5).

A variety of discrete theoretical distributions were used in different experiments and simulations by Boven (1976), Taylor (1984), and Taylor et al. (1991), including bell-shaped, bimodal, trimodal, triangular, and uniform distributions. Many of these distributions were sub-optimal, in that conversion to likelihood ratio would have improved performance. However, the human and pigeon observers were trained to used tonal frequency as the basis for decisions, rather than likelihood ratio, and GOC analysis generally recovered whatever ROC curves were consistent with the distributions of tonal frequency. Later experiments by Galvin et al. (1998) showed that GOC analysis could recover theoretical curves based on likelihood ratio if observers were trained to use likelihood ratio.

**Boven (1976).** Boven was the first to describe the concepts of unique and common noise, and to discuss the distinction between them and the concepts of internal and external noise.[29] He also extended Swets et al. (1959) derivations of unique-noise-affected performance, based on Gaussian common noise distributions of equal variance, to include Gaussian distributions of unequal variance. Boven showed that Swets et al.'s (1959) formula for estimating $k$, the ratio of unique and common noise variances, was unaffected if $d_z$ was used as a measure of sensitivity instead of $d'$. Boven (1976) ran multiple-observer frequency discrimination experiments, some of which were also published in Taylor et al. (1991).

**McAulay (1978).** McAulay investigated frequency discrimination of tonal signals using GOC analysis in a series of multiple-observer experiments, and compared the results with electronic simulations. She made use of GOC analysis, based on Boven (1976), but it was not the focus of her work. McAulay's experimental design did not involve predetermined distributions of frequency. Her experiments had two observation intervals per trial. A tone with a fixed standard frequency was always presented in the first interval, while a comparison tone of either the same or higher frequency was presented in the second interval. The results were analysed in terms of an SIFC task. Five observers ran one

[29] Earlier development of the concepts underpinning GOC analysis were influenced by analogies to signal-averaging, which can recover a signal pattern that repeatedly occurs in the presence of independent samples of noise (Whitmore, 1999, personal communication). Signal-averaging is often used to study evoked-potentials in sensory physiology (Regan, 1972).

replication each, where decisions were indicated on an 8-point rating scale. McAulay found variability in the ROC curves across observers, with $\mathcal{A}$ ranging over as much as 0.2 for the same experiment. In each experiment the area under the GOC curve was generally as good as, or better than, the area under the best individual's ROC curve. For two experimental conditions, McAulay presented a graph showing how the logarithm of $d'$, based on a GOC curve, increased with the logarithm of the number of replications contributing to the GOC curve. These may be the first examples in the literature of *functions of replications added* (FORAs) based on GOC analysis. McAulay partitioned her data set into blocks of 100 trials per block, and combined successive blocks to calculate her FORAs. She commented on the large amount of unique noise in her experimentation and concluded "If it is necessary to remove all the unique noise from human frequency discrimination data, many replications would be needed, or else a mathematical model of predicting the asymptotic level must be developed," (McAulay, 1978, p. 90). Such a model has now been developed, and is presented in Part II.

**Taylor (1984).** Taylor investigated the effects of unique noise on frequency discrimination in pigeons. He ran a series of multiple-replication, SIFC, binary-decision experiments, in which pigeons were trained to discriminate between two sets of tonal frequencies. The distribution of frequencies followed known, discrete distributions in some experiments, whereas only two tonal frequencies were used in other experiments. Taylor generally found the pigeon data very noisy, with estimates of the unique-to-common noise variance ratio, $k$, ranging from 0.5 to possibly unbounded values, depending on the individual bird and experiment. Detailed computer simulations were run for the pigeon experiments, in which a simulated observer had several layers of possible unique noise sources, including extra input noise, filter jitter, inattention and criterion variability. Taylor found that the level of extra input noise and filter jitter required to account for the pigeon results were unreasonably large, but that large but plausible levels of inattention and criterion variability were consistent with the data.

One of the practical problems that Taylor addressed in the course of his investigation was the variability inherent in combining successive replications in GOC analysis. GOC performance improves as more replications are added. Given a data set of multiple replications, there are many possible sequences of replications that may be generated, each of which results in a different FORA. The set of all possible FORAs shows a high degree of variability because of all the possible ways of combining the data. Taylor solved this problem in his development of *all combinations analysis* (ACA), where a GOC curve is calculated for each possible combination of replications, and average performance is calculated for each number of replications combined. This results in a FORA that is much more stable than the type of FORA calculated by adding successive single replications in GOC analysis. Taylor applied ACA and calculated FORAs for his pigeon experiments and

simulations. The FORAs either indicated that the pigeons' unique-noise-free performance was possibly unbounded, or if bounded, that many replications would be needed to remove the bulk of the unique noise. The material on FORAs developed in Part II of this thesis follows from Taylor's (1984) development of ACA.

**GOC analysis of Type II tasks.** Galvin et al. (1998) developed a theory of Type II ROC analysis, and ran a series of demonstration experiments involving GOC analysis. A Type I task is an SIFC task where an observer makes a decision about whether the $SN$ or $N$ event occurred during a trial. A Type II task requires an observer to decide whether or not their Type I decision was correct or incorrect (Clarke, Birdsall, & Tanner, 1959; Podd, 1975; Galvin, 1988). Galvin et al. (1998) showed that it is possible to use GOC analysis to recover known theoretical Type II ROC curves from a unique-noise-affected data set. The nature of the Type II task is such that ROC curves may lie well above the chance line, cross through the chance line, or lie well under the chance line, depending on an observer's decision axis and criterion in the Type I task. The experiments covered a range of possible Type II ROC curves, and covered both optimal and sub-optimal decision axes. GOC analysis provided better indications of theoretical performance than mean ROC analysis, even for conditions associated with very unusual ROC performance.

**Whitmore et al. (1993).** Whitmore et al. is a previously unpublished study involving 100 replications of an SIFC amplitude discrimination experiment, with 75 replications from one observer and 25 replications from another observer. As well as GOC, ACA, and FORA analyses, such a large data set allowed sample statistics of asymptotic FORA performance to be estimated. This is the topic of Chapter 7, and further details of Whitmore et al.'s (1993) study are available there.

**Lapsley Miller et al. (1998).** A series of unpublished experiments were run by Lapsley Miller et al. (1998), in a theoretical and experimental study of the relationship $\mathcal{A}_{\text{SIFC}} = P(\text{C})_{\text{2IFC}}$. One set of SIFC and 2IFC experiments involved frequency discrimination, based on known discrete theoretical distributions. Another set of experiments involved amplitude discrimination, for which the theory was not known. There was a substantial amount of unique noise in the data, which made any empirical relationship between $\mathcal{A}_{\text{SIFC}}$ and $P(\text{C})_{\text{2IFC}}$ ambiguous. GOC analysis was used to remove unique noise and, in the discrete case, known theoretical performance could be recovered almost exactly. Two of the experiments from Lapsley Miller et al. (1998) were analysed with ACA, and the results are given in Chapter 8. Further experimental details are available there.

**Lapsley Miller (1999).** In her doctoral thesis, Lapsley Miller (1999) investigated the role of bandwidth and duration in aural amplitude discrimination. Evaluation of these parameters of the auditory system is hampered by unique noise, where extraneous, possibly non-sensory processes affect performance in detection experiments. It was therefore desirable to remove the influence of unique noise, to obtain unambiguous measurement of performance. Lapsley Miller extensively used GOC, ACA and FORA analyses in experiments based on a wide range of stimulus parameters. Asymptotic, unique-noise-free performance was compared with a variety of simulations of amplitude discrimination, allowing unique-noise-free estimates of temporal and spectral limitations in human hearing.

With the large number of experimental conditions and signal-to-noise ratios used in the experiments, Lapsley Miller's (1999) project calculated more FORAs than in all of the other studies combined. One of the more remarkable results was that the observer with the best average single-replication performance did not necessarily have the best asymptotic performance, and that this pattern occurred almost one quarter of the time. Further experimental details and a summary of FORA results from Lapsley Miller (1999) are given in Section 8.4.

# Chapter 3

# Transform-average GOC analysis

In the preceding chapter, GOC analysis was based on the sum of ratings, taken across replications, where the ratings were integer-valued. In this chapter, GOC analysis is generalised to encompass transform-average mean ratings defined on arbitrarily-scaled ordinal rating scales.[1] The generalisation involves: (1) the type of average that is used to calculate mean ratings, and (2) the scaling of a rating scale. While these are ostensibly different ideas, it is shown that there is no practical distinction between them in GOC analysis—any difference is one of interpretation, and not of form. The generalisation developed here is called *transform-average GOC analysis*.

The preceding chapter showed that a GOC curve based on a sum of ratings is identical to a GOC curve based on an arithmetic mean rating. The arithmetic mean is not the only possible type of average. Other well known averages include the geometric mean and the harmonic mean. These, and other means, can be described within the same general framework, and are specific cases of a generalised mean, or transform-average. Given a transform function, the transform-average mean rating per stimulus forms the basis of a GOC curve.[2] It is shown in this chapter that transform-average GOC analysis is equivalent to conventional (sum-of-ratings) GOC analysis based on ratings that have been rescaled by an order-preserving transform of the rating scale. The choice of transform-average is synonymous with the choice of rating scale. A numerical example is given, which shows that the choice of scaling (or transform) in GOC analysis affects the ordering of stimuli on a rating scale, and hence, the resulting GOC curve. This is very different from single-replication ROC analysis, which is unaffected by the ordinal rescaling of a rating scale.

[1] This includes binary-decision (two-point) rating scales.

[2] It is assumed throughout this chapter that the generalised GOC algorithm is used to calculate GOC curves.

**Overview of chapter**

Section 3.1 describes transform-averages, and how they can be applied to GOC analysis. Section 3.2 presents more than two dozen different transform-average GOC curves for Taylor et al.'s (1991) continuous rating scale experiment. GOC analysis based on weighted sums of ratings is also investigated. Each transform that is used produces a different GOC curve, and some transforms are more effective than others. Possible reasons for discrepancies between transform-average GOC curves and the theoretical ROC curve are discussed in Section 3.3.

## 3.1 Generalised means in GOC analysis

A generalised, transform-average mean rating, calculated from a multiple-replication data set, is of the form

$$\overline{r}_j \quad = \quad g^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} g(r_{ji})\right) \tag{3.1}$$

where $g$ is a strictly monotonic transform, $r_{ji}$ is the rating for the $j^{th}$ stimulus in the $i^{th}$ replication, $m$ is the number of replications and $\overline{r}_j$ is the transform-average mean rating for the $j^{th}$ stimulus. Special cases of the generalised mean[3] includes the arithmetic mean $[g(r) = r]$, the geometric mean $[g(r) = \log_a(r), a > 0, a \neq 1, r > 0]$, the family of exponential means $[g(r) = b^r, b > 0, r \neq 0]$, and the family of power means $[g(r) = r^c, c \neq 0]$ including the harmonic mean ($c = -1$, so $g(r) = \frac{1}{r}$). For a given transform, $g$, the transform-average GOC curve can be obtained from the ordering of the stimulus set according to transform-average mean values, $\overline{r}_j$, using the generalised GOC algorithm (Section 2.4.2).

In describing transforms of ratings, it is useful to refer to the variety of possible scales in terms of some initial or original rating scale. This scale is defined by the domain of $r_{ji}$, and is typically a set positive integers, although it need not be. There is nothing special about an original rating scale, since it only reflects how an experimenter chose to initially code ratings.

If $g$ is a strictly monotonic *decreasing* function applied to a set of ratings using Equation 3.1, then $g$ can be replaced by its strictly monotonic *increasing* counterpart, $-g$, without changing the resulting mean value.[4] For reasons given in the next section, it is

[3] A comprehensive bibliography on generalised means is given by Norris (1976). Discussions and developments on the topic are found in Cargo (1965), Cargo and Shisha (1969), and Mantel (1969).

[4] To show this, let $R$ represent the entire domain of a rating scale (with specific values denoted $r$), and let $g$ be any continuous strictly monotonic *decreasing* function defined over all of $R$. Let $\gamma = -g$ be the strictly monotonic *increasing* counterpart of $g$, also defined over all of $R$. Since $g$ and $\gamma$ are strictly monotonic, they have unique inverse functions, $g^{-1}$ and $\gamma^{-1}$, respectively. For any specific $r \in R$, let $s = g(r)$, from which it follows that $-s = \gamma(r)$. If $g$ maps $R$ onto the range $s \in [\alpha, \beta]$, then $\gamma = -g$

convenient to only consider increasing transforms. Hence, any decreasing transform presented here will implicitly refer to its increasing counterpart. For example, the harmonic mean would usually be defined in terms of the decreasing transform $g(r) = \frac{1}{r}$, but here it will implicitly refer to the transform $g(r) = -\frac{1}{r}$, which is an increasing function.

### 3.1.1 The three-way equivalence

Given an original rating scale, then $g(r_{ji})$ in Equation 3.1 can be interpreted as a rating on a transformed rating scale. The original ratings, $r_{ji}$, fall on one rating scale with one particular scaling, and $g(r_{ji})$ falls on a new rating scale with a new scaling. The quantity $\sum_{i=1}^{m} g(r_{ji})$ is the *sum of ratings* defined on the rescaled rating scale, and $\frac{1}{m}\sum_{i=1}^{m} g(r_{ji})$ is the *arithmetic mean rating* on the rescaled rating scale. The three quantities:

*(A):* $\sum_{i=1}^{m} g(r_{ji})$,

*(B):* $\frac{1}{m}\sum_{i=1}^{m} g(r_{ji})$, and

*(C):* $g^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} g(r_{ji})\right)$,

are all strictly monotonic increasing (s.m.i.) with each other. *(B)* is s.m.i. with *(A)* because the function $\frac{1}{m}x$ is (linearly) s.m.i. with $x$. *(C)* is s.m.i. with *(B)* because the inverse transform, $g^{-1}$, is s.m.i. (since $g$ is s.m.i.). From these relationships, it follows that *(C)* is s.m.i. with *(A)*, since strict monotonicity is transitive. If a stimulus set is ordered according to any of these three quantities, the ordering is identical across quantities. This is called *the three-way equivalence.*

Given a transform, $g$, the three-way equivalence implies that the GOC curves based each of *(A)*, *(B)* or *(C)* are identical to each other, because GOC curves are based on the ordering of a stimulus set, and the orderings are identical.[5]

Just as any s.m.i. transform of a decision axis does not affect the resulting theoretical ROC curve (Egan, 1975), then any s.m.i. transform of the *sum* of transformed ratings does not affect the resulting GOC curve. What may seem like an unusual form of GOC analysis (based on transform-average mean ratings) can be interpreted in terms of sum-of-ratings GOC analysis based on a rating scale with a different scaling—the results of

maps $R$ onto the range $-s \in [-\beta, -\alpha]$. For any $s \in [\alpha, \beta]$, $s = g(r) = g(g^{-1}(s))$, and so $-s = -g(g^{-1}(s))$. Furthermore, $-s = \gamma(r) = \gamma(\gamma^{-1}(-s))$, which implies that $-g(g^{-1}(s)) = \gamma(\gamma^{-1}(-s))$. Since the outer functions of this equality ($-g$ and $\gamma$) are identical and are one-to-one functions, then their arguments are equal, which implies that $g^{-1}(s) = \gamma^{-1}(-s)$.

Equation 3.1 can now be re-expressed entirely in terms of $\gamma$ and $\gamma^{-1}$. Substituting $-\gamma(r)$ for $g(r)$ and $\gamma^{-1}(-s)$ for $g^{-1}(s)$, Equation 3.1 implies that $\overline{r}_j = \gamma^{-1}\left(-\frac{1}{m}\sum_{i=1}^{m}[-\gamma(r_{ji})]\right)$. The innermost negative sign can be carried through the arithmetic mean and cancels with the outer negative sign (within the argument of $\gamma^{-1}$) leaving $\overline{r}_j = \gamma^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}[\gamma(r_{ji})]\right)$. Hence, since $\gamma = -g$, then any *decreasing* function, $g$, in Equation 3.1 can be replaced by the increasing function, $-g$ (or vice versa), and the same value of $\overline{r}_j$ would obtain.

[5]The three-way equivalence would not hold if $g$ was a decreasing function, hence the earlier emphasis on increasing transforms. If $g$ was decreasing, then $g^{-1}$ would also be decreasing, which implies that the ordering based on *(C)* would be in the opposite direction to the ordering based on *(A)*, or *(B)*.

the two forms are identical, only their interpretation is different. Computationally, *(A)* is the best quantity to use because it requires the least calculation. *(B)* is only needed if the sorting-key in the generalised GOC algorithm (Section 2.4) must lie within the range of the transformed rating scale. *(C)* is only needed if the sorting-key must lie within the range of the original rating scale.

The simplest, trivial example of a transform-average mean is when $g(r) = r$. Under this transform, quantity *(A)* is the sum of ratings per stimulus, $\sum_{i=1}^{m} r_{ji}$, and quantities *(B)* and *(C)* are both the arithmetic mean rating per stimulus, which is a linear s.m.i. transform of the sum. GOC analysis based on sums of ratings (Watson, 1963; Taylor et al., 1991), and GOC analysis based on arithmetic mean ratings (Metz & Shen, 1992), are special cases of transform-average GOC analysis. They are formally equivalent to each other under the three-way equivalence.

Although the GOC curves based on each of the quantities within the three-way equivalence are identical for any single transform, GOC curves can differ for different transforms. If $g_1$ and $g_2$ are two different transforms, then the particular forms of $g_1$ and $g_2$ in Equation 3.1 *can affect the ordering of stimuli in a stimulus set*, and consequently, the resulting GOC curves can differ. Each transform results in its own three-way equivalence, which necessarily holds for any single transform, but which may or may not be identical across transforms. A consequence of this result is that the sum-of-ratings GOC curve on the original scale is not necessarily identical to the sum-of-ratings GOC curve on a transformed scale, because $\sum_{i=1}^{m} r_{ji}$ is not necessarily s.m.i. with $\sum_{i=1}^{m} g(r_{ji})$. This is shown in a numerical example in the following section.

Some families of transforms always result in the same GOC curve for different parameters. One example of this is the family of linear transforms [$g(r) = ar + b$, for constants $a$ and $b$], and the logarithmic transforms [$g(r) = \log_a(r)$] which result in the geometric mean. In general, the only situation where the order of means *must* be the same for *all* s.m.i. transforms of a data set is the degenerate case where the number of replications is $m = 1$.

### 3.1.2 Effect of different transforms

The purpose of this section is to show that the stimulus ordering, and hence GOC curve, is not *necessarily* the same for all transforms or scalings. Different transforms can result in different GOC curves. Assume four replications of data have been collected for a pair of stimuli, where the ratings were initially coded as positive integers. Assume further that the original ratings were 14, 9, 7 and 10 for the first stimulus and were 8, 12, 14 and 6 for the second stimulus. The three quantities for *(A)*, *(B)* and *(C)* of the three-way equivalence are given in Table 3.1 for each of the three transforms $g(r) = r$, $g(r) = \sqrt{r}$ and $g(r) = r^2$. Table 3.1 shows that the stimulus ordering is the same for each of *(A)*, *(B)* and *(C)* for any given transform, but the ordering changes across transforms.

| transform | sorting-key | Stimulus 1 | Stimulus 2 | order |
|---|---|---|---|---|
| $g(r) = \sqrt{r}$ | $A$ | 12.55 | 12.48 | $1 > 2$ |
| | $B$ | 3.14 | 3.12 | $1 > 2$ |
| | $C$ | 9.84 | 9.74 | $1 > 2$ |
| $g(r) = r$ | $A$ | 40 | 40 | $1 = 2$ |
| | $B$ | 10.0 | 10.0 | $1 = 2$ |
| | $C$ | 10.0 | 10.0 | $1 = 2$ |
| $g(r) = r^2$ | $A$ | 426 | 440 | $1 < 2$ |
| | $B$ | 106.5 | 110.0 | $1 < 2$ |
| | $C$ | 10.32 | 10.49 | $1 < 2$ |

TABLE 3.1: Example of order reversal in transform average GOC analysis. Values of different quantities in the three-way equivalence are listed for the rating scale example given in text. The order of the two stimuli according to each quantity is also shown ($j \in \{1, 2\}$ is the stimulus number, $i$ is the replication number). The quantities are: *A:* $\sum_{i=1}^{4} g(r_{ji})$; *B:* $\frac{1}{m}\sum_{i=1}^{4} g(r_{ji})$; and *C:* $g^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} g(r_{ji})\right)$.

The example in Table 3.1 only deals with one pair of stimuli. Over an entire experimental stimulus set, some pairwise orderings would change and others would remain the same, depending on the ratings for each stimulus and the transforms that were used. The example implies that the ordering of a stimulus set (including ties) depends on the type of transform chosen.

## 3.2 Transform-average GOC curves

Transform-average GOC analysis was applied to Taylor et al.'s (1991) continuous rating scale, frequency discrimination experiment, previously described in Chapter 2. An experiment such as this one is useful for evaluating transform-average GOC analysis because the theoretical ROC curve is known. If the theory was unknown, or uncertain, evaluation would be much more difficult.

Figures 3.1 to 3.5 show the 24-replication transform-average GOC curves based on 27 different s.m.i. transforms (or equivalently scalings) of a rating scale. Each graph displays the specific transform, $g(r)$, applied in Equation 3.1 to the original set of ratings (integers from 1 to 36). The resulting GOC curves differ across transforms. Some GOC curves follow the theoretical ROC curve along all of its length, but other GOC curves only follow the theoretical ROC curve along parts of its length. Interpretations of these results are discussed in Section 3.3.

Three GOC curves based on weighted averages, rather than transform-averages, are presented in Figure 3.5(b), (d) and (f), and are discussed separately.

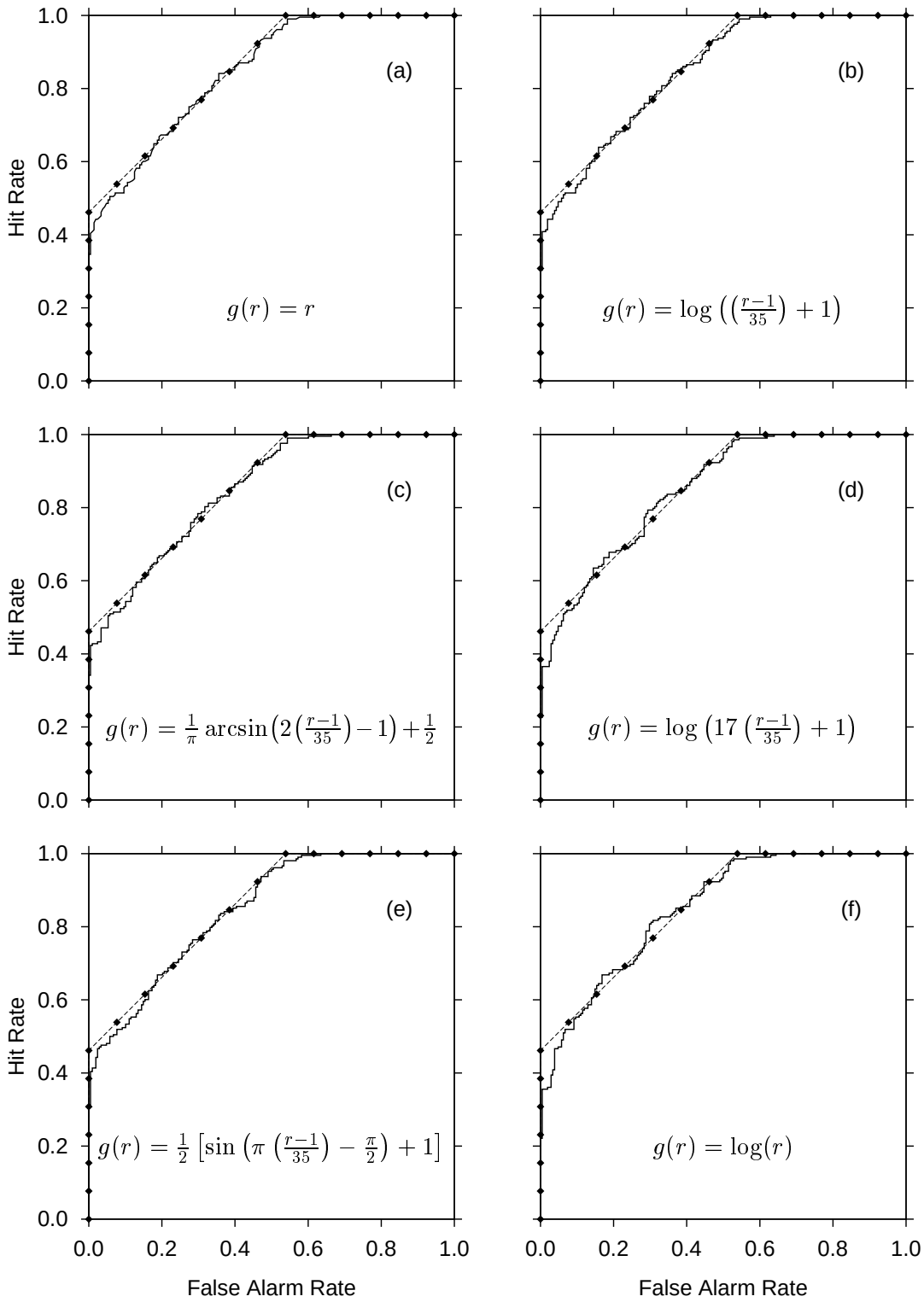


FIGURE 3.1: Transform-average GOC curves (solid lines) and the theoretical ROC curve (solid points and dashed lines). Each GOC curve is based on Equation 3.1, using the displayed transform, $g(r)$, where $r \in \{1, 2 \ldots 36\}$. Panels (a), (c) and (e) involve the *arithmetic mean*, the *arcsin-mean*, and the *sine-mean*, respectively. Panels (b), (d) and (f) involve *geometric means*, which are based on $g(r') = \log(r')$. Here, $r'$ is a linear transform of the original scale, $r$, onto the ranges $r' \in [1, 2]$, $r' \in [1, 18]$ and $r' \in [1, 36]$ in panels (b), (d) and (f), respectively.

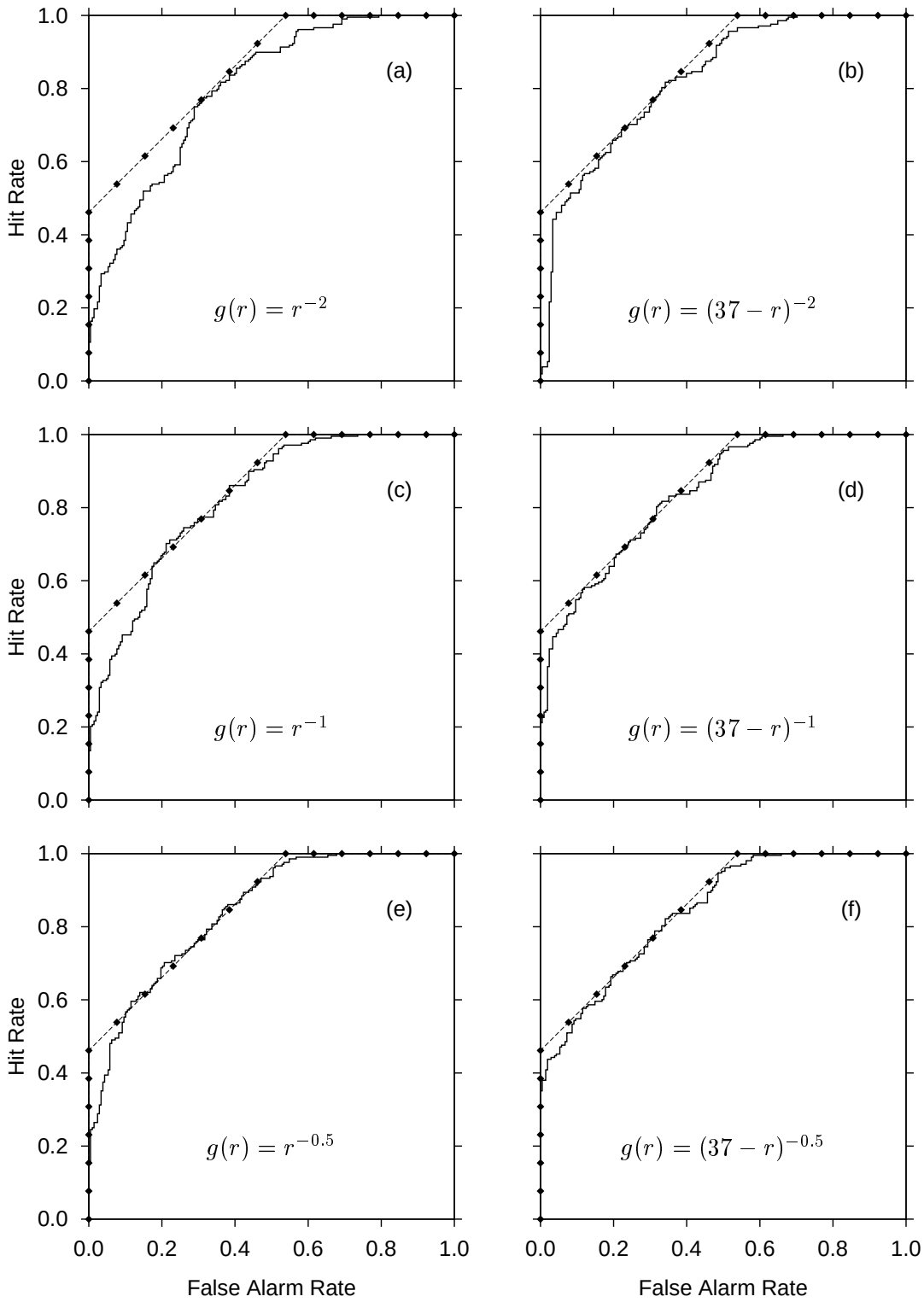


FIGURE 3.2: *Power-mean* transform-average GOC curves (solid lines) and the theoretical ROC curve (solid points and dashed lines). Each GOC curve is based on Equation 3.1, using the displayed transform, $g(r)$, where $r \in \{1, 2 \ldots 36\}$. Panels (a), (c) and (e) are based on forward-direction power means, $g(r) = r^c$, where $c$ equals $-2$, $-1$ and $-0.5$ respectively. Panels (b), (d) and (f) are based on reverse-direction power means, $g(r) = (37 - r)^c$, where $c$ equals $-2$, $-1$ and $-0.5$ respectively.

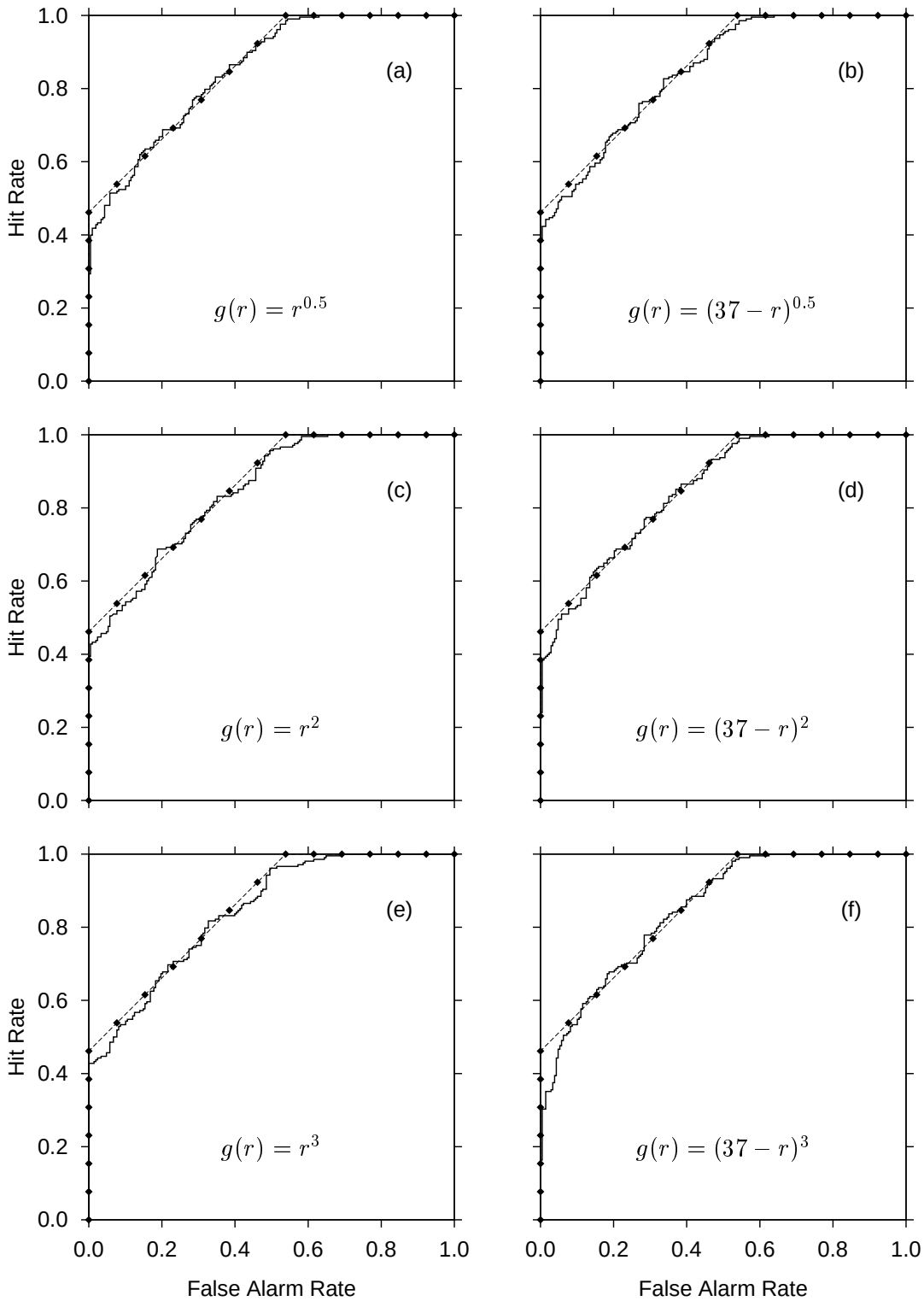


FIGURE 3.3: *Power-mean* transform-average GOC curves (solid lines) and the theoretical ROC curve (solid points and dashed lines). Each GOC curve is based on Equation 3.1, using the displayed transform, $g(r)$, where $r \in \{1, 2 \ldots 36\}$. Panels (a), (c) and (e) are based on forward-direction power means, $g(r) = r^c$, where $c$ equals 0.5, 2 and 3 respectively. Panels (b), (d) and (f) are based on reverse-direction power means, $g(r) = (37 - r)^c$, where $c$ equals 0.5, 2 and 3 respectively.

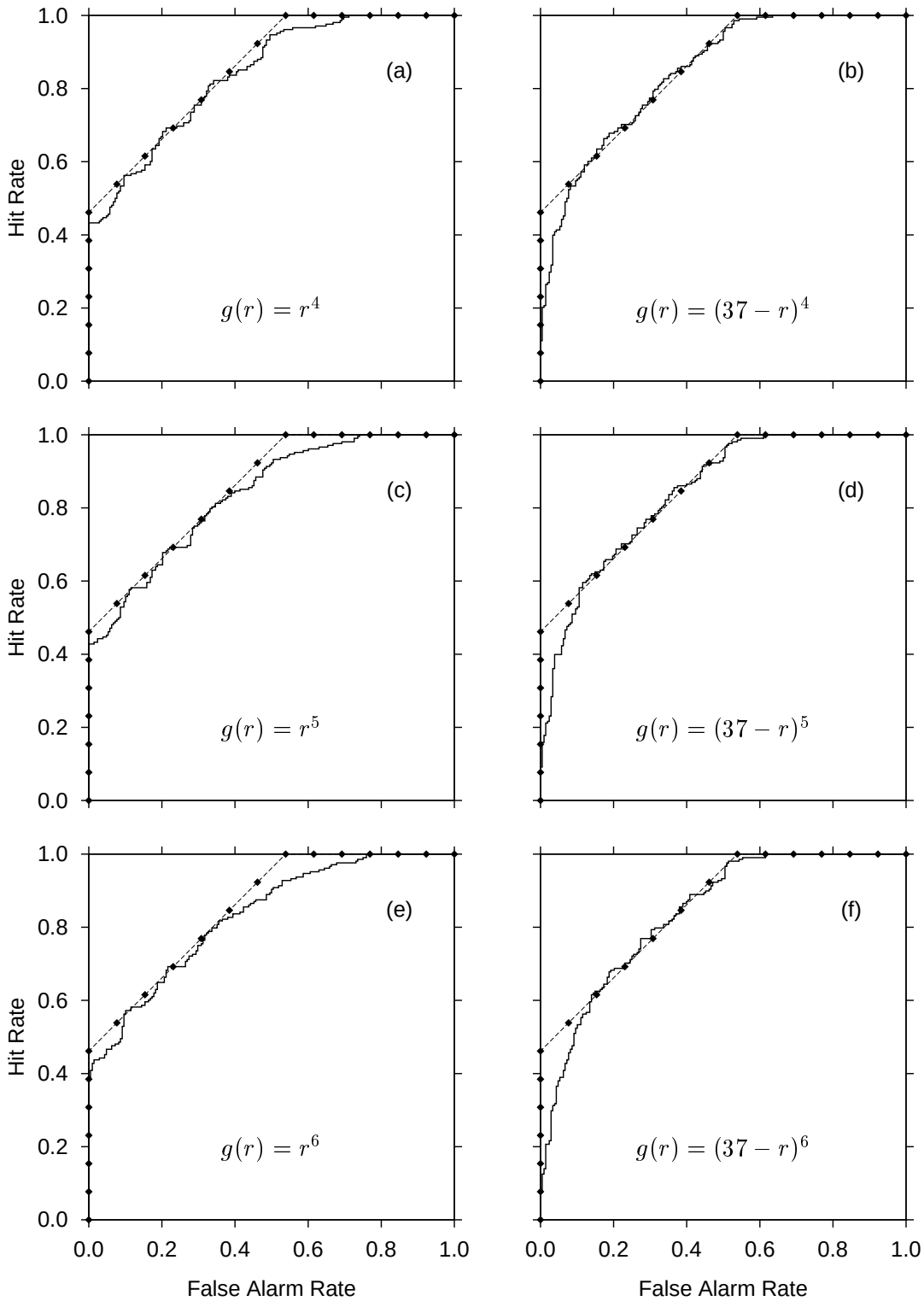


FIGURE 3.4: *Power-mean* transform-average GOC curves (solid lines) and the theoretical ROC curve (solid points and dashed lines). Each GOC curve is based on Equation 3.1, using the displayed transform, $g(r)$, where $r \in \{1, 2 \ldots 36\}$. Panels (a), (c) and (e) are based on forward-direction power means, $g(r) = r^c$, where $c$ equals 4, 5 and 6 respectively. Panels (b), (d) and (f) are based on reverse-direction power means, $g(r) = (37 - r)^c$, where $c$ equals 4, 5 and 6 respectively.

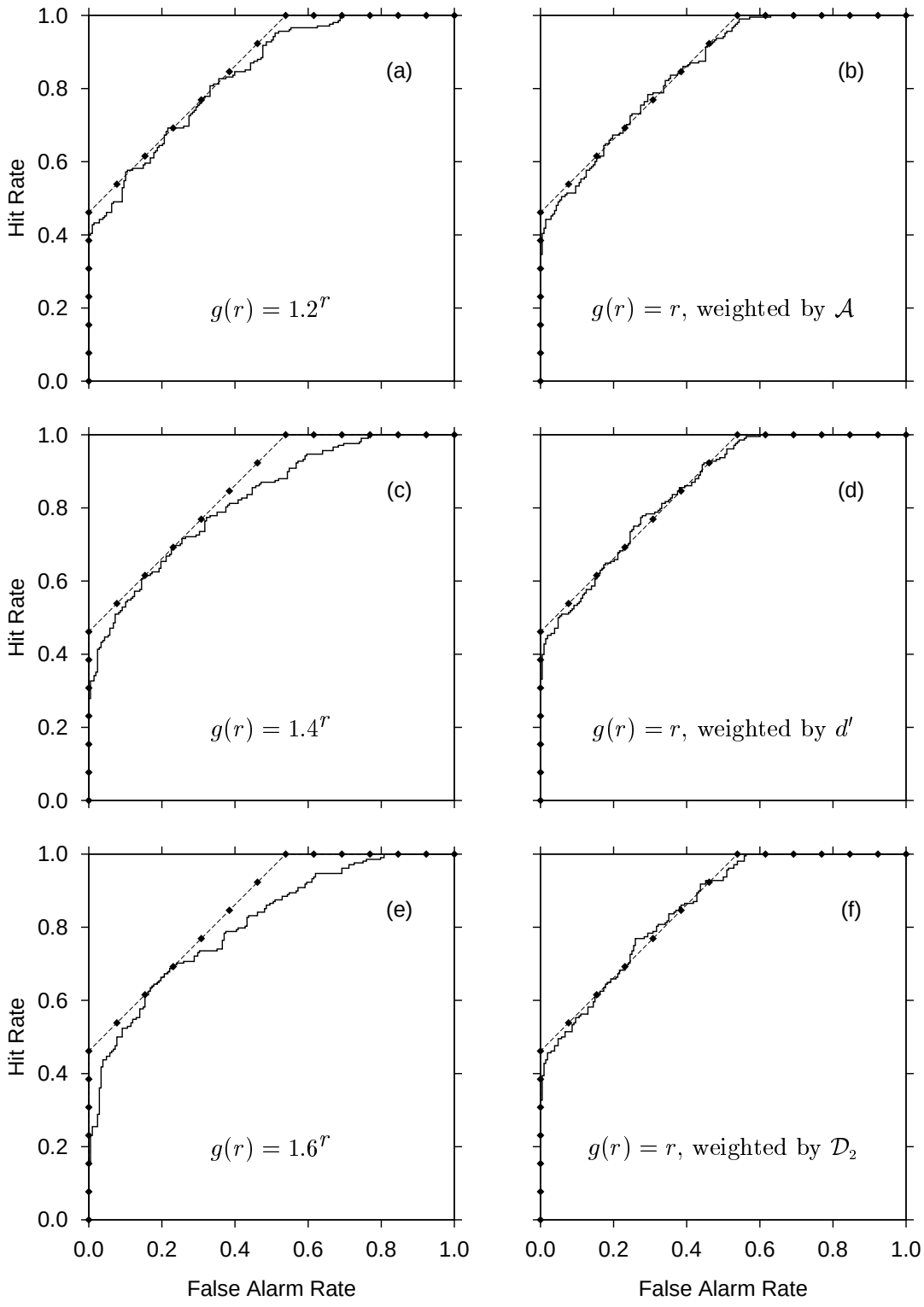


FIGURE 3.5: Transform-average GOC curves (solid lines) and the theoretical ROC curve (solid points and dashed lines). Each GOC curve is based on Equation 3.1, using the displayed transform, $g(r)$, where $r \in \{1, 2 \ldots 36\}$. Panels (a), (c) and (e) are based on *exponential-means*, $g(r) = b^r$, with $b = 1.2$, 1.4 and 1.6, respectively. Panels (b), (d) and (f) are based on *weighted arithmetic-mean ratings* (respectively weighted by single-replication values of $\mathcal{A}$, $d'$ and $\mathcal{D}_2$).

Figures 3.1(a), (c) and (e) respectively show the arithmetic-mean GOC curve[6] (from Figure 2.5), the arcsin-mean GOC curve, and the sine-mean GOC curve. The arcsin-mean is based on the transform $g(r) = \frac{1}{\pi}\arcsin\left(2\left(\frac{r-1}{35}\right)-1\right)+\frac{1}{2}$, and the sine-mean is based on $g(r) = \frac{1}{2}\left[\sin\left(\pi\left(\frac{r-1}{35}\right)-\frac{\pi}{2}\right)+1\right]$. The former transform involves a sigmoidal inverse transform in Equation 3.1, and the latter transform involves a sigmoidal inner transform. (Both transforms are described in further detail in Appendix A.) Although none of Figures 3.1(a), (c) and (e) are the same, all three GOC curves are very similar to each other, and all are good approximations to the theoretical ROC curve.

Figures 3.1(b), (d) and (f) show geometric-mean GOC curves, based on $g(r) = \log(r)$. What differs across curves is the range of the ratings prior to the logarithmic transform. The transforms are of the form $g(r) = \log\left(\frac{r-1}{35}c + d\right)$, where $d = 1$ for all three graphs, and $c$ equals 1, 17 and 35 for Figures 3.1(b), (d) and (f), respectively. For each GOC curve, the original scale was linearly mapped onto a different range prior to applying the logarithm. The ranges were [1,2], [1,18] and [1,36] for Figures 3.1(b), (d) and (f), respectively (where the range [1,36] is the original scale).

The figures show that GOC curves based on the geometric mean also follow the theoretical curve, although some scalings are better to use than others. The reason for this has to do with the nature of the transform over the range of ratings being used. The function $g(r) = \log(r)$ is almost linear in the range [1,2], so the similarity of Figure 3.1(b) to Figure 3.1(a) is not surprising. It appears that the less linear the transform, the less the agreement between the arithmetic-mean GOC curve and the geometric-mean GOC curve (the function $\log(r)$ is much more arced over the ranges [1,18] and [1,36] than over [1,2]).

These results *do not* suggest that the geometric mean is inherently worse to use than the arithmetic mean, although it was worse for this data set when the ratings were scaled over [1,36]. The area under the curve in Figure 3.1(b) is 0.8489, compared to 0.8483 for Figure 3.1(a), which shows that when the ratings were scaled over [1,2], the geometric mean GOC curve resulted in better overall performance than the arithmetic-mean GOC curve, albeit slighty.[7]

Figures 3.2, 3.3 and 3.4 all show power-mean transform-average GOC curves for different exponent values. Figure 3.2 shows power-mean GOC curves for negative exponents, where as those in Figures 3.3 and 3.4 show power-mean GOC curves for positive exponents. The range of the integer ratings is [1,36] for all of the curves. For each of Figures 3.2 to 3.4, the left-hand panels involve power means of the form $g(r) = r^c, c \neq 0$, whereas the right-hand panels all involve power means of the form $g(r) = (37-r)^c, c \neq 0$. To distinguish between them, the former are called *forward-direction* power means and the

[6] For brevity, terms such as "arithmetic-mean GOC curve" or "geometric-mean GOC curve" refer to GOC curves based on particular types of means applied to ratings. They do not refer to any averaging of GOC curves.

[7] This is a point of principle, rather than of practical benefit.

latter *reverse-direction* power means.[8] The function $37-r$ is a pre-transform which reflects any subsequent transform in the line $r = 18.5$ (i.e. the midpoint of the scale from [1,36]). A reverse-direction transform function is the mirror image (in $r = 18.5$) of any forward-direction transform function. The purpose of having both forward-direction and reverse-direction transforms was to see what effect compressing one end of the rating scale, versus the other end, may have had on GOC curves. Incorporating the function $37 - r$ into the power mean means that if the forward-direction transform stretches one end of the rating scale and compresses the other end, then the reverse-direction transform do the same, except to opposite ends of the scale.

From the top row to the bottom row in Figure 3.2, the power mean exponents in each pair of graphs are $-2$, $-1$ and $-0.5$ respectively. Figure 3.2(c) shows the harmonic-mean GOC curve, $g(r) = r^c, c = -1$, and Figure 3.2(d) shows its reverse-direction counterpart. The GOC curves improve as the transform becomes less extreme, although of the six curves presented, only the last one, the reverse-direction power-mean GOC curve with an exponent of $-0.5$, is close to the theoretical curve along all of its length. The other five curves meet the theoretical curve along some parts of their lengths, but fall short along other parts. This suggests that unique noise has been removed along some part of the rescaled rating scale, but remains in effect along other parts of the scale. *If* GOC analysis works for rating scales in general, then the results suggest that some transforms or scalings require more replications than others in order to fully recover the theoretical ROC curve.

The power-mean GOC curves in Figures 3.3 and 3.4 are based on positive exponents, whose values increase in progression from the top row of Figure 3.3 to the bottom row of Figure 3.4. The GOC curves progessively worsen as the exponent increases, but nevertheless, *all* of the GOC curves are consistent with the theoretical curve over some portions of their lengths. The reverse-direction GOC curves in the right-hand panels Figures 3.3 and 3.4 are approximate reflections, in the negative-diagonal, of the forward-direction GOC curves in the left-hand panels. This suggests that however the transforms may influence unique noise at one end of the scale, similar effects may be found at the other end of the scale, if the direction of the transforms is reversed. This pattern, found in power-mean GOC curves with positive exponents, does not extend to curves based on negative exponents in Figure 3.2.

Figures 3.5(a), (c) and (e) show exponential-mean GOC curves based on the very convex transform $g(r) = b^r$, with $b = 1.2$, 1.4 and 1.6 respectively. Although the range of parameters is small, the effect on the resulting curves is large. The transformed ratings range from 1 to $1.2^{36} \simeq 708.8$ for $b = 1.2$, but range from 1 to $1.6^{36} \simeq 2.2 \times 10^7$ for $b = 1.6$. Even with such an extreme transform, the portions of the GOC curve in Figure 3.5(e)

[8] All of the forward-direction power mean transforms with positive-valued power exponents are *increasing* functions, while those having negative-valued power exponents (i.e. hyperbolic transforms) are *decreasing* functions. The converse holds true for the reverse-direction power mean transforms. Also see footnote 4.

which match the theoretical ROC curve (portions in the middle and along the axes), match it almost exactly.

**Summary.** A variety of different transform-average GOC curves were calculated for Taylor et al.'s continuous rating scale experiment. All of the GOC curves were consistent with the known theoretical ROC curve, to a certain degree. Although many of the GOC curves fell short of the theoretical curve in some regions of the ROC space, none of the curves were entirely inconsistent with the theoretical curve. By the progressive adjustment of parameters for a given type of transform, it was shown that GOC performance deteriorated as the transforms became more extreme. Nevertheless, a wide range of transforms and parameters resulted in GOC curves that were very similar to the theoretical ROC curve. This suggests that there is no inherently-favoured scaling of a rating scale.

### GOC curves based on weighted sum of ratings

GOC analysis typically weights data from each replication equally. Figures 3.5(b), (d) and (f) are different from the other GOC curves presented in this chapter and in the rest of this thesis. Each curve is based on the *weighted* sum of ratings (or weighted arithmetic-mean ratings). The weighting was done according to single-replication sensitivity values, using $\mathcal{A}$, $d'$ and $\mathcal{D}_2$ values as the weights for Figures 3.5(b), (d) and (f) respectively. For example, let $d_i$ be the value of a performance measure derived from the single-replication ROC curve for the $i^{th}$ replication. The weighted sum, $\sum_{i=1}^{m}(d_i\, g(r_{ji}))$, replaces the transform-average in Equation 3.1 as the basis of the GOC curve. The weighted curves in Figures 3.5(b), (d) and (f) generally follow the theoretical ROC curve, although all differ in small detail.

The unweighted curve in Figure 3.1(a) was close to the theoretical curve, so any improvement from the unweighted curve could be only marginal. Weighted sums (or means) were also applied for a variety of GOC curves that fell below the theoretical curve over much of their lengths, to see if performance could be improved in cases where potential improvement was substantial. These GOC curves included the 8-replication curves from each observer based on sums of integer-valued ratings (Taylor et al., 1991, Figure 6(a), (b) and (c)), and a survey of the 24-replication transform-average GOC curves in Figures 3.1 to 3.5, including Figures 3.2(a) and 3.5(e), which had the poorest performance (based on unweighted mean ratings). Graphs of these various weighted GOC curves are not presented here, but a summary of results is given instead.

Three weighted GOC curves were derived for each of the unweighted curves described above, where the weightings used were single-replication values of $\mathcal{A}$, $d'$ or $\mathcal{D}_2$. In all cases, each of the three weighted GOC curves were very similar in gross form to the unweighted curve, and differences were only in the small details. Weighting sometimes resulted in a small increase in performance, and sometimes in a small decrease, but there was no general

improvement. In all cases, a single ROC curve fitted to the unweighted GOC curve could have described all three weighted GOC curves as well. The small scale changes were similar in nature to the differences between the unweighted curve in Figure 3.1(a) and the weighted curves in Figures 3.5(b), (d) and (f).

It intuitively seems that weighting each replication's data according to single-replication performance values should have enhanced GOC performance, but in fact it did not. It could be the case that single-replication performance values were too homogeneous to be effective as weightings, but the single-replication ROC curves (Figure 2.3) and their related measures were not homogeneous. Over 24 replications, $\mathcal{A}$ ranged between 0.6612 and 0.7901, $d'$ ranged between 0.5879 and 1.1410, and $\mathcal{D}_2$ ranged between 0.0763 and 0.2587. Whereas $\mathcal{A}$ varied only by a small amount, $d'$ varied by a factor of 2, and $\mathcal{D}_2$ varied by a factor of 3. If gross improvement could be achieved by weighted summation, then weighting by $d'$ or $\mathcal{D}_2$ should have produced it. The conclusion is that weighting ratings by single-replication performance has little overall effect on GOC performance.

## 3.3 Discussion

The main result shown in this chapter is that transform-average GOC analysis reduces unique noise for a variety of ordinal rating scales. Some implications of this are discussed here.

### Transform-average GOC curves compared to ROC curves

Unlike transform-average GOC curves, single-replication ROC curves are not affected by the scaling of a rating scale. The set of ROC curves remains *identical* for all s.m.i. scalings, and consequently the mean ROC curve is also identical across scalings. *All* of the GOC curves in Figures 3.1 to 3.5 provide much better performance than the mean ROC curve (Figure 2.5), including those based on extreme transforms which would not normally be considered in practice, such as $g(r) = r^6$ and $g(r) = 1.6^r$ (Figures 3.4(e) and 3.5(e) respectively). Scalings of a rating scale do affect results, but in a way that is not obvious from ROC analysis alone.

### Interpretation of variation across transforms

Some of the transform-average GOC curves were very close to the known theoretical ROC curve, while other GOC curves fell short of much of the theoretical curve. One possible explanation for these results is that in some unknown way, the common noise has changed, and that the common noise depends on the particular rating scale that is used in GOC analysis. Since 24 replications were enough to provide essentially unique-noise-free performance under the original scaling, it may be that each GOC curve reflected

unique-noise-free patterns of common noise that differed for different scalings. A second explanation is that the unique-noise-free, asymptotic GOC curve (attained as the number of replications tends to infinity) is the same for all scalings, but that some scalings were more efficient at removing unique noise than other scalings. The second explanation suggests that 24 replications were not enough to remove all of the unique noise, and that more replications would see all of the GOC curves tend towards the corners of the theoretical ROC curve.

The second interpretation is preferred over the first, primarily because the first interpretation runs counter to the idea that s.m.i. transforms of (unique-noise-free) scales result in the same ROC curve. Transform-average GOC curves do in fact differ across scales, but this is because they are affected by unique noise. It is expected that unique-noise-free performance reflects theoretical performance, given that the predominant source of common noise in Taylor et al.'s (1991) experiment was the reproducible stimulus set. The two interpretations of results could be tested using computer simulations of unique-noise-affected data sets (Taylor, 1984; Taylor et al., 1991), which would allow investigation over much larger numbers of replications than could be run experimentally.

### Integer ratings as ranks

Although some scalings of an ordinal rating scale are more useful than others, ROC analysis and transform-average GOC analysis treat all scalings as equally valid. While ROC analysis and mean ROC analysis are scale-invariant, transform-average GOC analysis is not. One procedure to reduce the arbitrariness of scales in GOC analysis is to convert ratings on any scale to *ranks*. This is equivalent to suggesting that equi-spaced positive integers (1,2, ... ) should be used, which is conventionally what is done. In this case, the interpretation of conventional GOC analysis is that it is based on the *sum-of-ranked-ratings*, or on the *arithmetic-mean-ranked-rating*.

While using rankings is a general way of standardising ratings, the arbitrary nature of rating scales is not removed, but merely hidden. With continuous rating scales, ranking ratings does not remove arbitrary scaling because the boundaries of successive rating categories and the number of categories are still set arbitrarily by the experimenter. Even if ranked ratings were used, any desired scaling (or s.m.i. transform) may be closely approximated (apart from a constant shift and scalar change) by judicious choice of the rating boundaries and number of categories. The result is that an arbitrary scaling can be introduced to a continuous rating scale, even if the resulting ratings are ranked. Although observers are often given instructions about how they *should* use a rating scale, experimenters do not control the way in which an observer *does* use a rating scale. This is important in GOC analysis, because how an observer uses a rating scale determines the effective scaling of a rating scale. For example, assume there are two theoretical observers which have identical decision axes (including unique and common noise effects), but which

differ in how evidence values on the decision axis map on to a rating scale, even if it is a ranked rating scale. One observer could effectively compress one portion of the decision axis onto the low end of the rating scale, while the other observer compresses the same portion of the decision axis onto the high end of the rating scale. As shown experimentally, (in Figures 3.3(e) and (f), for example) compression of different parts of a scale leads to different GOC curves, so each observer effectively chooses a different rescaling of the decision axis, even if the ratings are coded using the same numbers. Chapter 4 shows how relationships between decision axes and rating scales can be estimated from experimental data.

### The effect of equipment on scaling

A measurement apparatus may sometimes determine the scaling of a rating scale. The manipulandum used by Taylor et al. (1991) was a resistive slider, for which the voltage increased *linearly* with slider position. Since they partitioned the physical slider scale evenly into discrete categories, then the resulting ratings from 1 to 36 were a linear reflection of slider position. Another type of slider (such as a fader in a sound-mixing panel) could have voltage changing *logarithmically* with position. Had such a slider been used instead, then the conventional sum-of-integer-ratings GOC curve would have been the GOC curve shown in Figure 3.1(f) (i.e. the geometric-mean GOC curve based on linear-slider scaling). In that case, the smoother GOC curve shown in Figure 3.1(a) would have been an exponential-mean GOC curve, rather than the arithmetic-mean GOC curve that it is.

### Implications for a theory of GOC analysis

The above example illustrates the arbitrary nature of ordinal rating scales. It is not necessary to know the scaling underlying a rating manipulandum in order to perform a discrimination task. Furthermore, the partitioning of a continuous rating scale for data analysis is up to the experimenter, and is beyond the control of an observer. GOC analysis is applicable to any such scale. If the technique is to work on one arbitrary, ordinal scale, it should work on any ordinal scale, which is to say that GOC analysis should be transform-invariant. Evidence in this chapter suggests this may be so, albeit imperfectly, because all of the GOC curves followed the theoretical ROC curve to some degree. A theory showing how GOC analysis can work when there is arbitrary scaling of a rating scale is presented in Chapter 5.

### Sorkin and Dai's (1994) weighted summation model

Sorkin and Dai (1994) presented a model of group detection in a fundamental detection problem, where the group's decision is based on weighted sums of Gaussian random variables. Their block diagram shows multiple detectors with separate common and unique

noise inputs. There are multiple unique noise inputs to each detector, each of which consists of a sample from an independent Gaussian random variable. Particular unique noise inputs are shared only among a given subset of detectors, which provides a way to model partial correlations within a group. Each detector has a common noise input also, which is set to zero for the $N$ event, and is a constant (or a mean shift) for the $SN$ event. The mean shift may differ for different detectors, which is a mechanism for modelling a group of observers each of whom may have a different sensitivity or performance level in the task. The output of the $j^{th}$ detector, $X_j$ is weighted by $a_j$ and summed across detectors. A criterion-based decision rule is applied to $\sum_j a_j X_j$ to produce a binary-decision for the group. This is analogous to the weighted sum of ratings used to derive the GOC curves in Figures 3.5(b), (d) and (f).

The model incorporates both unique and common noise, and is very flexible with respect to individual sensitivity and correlation of unique noise within subsets of the group. Sorkin and Dai showed that appreciable improvement in group $d'$ is possible, depending on the parameter values and number of detectors in the modelled group. In spite of the similarities with weighted GOC analysis, Sorkin and Dai's model is not generally suitable as an analogy here because of two reasons: (1) the model is based on weighted sums of *evidence* values rather than *ratings*, and (2) the model assumes distributions that are Gaussian in form. Neither of these conditions apply to experimental data sets.

Sorkin and Dai's theory is an example of a model where unique noise is incorporated on the decision axis, and where the effects of unique noise are removed by averaging values on that decision axis. Whereas evidence values can be summed within a theory, this is a problem for an experimental data set, because only decisions or ratings are available in practice. Models such as Sorkin and Dai's need to be extended to incorporate rating scales (including binary-decision scales). It is possible to interpret Sorkin and Dai's model as a description of ratings on a rating scale, but this raises three further problems: (1) if a rating scale is discrete, with a relatively small number of categories, then the Gaussian form no longer applies; (2) even when a continuous rating scale is used and partitioned into a relatively large number of categories, experimental rating distributions probably will not be Gaussian;[9] and (3) ordered rating scales may be scaled arbitrarily by the experimenter.

**Summary.** GOC curves based on weighted sums of ratings differ from GOC curves based on unweighted sums, but the differences are reasonably small in the ROC space. Weighted sums of ratings did not change the overall form of a GOC curve, and did not result in much improvement in performance. Sorkin and Dai (1994) presented a model based on weighted sums of outputs of detectors, but the model in its current form needs modification to make the relationship between a decision axis and a rating scale explicit.

[9] The rating distributions on the original scale used in Taylor et al.'s (1991) experiment were almost bi-modal or multi-modal in form, depending on how each observer used the rating scale. Observers showed that they had favoured positions on the slider continuum.

**Summary of chapter**

Transform-average GOC analysis, based on different rating scales, can recover a theoretical ROC curve reasonably well, but extreme transforms may result in relatively poor GOC curves. Experimental results were consistent with the notion that, given enough replications, GOC analysis should be invariant with respect to the choice of rating scale or transform of that scale. The results suggested that 24 replications were enough to sufficiently recover the underlying theoretical curve for some scalings of Taylor et al.'s (1991) data set, but that more replications would be needed for other scalings. Compared to a mean ROC curve, transform-average GOC curves gave much better indications of the location, if not the shape, of the underlying theoretical curve, and did so over a wide variety of scalings. In experiments where the theoretical ROC curve is unknown, GOC estimates of the theoretical curve will depend on the number of replications combined and the scaling of the rating scale.

The main results presented in this chapter are that:

- Ratings can be averaged in a variety of different ways in order to derive a GOC curve (transform-average GOC analysis).
- A GOC curve based on sums-of-ratings from a rescaled rating scale is identical to a GOC curve based on transform-average mean ratings calculated on the original rating scale, because the processes involved in calculating the GOC curves are formally equivalent (the three-way equivalence).
- GOC analysis works in removing unique noise over a broad range of rating scales (or transform-averages), although some scalings (transforms) are better than others.
- A GOC curve depends on the scaling of a rating scale, whereas any single-replication ROC curve or mean ROC curve does not.
- Conventional GOC analysis based on sums of integer-valued ratings, or on arithmetic-means of integer-valued ratings, is a special case of transform-average GOC analysis.
- GOC analysis based on sums of ratings weighted by single-replication performance measures did not provide a substantial improvement over GOC analysis based on unweighted sums of ratings for Taylor et al.'s (1991) data set.

# Chapter 4

# Psychophysical transfer functions

Modelling observers in discrimination tasks is hampered by the fact that a decision axis is a hypothetical and unknown quantity, and that a there are an unlimited number of decision axes that result in the same theoretical ROC curve (Egan, 1975; Hanley, 1988). Furthermore, modelling the statistics of an observer's ratings depends not only on the form of a decision axis, but also on how evidence values translate into ratings. As Metz and Shen (1992) stated,

> Readers' [i.e. observers'] decision-variable outcomes cannot be measured. ... Ratings are related to the underlying decision variables in arbitrary (though ordinal) ways that vary from reader to reader, depending on how each reader uses his or her confidence-rating scale. (Metz & Shen, 1992, p. 70)

The relationship between a decision axis and a rating scale is called a *psychophysical transfer function*. It is implicit in models that partition a decision axis into sucessive intervals in order to derive or explain rating categories (McNicol, 1972; Hanley, 1988; Metz & Shen, 1992). Metz and Shen are correct in recognising that there is an arbitrary nature to transfer functions. However, *given* a decision axis, it is then possible to estimate a transfer function from a set of experimental data. Transfer functions are not arbitrary once a particular decision axis is assumed.

Section 4.1 shows how to estimate transfer functions from experimental results, by equating hit and false alarm rates on a GOC curve or ROC curve with theoretical hit and false alarm rates based on a model of performance on a decision axis. Transfer functions are estimated in Section 4.2 from Taylor et al.'s (1991) frequency discrimination experiment, based on assumed continuous decision axes. It is shown that functions can be estimated separately from both the GOC curve and the mean ROC curve, but that interpretation of the results is much easier for the former function than the latter function. The transfer function based on the GOC curve is used in Section 4.3 to estimate parameters of unique noise from the set of original ratings. Section 4.4 shows what happens to estimated transfer functions when inappropriate theoretical models are assumed. Section 4.5 shows an

alternative way of estimating transfer functions, by equating empirical and theoretical cumulative distribution functions, and discusses the problems and potential error associated with estimation. Finally, Section 4.6 deals with estimation when the assumed decision axis is discrete rather than continuous.

## 4.1 Estimation of a transfer function

A theoretical ROC curve results from a theoretical observer that makes decisions based on values on a decision axis, $X$. An empirical ROC curve results from the decisions of a real observer expressed on a rating scale, $R$. If a theoretical observer is used to explain the results of a real observer, then the real observer can be modelled, in part, as an input-output system that maps an evidence value, $x$, on the decision axis $X$ to a rating value, $r$, on the rating scale $R$. Let this mapping or transfer function be denoted as $R = h(X)$. A transfer function is hypothetical in that any decision axis $X$ is hypothetical, but once a particular decision axis is assumed, it is possible to estimate a transfer function from a set of empirical results on $R$. An estimated transfer function, $\widehat{h}$ provides information about how an observer may make decisions, and can be used to build a better model or simulation of an observer.

The term *theoretical* may be used somewhat loosely in the context of transfer functions. Preferably, the theoretical ROC curve and decision axis for a given experiment would derive from a statistical theory, or ideal observer, or simulated observer. In the absence of such a theory, however, the theoretical ROC curve could just be a curve that fits a given empirical ROC or GOC curve. Any decision axis that gives rise to the fitted curve can then be used to estimate a transfer function. If a fitted ROC curve is the only basis for assuming $X$, then the interpretation of $X$, and of the resulting transfer function, are less certain than if an ideal observer or substantive theory is involved.

A general assumption underlying the theoretical interpretation of empirical ROC analysis is the pairing of successive decision criteria on $X$ with successive criteria or cutoffs on $R$ (e.g. Wickelgren, 1968, Figure 1; McNicol, 1972, pp. 123-128; Hanley, 1988, Figure 4.) Criteria on $X$ and on $R$ are equivalent if the hit and false alarm rates on $X$ match those on $R$, which can only happen if the criteria on $X$ maintain the same relative order as on $R$. This requires the mapping $h$ to be a s.m.i. function or a monotonic increasing step function when dealing with a discrete rating scale (Bamber, 1975). Although $\widehat{h}$ is discrete in practice, it is assumed, for convenience, that the underlying transfer function, $h$, is s.m.i. and continuous, and that $X$ could be mapped continuously onto $R$ if only the resolution of $R$ could be fine enough. If a transfer function is based on an ROC curve or mean ROC curve, then the resolution of $R$ depends on the set of possible ratings. If a GOC curve is involved, however, then $R$ refers to the set of possible *average ratings* (or sums of ratings, if preferred). As more replications are added in GOC analysis, the number of possible

average ratings increases, and the effective resolution of $R$ increases substantially.

If a theoretical ROC curve matches a given empirical ROC or GOC curve, then empirical hit and false alarm rates can be equated with theoretical hit and false alarm rates, according to equivalent criteria, which gives the basis for estimating a transfer function. Let $HR_X(x)$ and $FAR_X(x)$ respectively denote theoretical hit and false alarm rates based on a criterion, $x$, and on a decision axis, $X$. Let $HR_R(r)$ and $FAR_R(r)$ respectively denote empirical hit and false alarm rates based on a criterion, $r$, applied to a rating scale, $R$. Equating theoretical with empirical hit rates requires finding a value, $x$, such that

$$HR_X(x) \quad = \quad HR_R(r)$$

holds for a given value of $r$ and its related hit rate, $HR_R(r)$. If $HR_X(x)$ has a unique inverse over the domain of $X$, then $x$ is found from

$$x \quad = \quad HR_X^{-1}(HR_R(r)) \tag{4.1}$$

As $r$ changes, $HR_R(r)$ changes, and so too does the estimated $x$. All such pairings of $r$ and $x$ define an estimated transfer function, $r = \widehat{h}_{\mathrm{SN}}(x)$. A similar process may be applied using false alarm rates, so that

$$x \quad = \quad FAR_X^{-1}(FAR_R(r)) \tag{4.2}$$

defines a separate estimated transfer function, $r = \widehat{h}_{\mathrm{N}}(x)$. Equations 4.1 and 4.2 imply that each separate point on an empirical ROC or GOC curve (i.e. a change in empirical hit or false alarm rate) corresponds to a point on $\widehat{h}_{\mathrm{SN}}(x)$, or a point on $\widehat{h}_{\mathrm{N}}(x)$, or both.

The functions $\widehat{h}_{\mathrm{SN}}$ and $\widehat{h}_{\mathrm{N}}$ presumably estimate the same underlying transfer function, $h$, so $\widehat{h}_{\mathrm{SN}}$ and $\widehat{h}_{\mathrm{N}}$ should be the same over any domain in $X$ that is common to both events. Egan (1975) stated that "an observer cannot apply one transformation for $x$-values from $N$-trials and a different transformation for $x$-values from $SN$-trials," (Egan, 1975, p. 48). In the context of transfer functions, this implies that the same $x$-value should map on to the same $r$-value regardless of which event is associated with any particular stimulus. If $\widehat{h}_{\mathrm{SN}}(x)$ and $\widehat{h}_{\mathrm{N}}(x)$ are different, then two different stimuli with the same $x$-value could result in different decisions on $R$, depending on the event that occurred. This implies that discrimination between the $SN$ and $N$ events is possible based on identical evidence, $x$, which would imply that the nominal decision axis is not the basis for decisions. A substantive discrepancy between $\widehat{h}_{\mathrm{SN}}(x)$ and $\widehat{h}_{\mathrm{N}}(x)$ would be grounds for rejecting the assumed conditional distributions on $X$ and trying another pair of distributions, or another theory.

**Strictly monotonic increasing transforms of the assumed decision axis**

There are any number of related transfer functions that may be estimated from the same cumulated empirical data. This is because any s.m.i. transform of an assumed decision axis results in the same theoretical ROC curve (Egan, 1975), and transfer function estimation allows any of these theoretical decision axes to be associated with the rating scale for a given experiment. This holds true regardless of whether or not the theoretical curve matches the empirical curve, and comes about because the scaling of both the rating scale and the decision axis may well be arbitrary.

Let $X_1$ be a decision axis that is assumed in order to estimate a transfer function. Let $X_2$ be the decision axis resulting from a s.m.i. transform of $X_1$. The estimated $x$-values based on $X_2$ are s.m.i. transforms of the estimated $x$-values based on $X_1$ and vice versa. This implies that the estimated transfer function based on $X_2$ results from a (horizontal) s.m.i. transform of the transfer function based on $X_1$. If the two event-conditional functions on $X_1$ are consistent with each other, then the event-conditional functions on $X_2$ are also consistent because both functions are transformed together. If they are not consistent on $X_1$, then they are not consistent on $X_2$, and no s.m.i. transform of $X_1$ would make it consistent.

The assumed decision axis used to estimate a transfer function may be the result of a theory involving an ideal observer for the particular discrimination task, or the assumed decision axis could be the result of curve-fitting to empirical ROC or GOC curves. In the former case, the theory should predict a particular decision axis having a particular scaling, and $X$ is set in the context of the theory (along with any units of measurement that may apply). In the latter case, *any* s.m.i. transform of an assumed decision axis is as valid as the original axis because there is no further information with which to justify the choice. In that case, the decision axis that is chosen is an arbitrary choice out of the set of ordinally related decision axes.

## 4.2 Experimental transfer functions

Transfer functions were estimated for Taylor et al.'s (1991) continuous rating scale experiment, which was described in Chapter 2. The 24-replication GOC curve based on arithmetic mean ratings and the arcsin-averaged mean ROC curve (Figure 2.5) are each used to estimate transfer functions. The pair of theoretical distributions assumed for each curve is different, because the curves themselves are different. Overlapping continuous uniform distributions are assumed for the GOC data, and Gaussian unequal distributions are assumed for the mean ROC data. Results are compared once both sets of transfer functions are presented.

### Transfer functions from the GOC curve, assuming a continuous uniform model

The decision axis in the experiment involved *discrete* overlapping uniform distributions (Figure 2.2). For computational convenience, the transfer functions presented here are based on assumed *continuous* uniform distributions which provide good approximations to the discrete distributions. Transfer functions based on a discrete uniform distributions are given in Section 4.6, along with a description of problems that arise when dealing with discrete decision axes. From a practical point of view, if the theory was not known for this experiment, then the continuous uniform model is a reasonable assumption, given the form of the GOC curve. The assumed probability density functions are

$$f(x|N) = \frac{1}{65}, \quad 595 \leqslant x \leqslant 660 \tag{4.3}$$

and

$$f(x|SN) = \frac{1}{65}, \quad 625 \leqslant x \leqslant 690, \tag{4.4}$$

where units of Hertz may be assigned to this decision axis. The bounds have been chosen so the hit and false alarm rates for the continuous distributions match the hit and false alarm rates for the discrete distributions at the frequencies at which probability was massed. This results in the same theoretical ROC curve as shown in Figure 2.5. (Note that the 13 distinct frequencies of each discrete uniform distribution ranged over 60 Hz. The range of each continuous distribution must be $13 \times 5$ Hz $= 65$ Hz, however, in order for the discrete and continuous hit or false alarm rates to match each other.)

Figure 4.1 shows the estimated event-conditional transfer functions,[1] which are presented separately in Figures 4.1(a) and 4.1(b). The two functions are relatively consistent with each other, as can be seen in Figure 4.1(c). The function estimated in Figure 4.1(c) is reasonably linear, but with kinks corresponding to the ends of the event-conditional domain boundaries around 625 Hz and 660 Hz. The entire function could be approximately fitted by one straight line, or more precisely by three joined line segments (as is done in Section 4.3). The graphs show that frequency in Hz was mapped on to the rating scale by an approximately linear transform.

### Transfer functions from the mean ROC curve, assuming a Gaussian unequal variance model

A transfer function can be estimated from a single-replication ROC curve or a mean ROC curve as well as a GOC curve. The ROC curves for the experiment vary considerably

[1] Strictly speaking, $x$ should be plotted as a function of $r$, because $x$-values are estimated from $r$-values, rather vice versa. However, $R$ is seen as a function of $X$ in a theoretical context, so transfer functions are presented with $x$ as the independent variable and $r$ as the dependent variable.

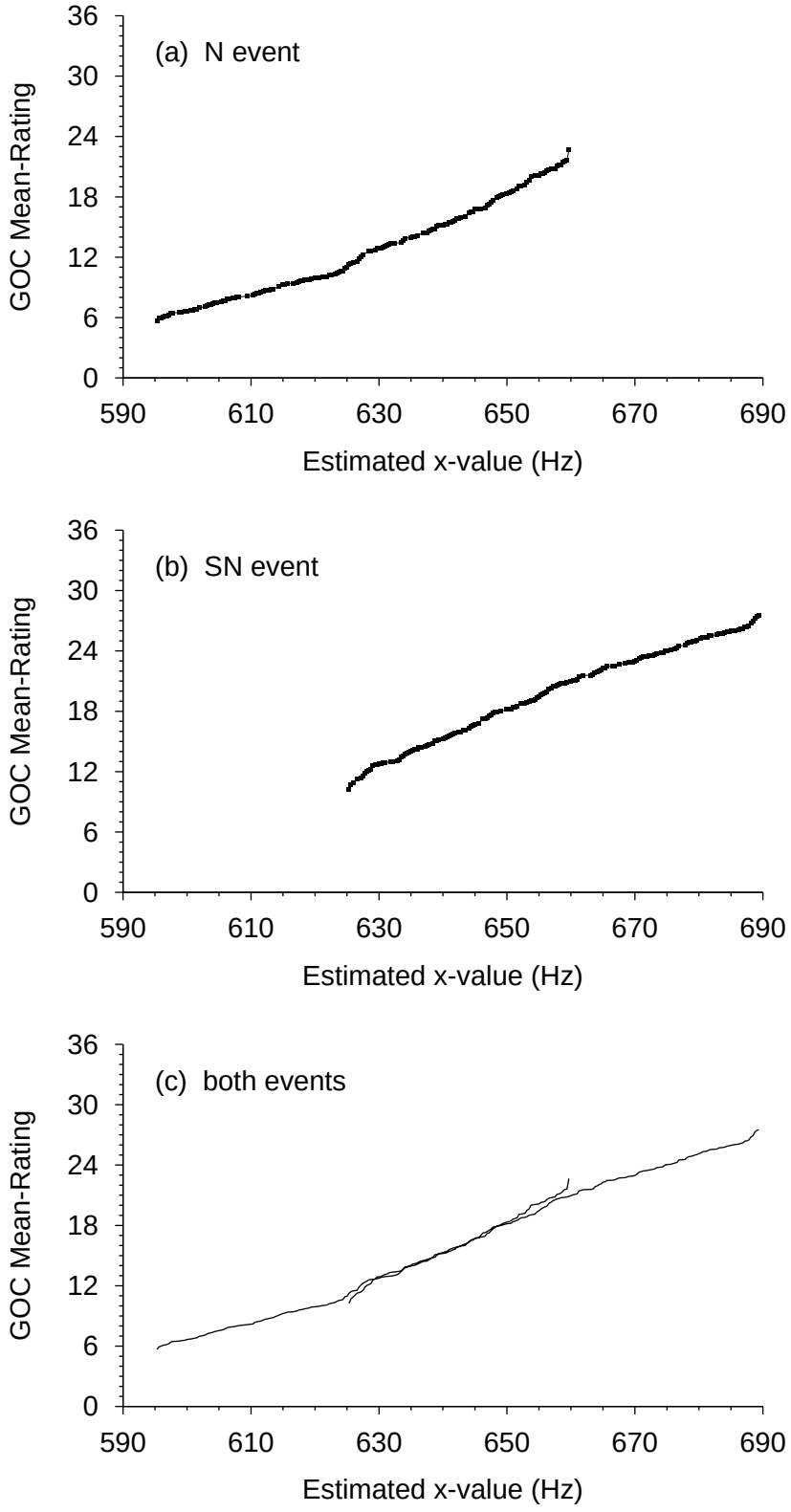


FIGURE 4.1: Transfer functions estimated from the GOC curve, assuming continuous uniform distributions. (a) Based on the $N$ event only. (b) Based on the $SN$ event only. (c) Both the $N$ and $SN$ functions together (shown as thin lines, for clarity).

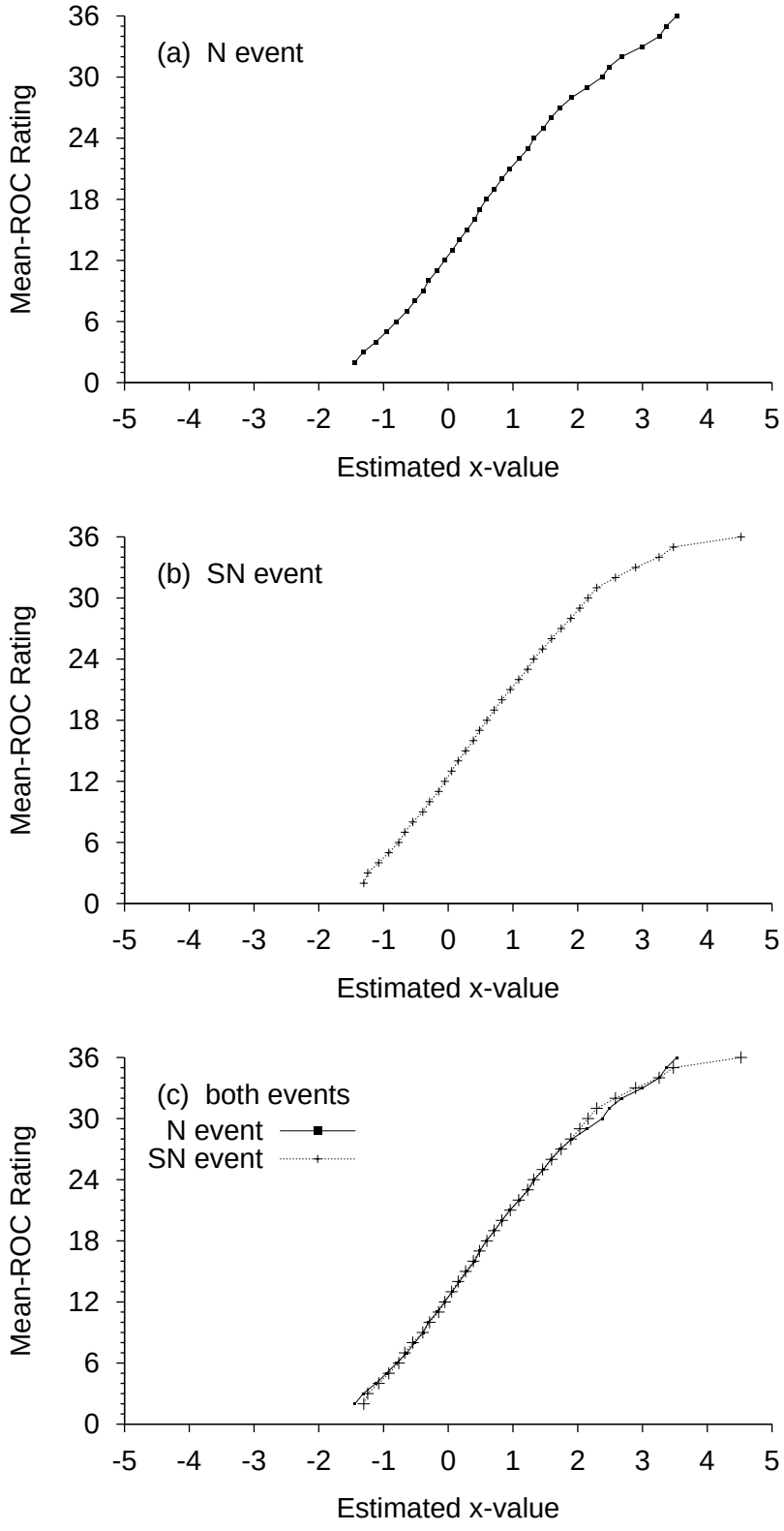


FIGURE 4.2: Transfer functions estimated from the mean ROC curve, assuming Gaussian unequal variance distributions. (a) Based on the $N$ event only. (b) Based on the $SN$ event only. (c) Both the $N$ and $SN$ functions together.

across replications (Figure 2.3), and although a different theoretical ROC curve could be fitted to each one, it would be difficult to reconcile the variety of estimated decision axes and transfer functions across replications. A transfer function is estimated here from the 24-repln mean ROC curve shown in Figure 2.5, since it represents typical single-replication ROC performance.

As was noted in Section 2.3.3, the mean ROC curve for this experiment can be fitted well by a theoretical ROC curve based on a Gaussian unequal variance model, with means and standard deviations of $\mu_N = 0.0$, $\sigma_N = 1.0$, $\mu_{SN} = 0.94$, and $\sigma_{SN} = 1.07$, giving $d_z = 0.9077$. The estimated event-conditional transfer functions based on the mean ROC curve, and assuming these distributions, are presented in Figure 4.2. The functions are reasonably linear, and are essentially consistent across events, except at the highest $x$-value.

The apparently large discrepancy between estimates at the highest $x$-value is only a reflection of the sensitivity of the specific (Gaussian) model to very small differences at extreme probability values. The discrepancy relates to the highest rating category, and stems from tiny differences between the mean ROC curve and the fitted ROC curve in the region of the ROC space closest to the origin. The curves differed in hit and false alarm rate by less than 0.001, which corresponded to $z$-scores of about 3.5. This shows that when evaluating the agreement between event-conditional transfer functions, values that relate to the tails of assumed distributions are not as important as values that relate central regions of the distributions. In Figure 4.2, this corresponds to estimated $x$-values between about $-1$ and 2 (or $\mu_N$ minus one standard deviation to $\mu_{SN}$ plus one standard deviation), over which the agreement is excellent.

### Comparison and interpretation

The pairs of event-conditional transfer functions estimated from the GOC curve and from the mean ROC curve are each self-consistent across events, and all of the functions are generally linear. Apart from these properties, the functions based on the GOC curve and on the mean ROC curve are not comparable to each other. Figure 4.1 was based on the assumption of uniform distributions, whereas Figure 4.2 was based on the assumption of Gaussian distributions. How each of these axes relate to each other is not clear. The apparent solution might be to adopt one common axis, and to use it to estimate transfer functions from both the GOC curve and the ROC curve. This is no solution, however, because the resulting event-conditional transfer functions cannot be consistent for both the GOC and mean ROC data for this experiment. If a single common axis existed, then the pair of theoretical distributions that defined the common axis would result in a curve that is consistent with both the GOC curve and the mean ROC curve. Such a curve is impossible because the GOC curve and mean ROC curve for this experiment are so different.

The transfer functions shown in Figures 4.1 and 4.2 are the consequence of two different models that are tied to different aspects of the same data, namely unique-noise-free and unique-noise-affected decisions. Each model is appropriate within its own context. Since the experiment involved a frequency discrimination task, it makes sense to interpret any decision axis that is proposed in terms of aural frequency (in Hz), or some s.m.i. transform of it. Furthermore, the rating value on each trial was a linear function of the position of the rating slider, with three rating categories per centimeter (approximately). This implies that units of measurement can be assigned to both axes in Figure 4.1, indicates how frequency in Hz maps, *on average*, to slider position (in 1/3-centimeter steps). Assigning units of measurement to the decision axis defined by uniform distributions,[2] is reasonably straightforward, because of the way in which the experiment was designed. In contrast, it is not clear how to assign units of measurement to the decision axis in Figure 4.2, other than interpreting it in terms of standard scores.

It is curious that the transfer functions are so linear. In principle, the only assumption that needs to be satisfied in order for ROC analysis to apply is that the rating scale is an s.m.i. function of the decision axis. The functions could have been concave, convex, or sigmoidal, without upsetting any fundamental theoretical assumptions. Transfer functions of other shapes are shown in Section 4.4, although they result from inappropriate models being used in the estimation procedure. The fact that the functions for Taylor et al.'s (1991) experiment were linear, should not be generalised, because they only derived from a single experiment.[3]

## 4.3 Use of transfer functions to quantify unique noise

Unique noise has been modelled as extra variability associated with each individual stimulus (e.g. Siegel, 1979; Siegel & Colburn, 1989). Under such models, each stimulus is associated with a distribution *across replications* of unique-noise-affected evidence values that lie on the decision axis. On the $i^{th}$ replication and for the $j^{th}$ stimulus, an evidence value is sampled from the distribution for the $j^{th}$ stimulus. A decision rule is applied to the unique-noise-affected evidence value, which results in the decision, or rating, for the $i^{th}$ stimulus on the $j^{th}$ replication. The unique-noise-affected evidence values differ across replications for the same stimulus, and hence the decisions that are made differ across replications. Once a transfer function has been estimated, it is possible to work backwards from an experimental data set of ratings to derive sample distributions and statistics of unique noise defined on a supposed decision axis. This was done for Taylor et al.'s (1991) data set.

[2] Including continuous approximations to the discrete uniform distributions.

[3] Preliminary analysis of Whitmore et al.'s (1993) amplitude discrimination experiment suggested that non-linear transfer functions apply in other experiments.

In order to convert $r$-values into $x$-values, it is convenient to first fit a function to the estimated transfer function. A function composed of three joined line segments was fitted by eye to the estimated transfer functions that are shown in Figure 4.1(c). The line segments were defined by the coordinates $(595, 5.8)$, $(625, 10.75)$, $(660, 21.1)$, and $(690, 26.9)$, and provided a reasonable approximation. The transfer function based on these points is defined as

$$r = \begin{cases} 0.1650x - 92.375, & x \leqslant 625, \\ 0.2957x - 174.071, & 625 < x \leqslant 660, \\ 0.1933x - 106.500, & x > 660, \end{cases}$$

where $r$ is a value on the rating scale, and $x$ is value of frequency in Hz. The transfer function in Figure 4.1(c) was based on arithmetic mean ratings, and only ranged in value between approximately 5.8 and 26.9 on the rating scale. The regression function may be extrapolated, by linearly extending the outer two line segments to cover the full range of the rating scale. According to the extrapolation, a rating of one equated to a frequency of 566 Hz, and a rating of 36 equated to a frequency of 737 Hz. Both values lie well beyond the frequency bounds that were used (595–685 Hz).

In order to estimate an $x$-value for any given rating $r$, the inverse of the regression function is required, namely

$$x = \begin{cases} 6.0606r + 559.848, & r \leqslant 10.75, \\ 3.3816r + 588.647, & 10.75 < r \leqslant 21.1, \\ 5.1724r + 550.862, & r > 21.1. \end{cases}$$

This inverse transfer function can be used to convert the entire data set of ratings into a set of estimated $x$-values. There are several possible uses for a set of estimated $x$-values. For example, assumptions about the form of unique noise distributions on $X$ may be checked, simulations of observers can be adjusted to better account for data, and parameters that characterise the unique noise may estimated.

Estimated $x$-values from the experiment were used to estimate the unique noise variance on $X$. Since there were 24 replications, each stimulus was associated with a set of 24 ratings, which were converted into a set of 24 estimated $x$-values. The sample variance of estimated $x$-values was calculated for each stimulus. Table 4.1 gives the average variance per stimulus, calculated for each event separately, and across both events. There were 208 stimuli per event, so the average variances in Table 4.1 resulted from averaging 208 variances per event, and 416 variances across both events. Table 4.1 also gives the square root of the average variance. This is the root-mean-square (r.m.s.) of sample standard deviation values, given in units of Hz. It indicates the effective amount of spread of unique noise on $X$. The average variances are estimates of the variance of unique noise, $\sigma_{\text{u}}^2$.

| | average variance | (average variance)$^{\frac{1}{2}}$ (Hz) | $\sigma_u^2/\sigma_c^2$ |
|---|---|---|---|
| $N$-event | 883.25 | 29.72 | 2.51 |
| $SN$-event | 1002.01 | 31.65 | 2.84 |
| both events | 942.63 | 30.70 | 2.68 |

TABLE 4.1: Sample statistics of estimated $x$-values for Taylor et al.'s (1991) experiment, and estimates of $k = \sigma_u^2/\sigma_c^2$ based on $\sigma_c^2 = 352.08$ Hz$^2$.

The variance of each of the continuous uniform distributions used to estimate the transfer function was $\sigma_c^2 = 352.08$ Hz$^2$ (i.e. $(65^2)/12$, where the range was 65 Hz).

The r.m.s. sample standard deviation was on the order of 30 Hz, whereas the standard deviation of each continuous uniform distribution was 18.7 Hz, by comparison. Clearly, unique noise dominated common noise in this experiment, which reflects the fact that extra unique noise was deliberately introduced into this experiment (as explained in Section 2.2). The estimated value of $k = \sigma_u^2/\sigma_c^2$ was approximately 2.7, when estimated across all stimuli from both events. In light of the amount of unique noise, the fact that GOC analysis removed most of its effects (Figure 2.5) is impressive.

The estimated value of $k$ was slightly larger for the $SN$ event, than for the $N$ event, because the estimate of $\sigma_u^2$ was larger for the $SN$ event. The ratio of estimated standard deviations for the $SN$ event relative to the $N$ event was $31.65/29.72 = 1.065$. This is very close to the ratio of standard deviations for the Gaussian unequal variance model that was fitted to the mean ROC curve, which was 1.070. Although this superficially suggests that unique noise characteristics were different for each event, it is more likely that the unique noise variance increased as a function of $x$. Such a result poses a problem for any model based on the addition of a single unique noise random variable to common noise values on $X$ (e.g. Wickelgren, 1968; McNicol, 1972), because the same random variable can not account for differences across events.

### 4.3.1 Implications

Estimated transfer functions were used as part of a new method of quantifying unique noise. The estimate of $\sigma_u^2$, however, is only as firm as the assumptions on which the estimation are based, and rests especially on the choice of assumed theoretical decision axis. As noted in Section 4.1, the transfer function based on one assumed decision axis, $X_1$, may be converted into a transfer function based on a second decision axis, $X_2$, that is an s.m.i. transform of $X_1$. Without a firm basis for choosing one axis over the other, either is acceptable. Preferably, the basis for choice should be *a priori* rather than *post hoc*. If the assumed axis is based solely on an ROC fit to a GOC curve, then any axis that results in the fitted curve is as acceptable as any other. The data set examined here is unusual

in that the theoretical distributions were known, and could be defined *a priori* based on experimental design independently (for the most part) of an observer. Even for this experiment, however, there is no guarantee that aural frequency, $X$, was used, rather an s.m.i. transform of it, such as $\log(X)$.

The choice of possible decision axis has consequences for quantifying unique noise. Quantities such as $\sigma_u^2$ and $\sigma_c^2$ must be affected by non-linear s.m.i. transforms of a decision axis, and the estimated ratio, $k$, must depend on the assumed axis. If $X_1$ and $X_2$ are non-linear s.m.i. transforms of each other, and if $\sigma_u^2$ is constant as a function of $x$ on $X_1$, then $\sigma_u^2$ is generally *not* constant as a function of $x$ on $X_2$. For example, let a single unique noise variable, $U_1$, be defined on the frequency decision axis, $X_1$, where $U_1$ has a fixed variance for stimuli of all frequencies. Also, let the mean value of $U_1$ be the frequency (in Hz) of a given stimulus. If $X_2 = \log(X_1)$, then the variance of $U_2 = \log(U_1)$ on $X_2$ will increase as the logarithm of the stimulus frequency increases, which is quite different from the pattern found on $X_1$. This calls into question the concept of $k$, since $k$ relies on the assumption that $\sigma_u^2$ is a fixed value with respect to a decision axis. Although this is the case on $X_1$, it is not the case on $X_2$. For a given data set, there is no guarantee that a decision axis exists which has unique noise of a fixed variance, nor is there a guarantee that such a decision axis could be found, even if it did exist.

## 4.4 Transfer functions based on inappropriate models

The transfer function based on GOC results assumed theoretical overlapping uniform distributions whereas that based on the mean ROC curve assumed theoretical overlapping normal distributions. In both cases, the assumed theoretical ROC curve matched the empirical curve. The purpose of this section is to show what happens to the estimated transfer function when the assumed theoretical ROC curve does not match the empirical curve.

### Transfer functions from the mean ROC curve, assuming uniform distributions

Transfer functions were estimated from the mean ROC curve, assuming overlapping continuous uniform distributions—a deliberately inappropriate combination. Two pairs of uniform distributions were used in turn. The first pair is that of the earlier GOC model given by Equations 4.3 and 4.4, which result in a theoretical ROC curve that is well above the mean ROC curve. The second pair of uniform distributions had the same standard deviations as the first pair, but with a smaller separation between the means, chosen so the area under the theoretical ROC curve was equal to the area under the mean ROC curve. Under this assumed model, the probability density function for the $N$ event was

still as given in Equation 4.3, but the density function for the $SN$ event was

$$f(x|SN) = \frac{1}{65}, \quad 613.13 \leqslant x \leqslant 678.13. \tag{4.5}$$

The theoretical ROC curve in this case is a straight line parallel to the chance line that intersects the curved mean ROC curve twice, and is a very poor fit to the mean ROC curve.

Figure 4.3(a) shows the estimated transfer functions based on the mean ROC curve and assuming the continuous overlapping uniform distributions of Equations 4.3 and 4.4, while Figure 4.3(b) shows those for the distributions of Equations 4.3 and 4.5. The two event-conditional transfer functions are quite different from each other in both cases. This clearly indicates that there is something wrong with the assumed decision axes.

Given that the transfer functions in Figure 4.3(a) and (b) are all based on overlapping uniform distributions (with different parameters), it is not surprising that the functions are so similar to each other. The $N$ event function is the same for both pairs because the $N$ distributions are the same. The two $SN$ event functions are just shifted versions of each other, because the effect of changing the location of the assumed $SN$ distribution is to change the location of the estimated $SN$ transfer function. However, there is no possible horizontal shift of the $SN$ distribution such that the $N$ and $SN$ transfer functions will match each other. The overlap between the functions shown in Figure 4.3(b) is approximately the best that can be achieved based on the mean ROC curve and assuming uniform distributions of equal variance.

### Transfer functions from the GOC curve, assuming normal distributions

Transfer functions were estimated from the GOC curve assuming overlapping Gaussian distributions. These are deliberately inappropriate assumptions because any Gaussian theoretical ROC curve is curved in the ROC space whereas the GOC curve is reasonably straight. Two pairs of Gaussian distributions were used. The first pair were random variables of *equal* variance, with $d' = 1.4553$ (with $\sigma_{\mathrm{N}} = \sigma_{\mathrm{SN}} = 1.0$, $\mu_{\mathrm{N}} = 0$, and $\mu_{\mathrm{SN}} = 1.4553$). These parameters were set so that area under the theoretical curve was the same as the area under the GOC curve. The second pair of distributions were the specific *unequal* variance random variables, described in Section 4.2, that provided a good fit to the mean ROC curve and resulted in $d_z = 0.9077$.

Figure 4.4(a) shows the estimated transfer functions based on the GOC curve assuming the Gaussian equal variance model with $d' = 1.4553$. Figure 4.4(b) shows the estimated transfer functions based on the GOC curve and assuming the Gaussian unequal variance model with $d_z = 0.9077$. In both cases, the event-conditional functions are not consistent. The functions in Figure 4.4(a) only meet at two points. These correspond to the points where the theoretical ROC curve intersects the GOC curve (i.e. where theoretical hit

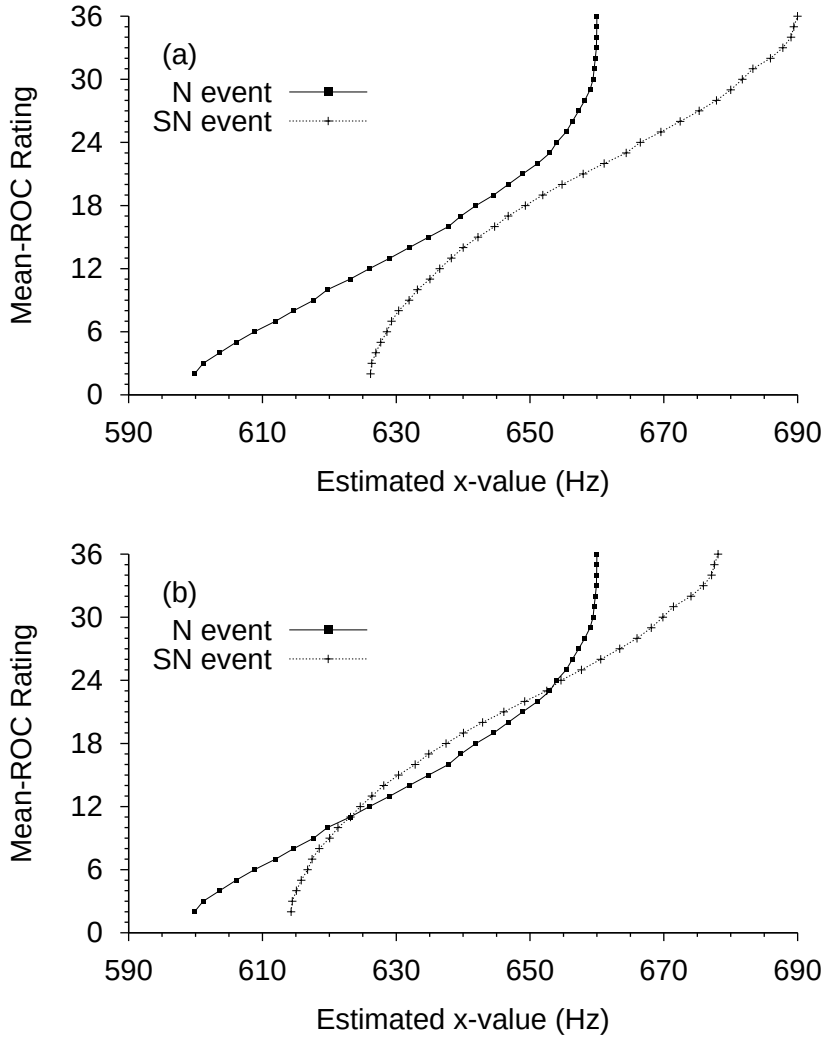


FIGURE 4.3: Transfer functions estimated from the mean ROC curve, assuming inappropriate uniform models. The $N$ event function (solid line) is the same in both panels, and assumes $X_{\mathrm{N}}$ ranges from 595 Hz to 660 Hz (Equation 4.3). (a) Assuming $X_{\mathrm{SN}}$ ranges from 625 Hz to 690 Hz (Equation 4.4). (b) Assuming $X_{\mathrm{SN}}$ ranges from 613 Hz to 678 Hz (Equation 4.5).

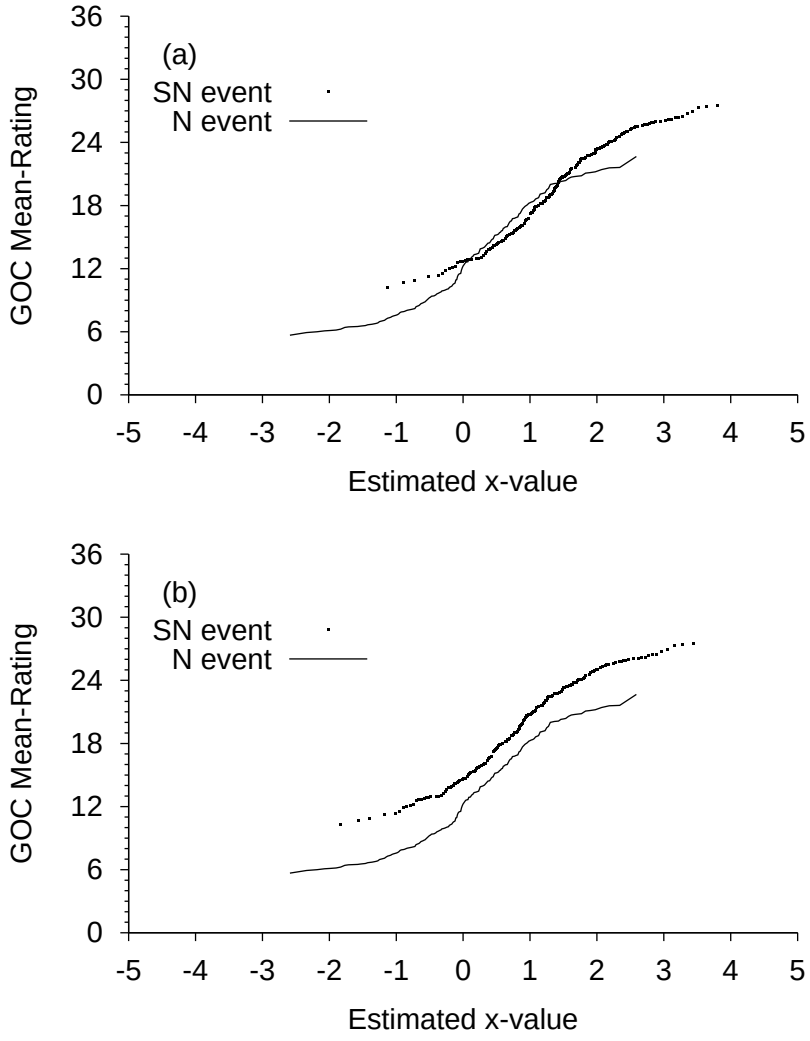


FIGURE 4.4: Transfer functions estimated from the GOC curve, assuming inappropriate Gaussian models. (a) Assuming a Gaussian equal variance model with $d' = 1.4553$. (b) Assuming a Gaussian unequal variance model with $d_z = 0.9077$. The $N$ event function (lower curve in each panel) is the same in both panels.

and false alarm rates match theoretical hit and false alarm rates). The two functions in Figure 4.4(b) do not intersect, reflecting the fact that the theoretical fit to the mean ROC curve does not intersect with the GOC curve. Although the $N$ and $SN$ functions are similar in form in both cases,[4] no possible horizontal shift would make the two functions match each other. A horizontal *and* vertical shift might match up the two functions, but a vertical shift implies selectively changing the rating distribution for one event while leaving the other untouched. Any transform applied to $R$ could not do this selectively for one event only and not the other. The transfer functions show that the Gaussian models are not appropriate to the GOC curve (which is, of course, already apparent in the ROC space).

In summary, the two event-conditional transfer functions that can be estimated from a data set are not consistent when the theoretical ROC curve resulting from an assumed model does not match an empirical ROC curve or GOC curve. The converse was already shown in Section 4.2. When the theoretical ROC curve used to estimate a transfer function was a good fit to the empirical curve, the event-conditional transfer functions were consistent with each other, for this particular data set.

## 4.5 Transfer functions estimated from cumulative distribution functions

Another way of estimating a transfer function is by equating the empirical and theoretical cumulative distribution functions (c.d.f.'s) of $R$ and $X$. This requires cumulating probability and proportions from below, whereas using hit and false alarm rates requires cumulating from above. If the distributions on both $R$ and $X$ are continuous, then the transfer functions are the same using either direction of cumulation. If the distributions on $R$ are discrete (while still assuming that $X$ is continuous), then the transfer functions estimated by cumulating from above and from below are systematically different from each other. The estimated function provided by each approach is as valid as the other, which implies that there must be some error in estimation, since the same function is supposed to be estimated by each approach. Equations are presented that show how the differences arise, and that they can be minimised by increasing the number of rating categories, or by increasing the number of replications when using GOC analysis. Transfer functions calculated by cumulating from below are estimated for Taylor et al.'s (1991) experiment, and are compared with the functions calculated by cumulating from above (Figures 4.1

[4] The $N$ distribution is the same for both Figures 4.4(a) and 4.4(b). Compared to the transfer function for the $SN$ function in Figure 4.4(a), the $SN$ function in Figure 4.4(b) is shifted to the left (so that $\mu_{SN} = 0.94$ rather than 1.46) and is horizontally stretched by the factor 1.07 (which is the standard deviation of the assumed $SN$ distribution).

and 4.2). The differences between the two estimated functions are small to negligible for this experiment, but could be much larger for other experiments.

If it is the case that the distributions on both the decision axis and the rating scale are continuous then $P(X \leqslant x) = 1 - P(X \geqslant x)$ and $P(R \leqslant r) = 1 - P(R \geqslant r)$, since $P(X = x) = 0$ and $P(R = r) = 0$. This implies that the same transfer function is estimated by cumulating from above as by cumulating from below.[5] A rating scale is truly continuous only in a theory or in a model. In practice, however, rating data are discrete, even if based on a continuous rating scale. Consequently,

$$P(R \leqslant r) \quad \neq \quad 1 - P(R \geqslant r)$$

holds in practice, since $P(R = r)$ may be non-zero.

For a given rating value $r$, let $x_1$ be the $x$-value estimated by equating theoretical and empirical hit rate and false alarm rate values (Equations 4.1 and 4.2). For the same rating value $r$, let $x_2$ be the $x$-value estimated by equating the c.d.f.'s of $R$ and $X$, $F_R(r)$ and $F_X(x)$ respectively. Here, $x_2$ is found via the equality

$$\begin{aligned} F_X(x_2) &= P(X \leqslant x_2) \\ &= P(R \leqslant r) \\ &= F_R(r). \end{aligned} \tag{4.6}$$

Applying the inverse c.d.f. to both sides of Equation 4.6, and considering the event-conditional forms, then

$$x_2 \quad = \quad F_X^{-1}(\, F_R(r|N)\,|N) \tag{4.7}$$

and also

$$x_2 \quad = \quad F_X^{-1}(\, F_R(r|SN)\,|SN)\,. \tag{4.8}$$

In practice, $x_1 \neq x_2$, because $P(R \leqslant r) \neq 1 - P(R \geqslant r)$, and $R$ is discrete. Hence the transfer function estimated using cumulation from above is different from the function estimated using cumulation from below (Equations 4.1 and 4.2, and 4.7 and 4.8, respectively).

When estimating transfer functions by cumulating either from above or from below, cumulative probability values of 0 or 1 should not be used. This is because their associated $x$-values are *always* the bounds of the domain of each of the assumed event-conditional distributions on $X$, and these bounds are independent of the observer and the observer's decisions.

Figures 4.5 and 4.6 both show transfer functions estimated by cumulation from above compared to transfer functions estimated by cumulation from below. Figure 4.5 derives

[5]For convenience, the event-conditional transfer functions are assumed to be consistent, and event-

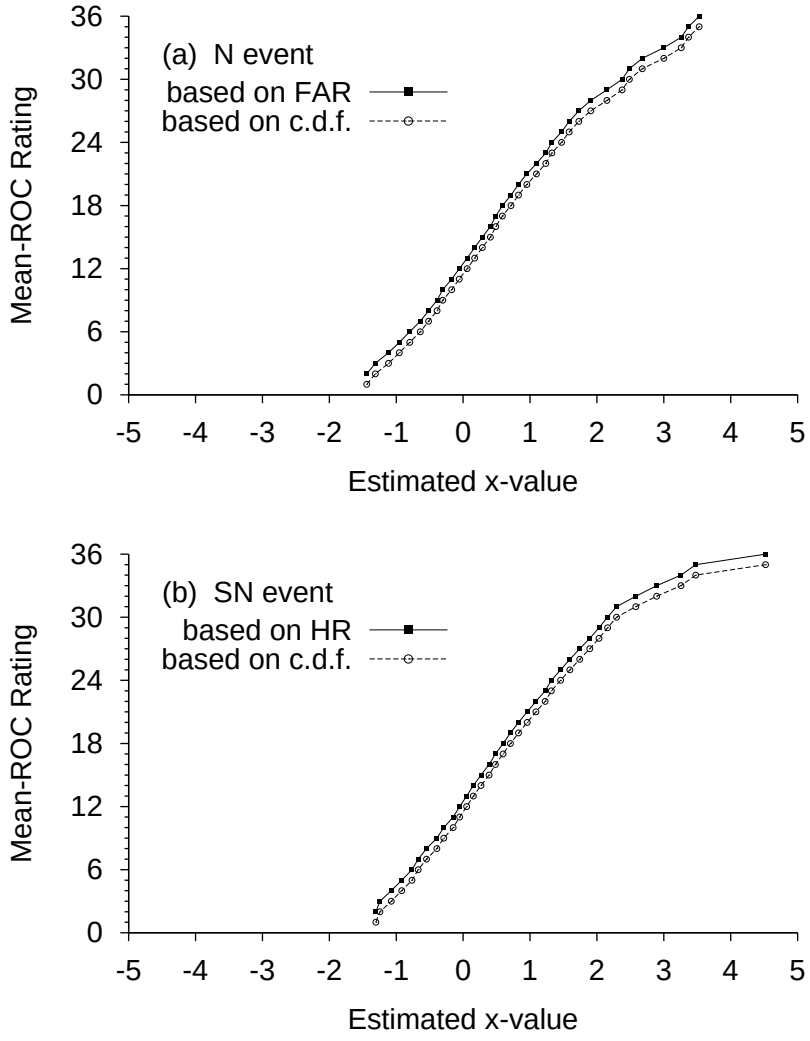


FIGURE 4.5: Transfer functions estimated from the mean ROC curve, assuming Gaussian unequal variance distributions. The functions were estimated from hit and false alarm rates (upper series in each panel) and from cumulative distribution functions (lower series in each panel). (a) Based on the $N$ event only. (b) Based on the $SN$ event only.

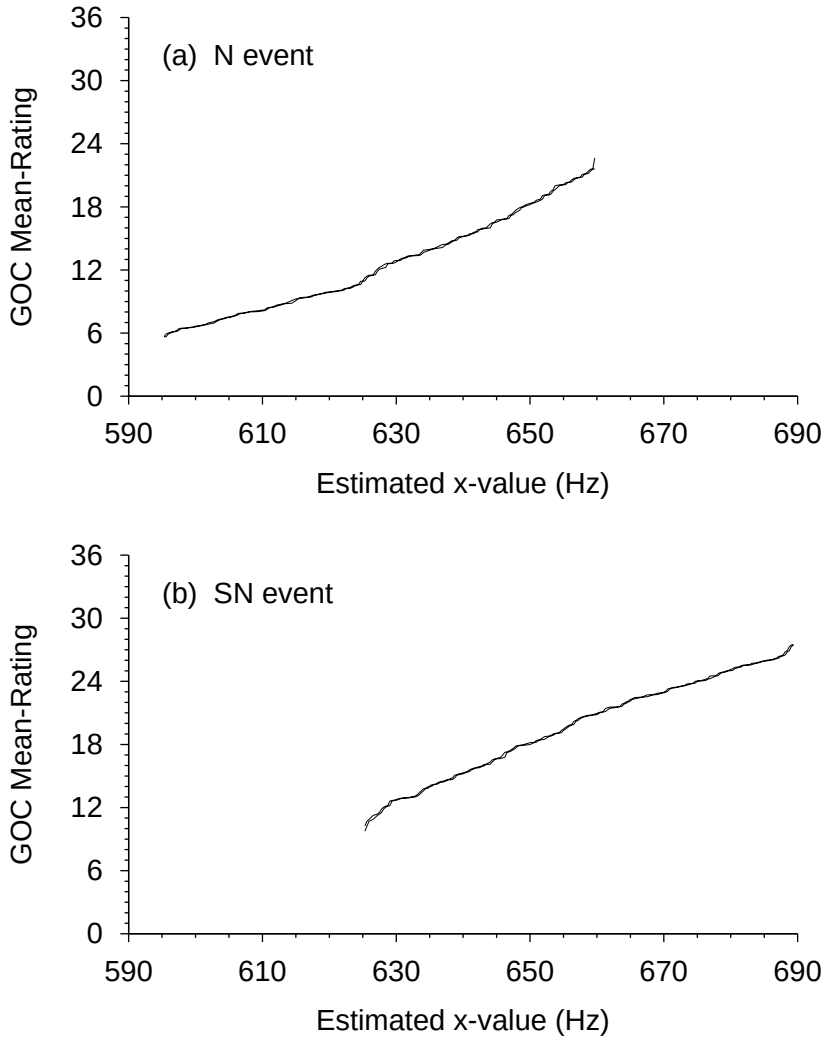


FIGURE 4.6: Transfer functions estimated from the GOC curve, assuming continuous uniform distributions. The functions were estimated from hit and false alarm rates (upper series in each panel) and from cumulative distribution functions (lower series in each panel). (a) Based on the $N$ event only. (b) Based on the $SN$ event only. (There are two data series in each panel, but the difference between them is negligible.)

from the mean ROC curve, under the assumption of the Gaussian unequal variance model with $d_z = 0.9077$, whereas Figure 4.6 derives from the GOC curve under the assumption of the continuous uniform model fit to the GOC curve. In both figures, the functions estimated via c.d.f.'s are vertically shifted down from the functions estimated via hit and false alarm rates. The shift is small, but constant, for the mean ROC functions in Figure 4.5, but hardly visible for the GOC functions in Figure 4.6. The upper and lower functions in Figures 4.5 and 4.6 place bounds on where the underlying transfer function may lie (assuming that it exists). If the gap between functions is large, then many possible functions may be supposed, whereas if the gap is small, then the constraints on the hypothesised function are much greater.

The vertical shift reflects the size of rating increments, or resolution of the rating scale. It is much smaller for a GOC curve, which is based on average ratings, than for a mean ROC curve, which is based on the original ratings. The vertical shift depends on which rating categories are used,[6] and how the rating scale is scaled. Not all categories are necessarily used, particularly in GOC analysis or when a rating scale contains a large number of possible points.

Let $r_k$ and $r_{k+1}$ be the values of the $k^{th}$ and $(k+1)^{th}$ rating categories that are used. Since $R$ is discrete, then

$$P(R \leqslant r_k) \quad = \quad 1 - P(R \geqslant r_{k+1}). \tag{4.9}$$

Assume that a transfer function has been estimated via c.d.f.'s and that $X$ is continuous. For a given rating value, $r_k$, let $x_k$ be such that

$$P(R \leqslant r_k) \quad = \quad P(X \leqslant x_k). \tag{4.10}$$

Since $X$ is continuous, then

$$P(X \leqslant x_k) \quad = \quad 1 - P(X \geqslant x_k)$$

which, together with Equation 4.9 implies that

$$P(R \geqslant r_{k+1}) \quad = \quad P(X \geqslant x_k). \tag{4.11}$$

Equation 4.10 states that $x_k$ is the estimated $x$-value for $r_k$ when using cumulation from below, but Equation 4.11 shows that $x_k$ is also the estimated $x$-value for $r_{k+1}$ when using cumulation from above. Hence two different $r$-values result in the same estimated $x$-value when $R$ is discrete and $X$ is continuous.

For the $k^{th}$ rating, the difference between the upper and lower estimated transfer

conditional notation is omitted where possible.

[6] This refers to categories on the original rating scale, if dealing with an ROC or mean ROC curve, and to mean-rating (or sum-of-ratings) categories, if dealing with a GOC curve.

functions is $r_{k+1} - r_k$. This difference does not depend on the assumed distributions on $X$, nor on the cumulative probabilities on $R$, but it does depend on the scaling of the rating scale and the how an observer uses a rating scale, particularly which ratings (or mean-ratings) are not used.

For the mean ROC curve, or even just a single ROC curve, there are relatively few potential categories, but most if not all of them are typically used. The rating categories for the data set were coded as successive integers. This implies that for ROC and mean ROC data, $r_{k+1} - r_k = 1$ for all $k$, and hence the functions in Figure 4.5 appear shifted by a constant. If the ratings were coded using squares of integers, for example, the discrepancy between the pairs of transfer functions in Figure 4.5 would diverge as a function of $x$ (since $r_{k+1} - r_k$ would diverges as a function of $r$).

With GOC curves, there are many potential mean-rating categories, and not all of them are usually used. However, the difference between successive mean-ratings that *are* used is typically much smaller than the difference between successive categories on the original rating scale. In contrast to Figure 4.5, the transfer functions estimated from GOC data in Figure 4.6 are very much closer together, and a very close inspection of Figure 4.6 reveals that the vertical shift does in fact vary over the length of the function.

Potential problems arise in estimating a transfer function when the number of categories is small, because the vertical shift can be relatively large. For example, if the ratings are coded as integers, from 1 to $q$, and a transfer function is estimated based on an ROC curve or a mean ROC curve, then the vertical shift is a minimum of $1/q$ of the range of the rating scale. If $q$ is small (less than 20, say), as is most often the case in psychophysics, then the vertical gap between estimated transfer functions is large. Hence, there may be much uncertainty about the location, and possible shape, of any the underlying transfer function. In Figure 4.5, the gap is reasonably small because $q = 36$ in Taylor et al.'s (1991) experiment. As Figure 4.5 shows, the gap can be made virtually non-existent if multiple replications are run.

**Summary.** Four transfer functions can be estimated by equating any of four cumulative proportions on the rating scale with cumulative probabilities on the decision axis. The quantities are the hit rate, the false alarm rate; the c.d.f. conditional on *SN*, and the c.d.f. conditional on *N*. The first two quantities involve cumulation of probability from above, whereas the last two quantities involve cumulation from below. Ideally, the estimated transfer functions based on all four quantities would be identical, since, they presumably estimate the same underlying function. In practice, a set of ratings is discrete, and the functions estimated by cumulation from above must be different from the functions estimated by cumulation from below. The amount of discrepancy depends on the resolution and scaling of the rating scale and, for GOC data, also on the number of replications run. This provides a means of determining the number of categories required to achieve a desired resolution.

## 4.6 Transfer functions based on discrete decision axes

The estimation of a transfer function is more complicated when the decision axis is discrete because there may not be any $x$-values and $r$-values for which cumulative probabilities and proportions are exactly equal on the two axes. As well as that, the cumulative probability functions on $X$ that are needed for the estimation (Equations 4.1, 4.2, 4.7 and 4.8) do not have unique inverses, which implies that there are potentially many ratings that result in the same estimated $x$-value.

Transfer functions from GOC data were estimated based on the assumption of *discrete* theoretical uniform distributions that were shown in Figure 2.2. The discrete distributions are the appropriate model for Taylor et al.'s (1991) experiment, rather than the continuous uniform approximation used to this point. The transfer functions based on discrete $X$, and using cumulation from above, are presented in Figure 4.7, along with the functions based on the continuous uniform model from Figure 4.1. The hit and false alarm rates of the continuous model were equal to those of the discrete model at the points where probability in the discrete model is massed. This is reflected in each panel of Figure 4.7 at the points where the two functions meet. If the continuous approximation were based on equal c.d.f. values (instead of hit and false alarm rates), the function based on continuous $X$ (dashed lines in Figure 4.7) would be shifted to the left by 5 Hz, and would match the top of each vertical step, rather than the bottom. A continuous approximation to any discrete $X$ provides a means to interpolate the transfer function based on the discrete decision axis.

The stairstep functions in Figure 4.7 reflects that the transform from $R$ to $X$ was many-to-one. In such a case, some convention must be adopted. The convention adopted for Figure 4.7 was that for a given empirical hit rate, $HR_R(r)$, the associated $x$-value that is paired with $r$ is the *largest* value of (discrete) $X$ such that

$$HR_X(x+5) < HR_R(r) \leqslant HR_X(x). \tag{4.12}$$

Similarly, for a given empirical false alarm rate, $FAR_R(r)$, $x$ is such that

$$FAR_X(x+5) < FAR_R(r) \leqslant FAR_X(x). \tag{4.13}$$

For example, for the *SN* event, $HR_X(685) = \frac{1}{13} \simeq 0.077$, and $HR_X(680) = \frac{2}{13} \simeq 0.154$. For any $r$ such that $0.077 < HR_R(r) \leqslant 0.154$, the estimated $x$-value is equal to 680 Hz. If $HR_R(r)$ is equal to 0.077, then the estimated $x$-value is 685 Hz.

The theoretical distributions were known to be discrete for Taylor et al.'s (1991) experiment. In more substantive experiments, the appropriate theoretical distributions are usually unknown. Continuous models would probably be assumed for such experiments, unless there is good reason or evidence to assume a discrete model (for example, if a neural counting model is used (e.g. McGill & Teich, 1991), or if a GOC curve shows distinct clus-

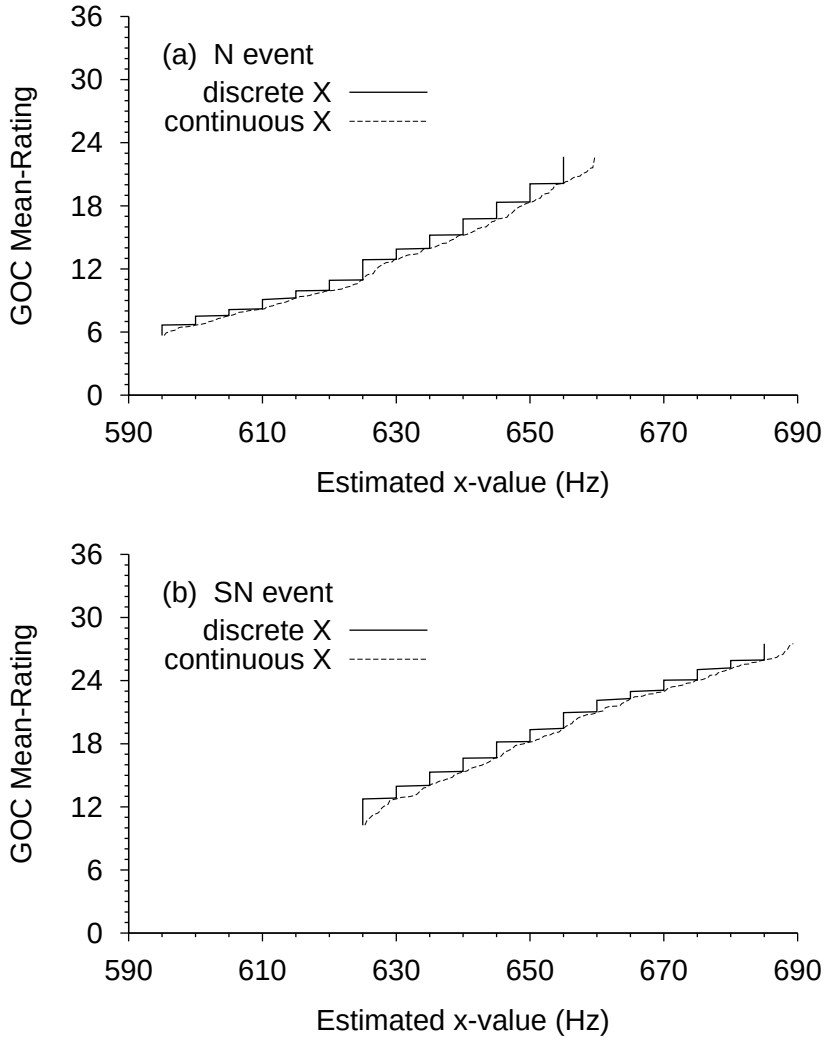


FIGURE 4.7: Estimated transfer functions based on the GOC curve, assuming discrete and continuous uniform models. Stairstep functions are based on discrete theoretical uniform distributions, whereas the smoother increasing functions are based on the continuous uniform model (from Figure 4.1). (a) Based on the $N$ event only. (b) Based on the $SN$ event only.

terings of points in the ROC space). Continuous models are more convenient than discrete models because the inverse cumulative functions for continuous can provide a one-to-one function. Discrete models result in many-to-one inverse cumulative functions and are not as simple to deal with.

## 4.7 Summary

A psychophysical transfer function is the s.m.i. relationship between a theoretical decision axis, $X$, and an empirical rating scale, $R$. Transfer functions can be estimated by fitting a theoretical ROC curve to an empirical GOC or ROC curve, and equating the criteria on $X$ and on $R$ that result in the same hit rate, or in the same false alarm rate. Transfer functions can also be estimated by equating theoretical and empirical c.d.f.'s.

Transfer functions have event-conditional forms. If the theoretical ROC curve used to estimate transfer functions provides a good fit to an empirical GOC or ROC curve, then the event-conditional functions are consistent with each other, otherwise they are not.

Systematic discrepancies occur between estimated transfer functions based on hit and false alarm rates, and those based on c.d.f.'s. The discrepancies are a practical consequence of analysing discrete rating data. The discrepancies are large when the resolution on the rating scale is low. The discrepancies are small when the resolution on the rating scale is high, and when GOC analysis is applied to an already high-resolution rating scale, then the discrepancies are negligible.

Transfer functions were estimated for Taylor et al.'s (1991) continuous rating scale experiment, based on the 24-replication GOC curve, assuming continuous uniform distributions, and based on the mean ROC curve, assuming Gaussian distributions. In both cases, ratings were reasonably linear with decision axis values. The transfer function based on the GOC curve showed that the rating slider position was, on average, a linear function of frequency. Interpretation of the Gaussian decision axis for the mean ROC curve was uncertain, because it was not clear how the decision axis related to frequency.

If a decision axis is discrete, then there are unavoidable problems in the estimating a transfer function, because the inverse function may be many-to-one. A continuous approximation to the discrete distributions is a practical solution, although exactly how the approximation is defined determines the resulting transfer function.

Transfer functions can be used to estimate unique noise parameters. A set of ratings may be converted into a set of estimated $x$-values lying on a unique-noise-affected decision axis, and statistics of the $x$-values may be calculated. For Taylor et al.'s (1991) experiment, the unique-to-common noise variance ratio was estimated to be approximately 2.7, which indicated that a relatively high amount of unique noise was present.

Since any arbitrary, order-preserving, s.m.i. rescaling of a given decision axis, $X$, produces the same theoretical ROC curve, any and all s.m.i.-related axes could be used as the basis for estimating a transfer function (if the nature of the decision axis is not known). Any such rescaling of $X$ would also transform unique noise as well as common noise. If the unique noise on one decision axis is of the same form and fixed variance, $\sigma_u^2$, for all values of the decision axis prior to such a transform, the unique noise is unlikely to have these properties after the transform. This calls into question the concept of the ratio of unique to common noise, $k$, if the unique noise variance changes as a function of $x$.

# Chapter 5

# The theory of GOC analysis

Group operating characteristic analysis is effective in improving performance in discrimination tasks by removing the effects of unique noise (Watson, 1963; Metz & Shen, 1992) and in recovering known theoretical ROC curves (Taylor et al., 1991; Lapsley Miller et al., 1998). Given enough replications, even very complicated ROC curves can be recovered (Taylor, 1984; Taylor et al., 1991). This chapter presents a theory of GOC analysis that explains the statistical properties that are necessary in order for GOC analysis to work.

## 5.1 Models of unique-noise-affected observers as analogies for GOC analysis

Unique noise may affect a discrimination task at any stage of the decision process. It may be pre-observer input such as extra environmental noise (Ronken, 1969; Tanner & Sorkin, 1970), it may be internal noise, for example, criterion variability (Wickelgren, 1968; McNicol, 1972), or it might even be post-observer error, such as error due to partitioning a continuous rating scale for data analysis. Models of inconsistent observers incorporate unique noise at many different stages in the decision process (Taylor, 1984; Durlach et al., 1986). The simplest and most common type of theory or model involves additive Gaussian unique noise on a decision axis (Swets et al., 1959; Siegel, 1979; Metz & Shen, 1992; Richards & Zhu, 1994; Sorkin & Dai, 1994). The mathematical benefit of such models is that the decision process can be expressed in terms of operations on random variables (Section 2.1.2). If both the unique and common noise are assumed to be Gaussian, then their additive mixture is also Gaussian, and performance can be described in terms of familiar and well-established theory. There are a wide variety of models of unique-noise-affected observers that could be used to simulate unique-noise-affected data. While GOC analysis can be applied to such data, the models themselves do not explain how GOC analysis works. There is no existing theory of GOC analysis.

The first attempt to model GOC analysis was made by Watson (1963), following a suggestion by Egan. Watson (1963) extended Tanner's dice game (Swets et al., 1961; Green & Swets, 1974) to show how GOC analysis may work. The original dice game involved a fundamental detection problem in which evidence values were generated on each trial by throwing dice and summing the results. One of two events (say *SN* or *N*) occurs during a trial. If the *SN* event occurs, a constant is added to the dice sum. The resulting sum forms an evidence value, and an observer must decide which event occurred, based on the evidence. Two overlapping event-conditional evidence distributions are generated. In the original dice game, these are of the same form but have different means for each event.

Watson's (1963) extension of the dice game involved multiple observers, in which each observer is associated with an additional, separate unique noise die. On each trial, a value is generated like in the original dice game, which represents common noise in the extended dice game. For each observer, the unique noise die for the observer is also thrown, and the result added to the common noise value. The evidence presented to each observer is the sum of the common noise value, which is the same for all observers, plus the unique noise value, which is individual to each observer. The observers' task is still the same, to discriminate between events, but the evidence is not necessarily the same across observers (although it may be if two observers happen to get the same unique noise value).

After describing the dice game, Watson (1963) discussed summing (or averaging) unique-and-common-noise-mixed values across observers. Consequently, unique noise is averaged out and removed by this analysis, while common noise remains. Although the extended dice game provides an easily understood analogy for how unique noise can influence decisions, by adding noise to evidence values, Watson's analysis is not equivalent to GOC analysis. The extended dice game, and many similar models, operate by averaging evidence values on a decision axis, whereas GOC analysis operates by averaging ratings on a rating scale. These are not equivalent, because unless an experiment is highly transparent, such as a dice game, experimenters do not have free access to an observer's evidence values. Instead, experimenters can only deal with rating scale data (including binary-decision data).

Two recent statistical models of unique-noise-affected observers have been proposed by Metz and Shen (1992) and by Sorkin and Dai (1994). Although the models are more detailed than in the dice game, they both share the same fundamental emphasis as Watson's (1963) analogy—the models remove unique noise on a decision axis rather than on a rating scale. Neither model deals with the problem of how individual observers transform input to produce ratings, and the consequences for the removal of unique noise.

The questions raised by existing models of observer inconsistency, plus the material presented in earlier chapters here, suggest that there are three interrelated topics that should be incorporated within any theory of GOC analysis. These are:

1. the relationship between unique-noise-affected evidence values and unique-noise-affected ratings, that is, the role of the transfer function in models of observer inconsistency (Chapter 4),

2. the potentially arbitrary scaling of a rating scale (Chapter 3), and how that affects the removal of unique noise, and

3. how the removal of unique noise on a rating scale relates to removal of unique noise on a decision axis.

All of these points were implied in the following passage by Metz and Shen (1992), who said that[1]

> Confidence ratings obtained in an image-reading experiment represent a reader's decision-variable outcomes after they have been categorized on a discrete, ordered scale in some unknown way. If we think of the confidence rating from each reading of a case as a crude *approximation* [original emphasis] to the continuous decision-variable outcome, an obvious scheme for combining ratings is simply to average them and then use the result as a decision variable for final image interpretation. The gains in accuracy that are obtained in this way clearly depend on two things, however: the number of categories in the rating scale (which determines the "fineness" or "coarseness" of the categorization process), and the way in which the category boundaries are distributed over the continuous decision-variable scale (which both determines the breadth of each category and effects a generally non-linear transformation of the decision-variable scale). (Metz & Shen, 1992, p. 72)

This passage shows that Metz and Shen (1992) recognised the role of the transfer function in discrimination tasks, the arbitrary scaling of a rating scale (in their case, a discrete

[1] Metz and Shen (1992) were interested in GOC analysis (or, in their terms, *mean-rating ROC analysis*) in the context of medical diagnostic tasks. In their context, an observer was a *reader* (of x-ray images), and the *reading of a case* was equivalent to the observation of a stimulus.

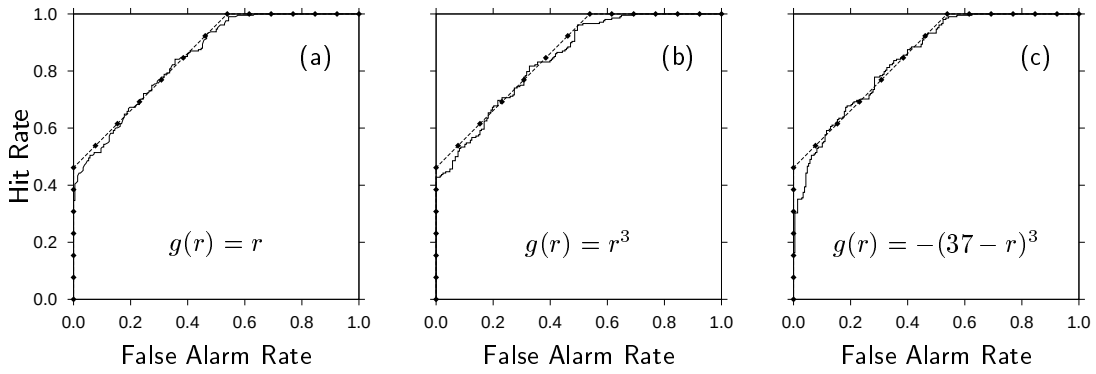


FIGURE 5.1: Transform-average GOC curves re-presented from Chapter 3. Each GOC curve was based on a transform of ratings $g(r)$, where $r \in \{1, 2 \ldots 36\}$. Panels (a), (b) and (c) were originally presented in Figures 3.1(a) 3.3(e), and 3.3(f), respectively.

scale), and the question of how unique noise is removed under the circumstances. Metz and Shen noted that there is a "generally non-linear transformation" from the decision axis onto a rating scale, and that the production of ratings on an ordered scale is achieved "in some unknown way," implying that the transfer function, and consequent scaling of a rating scale, are somewhat arbitrary, but ordinal.

Arbitrary but ordinal scaling of a rating scale can in fact be induced by transform-average GOC analysis, which was applied to Taylor et al.'s (1991) continuous rating scale data in Chapter 3. There, two rather contradictory patterns of results were apparent: (1) order-preserving strictly monotonic increasing (s.m.i.) transforms of rating scales definitely affected the resulting GOC curves, but (2) all of the GOC curves were consistent with the theoretical ROC curve over part of their lengths. The first result demonstrated that GOC results depended on the scaling of a rating scale. The second result indicated that GOC analysis was transform-invariant to some degree.

The last of the three points (noted on the preceding page) was that the theory of GOC analysis should explain how removing unique noise from rating data relates to removing unique noise on a decision axis. Metz and Shen (1992) did not have a solution to this. They recognised the problem though, and attempted to address it by suggesting that a confidence rating could be thought of as a "crude approximation" to a value on a decision axis. A problem with this is that decision axis values and random variables are not known, and possibly cannot be known, so experimenters cannot know when an approximation is appropriate or not.

It can be shown by example that a crude approximation of ratings to evidence values is not necessary in order for GOC analysis to work. This can be seen in transform-average GOC analysis of Taylor et al.'s (1991) data set, where each GOC curve is interpreted as based on a rescaling of a rating scale. Three of the 24-replication transform-average

GOC curves from Chapter 3 are reproduced in Figure 5.1, for ease of comparison. (Here, the original ratings were integers from 1 to 36, and $g(r)$ represents the rating transform.) Figure 5.1(a) is based on $g(r) = r$, and Figures 5.1(b) and (c) were based on $g(r) = r^3$ and $g(r) = -(37 - r)^3$, respectively.[2] All three GOC curves are very similar, especially in contrast to the single-replication ROC curves (Figure 2.3), which remain unaffected by different order-preserving rescalings of the rating scale. The rating distributions underlying each of the GOC curves are very different, however. The transform $g(r) = r^3$ is very convex when applied to the original rating scale (illustrated by the values: $g(1) = 1$, $g(5) = 125$, $g(9) = 729$, $g(18) = 5832$, and $g(36) = 46656$). Similarly, the transform $g(r) = -(37 - r)^3$ is very concave. If the rating data on the original scale was an approximation to distributions on the decision axis, as Metz and Shen (1992) suggest it could be, then the rating distributions on either of the transformed scales would certainly not be, because the new distributions would be *very* distorted. These examples in Figure 5.1(a) illustrate that if an approximation held for one scale, it would disappear under some other scale, yet very similar GOC curves can result. This demonstrates that rating distributions need not even be crude approximations to theoretical distributions of evidence values in order for GOC analysis to work. Many similar examples could also be drawn from the results in Chapter 3.

The potential distortion of distributions by transform averaging may or may not matter. A GOC curve is based on the order of a stimulus set according to the *mean rating* per stimulus, calculated for all stimuli in a stimulus set. This ordering is not affected by actual values of mean ratings, only their relative placement on the rating scale of interest. The transformations used in Figure 5.1 clearly do affect the ordering of the means, because the GOC curves are different. The ordering may not have changed by much, however, because the GOC curves are quite similar.

## 5.2 The equivalent statistical observer

A broad outline of an ESO was first introduced in Section 2.1.2 (Figure 2.1), as part of the definition of unique and common noise. The way in which unique and common noise contributed to decision making was left unspecified. This section presents an ESO that provides specific details omitted in Section 2.1.2.

Figure 5.2 outlines the key features of an ESO which consists of a common noise source, a unique noise source, a mixer, a transfer function and an optional quantising function. In a concrete application of the model, reproducible stimuli may contribute to most, if not all, of the common noise. The common noise source could be some type of black box

[2]The third transform listed here, $g(r) = -(37 - r)^3$, is strictly monotonic increasing, whereas the transform originally presented in Figure 3.3(f) in Chapter 3, $g(r) = (37 - r)^3$, was strictly monotonic decreasing. The two are equivalent with regards to transform-average GOC analysis, for reasons described in Section 3.1, and Figures 5.1(a) and 3.3(f) are identical.

discriminator or detector, like an ideal observer in models of unique-noise-free detectors (e.g. Jeffress, 1964; Green & Swets, 1974; Gilkey & Robinson, 1986). Sources of unique noise could involve both internal and external noise (Section 2.1.2).

Say the $j^{th}$ stimulus is presented to an observer on a given trial in the $i^{th}$ replication. Figure 5.2 represents the details of the $i^{th}$ branch of Figure 2.1, for the $j^{th}$ stimulus. The stimulus may influence either the common noise only, or the unique noise only, or both the common noise and unique noise. In the model, the $j^{th}$ stimulus is associated with a single common noise value, $x_j$, and a particular unique noise random variable, $U_j$. A unique noise value, $u_{ji}$, is sampled from $U_j$ on the $i^{th}$ replication. The values $x_j$ and $u_{ji}$ are then combined in the mixer to provide a *unique-noise-affected evidence value*, $y_{ji} = x_j \oplus u_{ji}$, where "$\oplus$" denotes some form of mixing, such as additive or multiplicative mixing. Formally, the mixer is a function[3] that takes on two arguments, $x_j$ and $u_{ji}$, and produces one value, $y_{ji}$. The value $y_{ji}$ is transformed by a continuous s.m.i. transfer function, $h$, that maps the decision axis onto a rating scale. This produces a rating value, $r_{ji}$, on a continuous rating scale. When modelling a discrete rating scale, $r_{ji}$ is converted by a monotonic increasing step function, $\Lambda$, into a value, $q_{ji}$, which lies on the discrete rating scale. The quantising function is optional, and is not required when modelling a continuous rating scale. In terms of random variables, $Y_j = x_j \oplus U_j$, then $R_j = h(Y_j)$, and $Q_j = \Lambda(R_j)$. In terms of sample values, $y_{ji} = x_j \oplus u_{ji}$, then $r_{ji} = h(y_{ji})$, and $q_{ji} = \Lambda(r_{ji})$. These random variables are defined for the $j^{th}$ stimulus only, and are distributed across replications (not stimuli). The value $x_j$ may be viewed as a sample value from some random variable, $X$, which has event-conditional forms $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$. The value of $x_j$ is constant across replications (hence the single subscript), although it could (and generally would) differ for different stimuli. The random variable $U_j$ describes the distribution of unique noise values, across replications, for the $j^{th}$ stimulus. All of the $U$-variables may be identically distributed for all stimuli, or each $U$-variable may take on a different form for each stimulus. In the ESO, unique noise may be independent of common noise, although it need not be. The set of $U$-variables may or may not be identically distributed for all stimuli, and $U_j$ may depend on the value of $x_j$ (e.g. $x_j$ could be a parameter in the distribution of $U_j$).

In general, $X$ refers to a common noise decision axis (i.e. a unique-noise-free decision axis), $Y$ refers to a unique-noise-affected decision axis, $R$ refers to a continuous rating scale, and $Q$ refers to a discrete rating scale. $U$-variables do not need to be defined on a separate axis, although they could be. At each stage in Figure 5.2, there is one random variable per stimulus, and an experimental stimulus set is associated with a family of random variables lying on each particular axis. Apart from $X$, the other random variables in the ESO are not event-conditional.

[3]The mixer is only characterised as an operator rather than a function, by analogy with the operators for addition and multiplication.

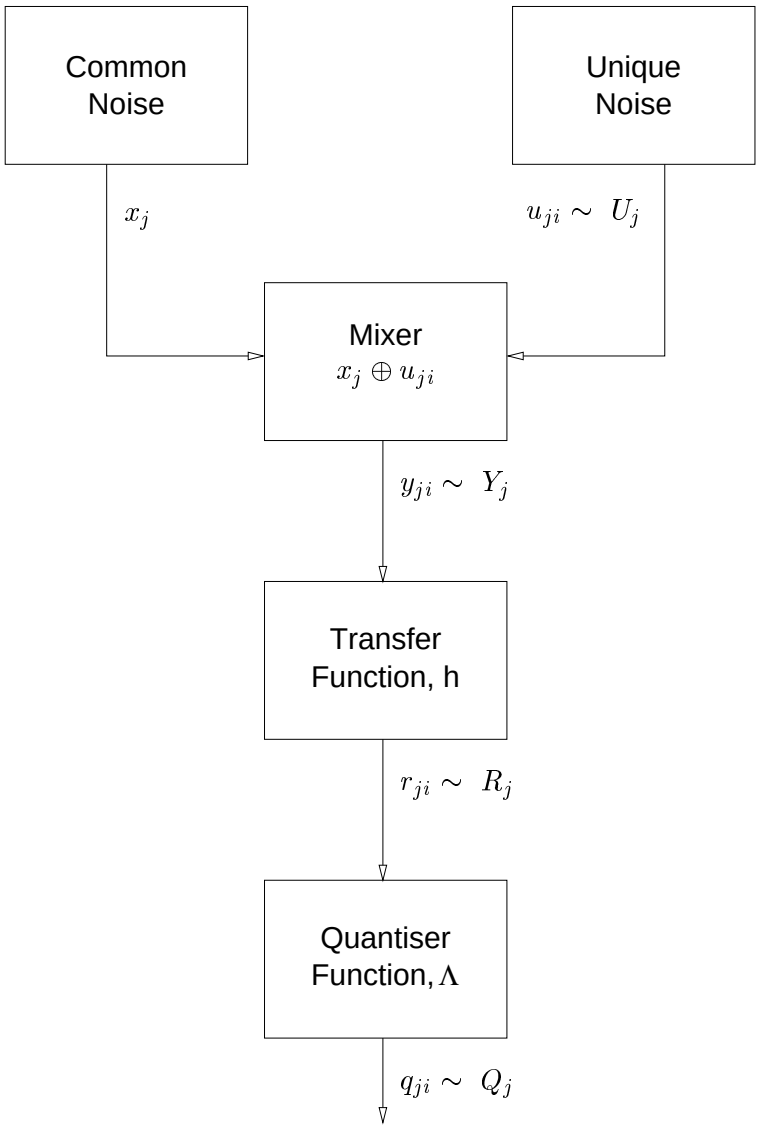


FIGURE 5.2: General model of a unique-noise-affected equivalent statistical observer, presented with the $j^{th}$ stimulus on the $i^{th}$ replication, showing the model components, sample values and random variables that are involved in the decision process. The sample values, $y_{ji}$, $r_{ji}$, and $q_{ji}$, are sampled from ($\sim$) random variables, $U_j$, $R_j$, and $Q_j$, respectively. The value, $x_j$, represents common noise, and is a constant across replications. $Y_j$ falls on a unique-noise-affected decision axis, $R_j$ falls on a continuous rating scale, and $Q_j$ falls on a discrete rating scale. The quantising function, $\Lambda$, is optional, and is only needed if modelling a discrete (rather than continuous) rating scale. Further details and interpretation are given in the text.

**Discrete rating scales.** Discrete rating scales are viewed as the result of a partition of a decision axis into adjacent intervals by applying a set of criteria on a decision axis (McNicol, 1972; Green & Swets, 1974). In Figure 5.2, the quantising function partitions $R$ into separate intervals, and assigns a value in the domain of $Q$ to each interval on $R$ in a manner that is monotonic increasing with $R$. This is equivalent to transforming $R$-variables into $Q$-variables by applying a monotonic increasing step function, $\Lambda$, to the domain of $R$. $Q$-variables are always discrete, regardless of the nature of the $Y$-variables or $R$-variables. A formal definition of $\Lambda$ is given in Section 5.4.3.

In principle, the quantising function is optional, and is required only if modelling a discrete rating scale, rather than a continuous rating scale. In practice, all rating scales are analysed as discrete scales (Section 1.3.1), partly because a continuous data set cannot be generated in practice. However, a rating scale may be *modelled* either as a continuous scale on $R$, without the need for a quantising function, or as a discrete scale on $Q$, depending on the details of the observer and experiment being modelled. Since the transfer function is continuous, then whether the random $R$-variables are continuous or discrete depends only on the nature of the $Y$-variables, regardless of whether there is a quantising function or not.

Together, $\Lambda$ and $h$ act like a single monotonic increasing step function, $\Lambda_0(y) = \Lambda(h(y))$, so that $Q_j = \Lambda_0(Y_j)$ is the same random variable as achieved by separately transforming $Y_j$ to $R_j$, and then $R_j$ to $Q_j$. Any rescaling of $Y$ by $h$ can be subsumed within $\Lambda_0$, obviating the need for a separate intermediate transform, $h$, or rating scale, $R$. Keeping $h$ separate from $\Lambda$ may be useful. If a continuous rating scale is used, the scaling of $R$, and the ultimate form of $h$, are determined jointly by the observer and by the experimenter. For example, $h$ may represent the mapping from an observer's decision axis, $Y$, onto the position of a rating slider, represented by $R$. $\Lambda$ may then represent the partition of the slider scale and coding into a data set, represented by $Q$. Slider partitioning and coding (i.e. the mapping from $R$ to $Q$) is determined by an experimenter or experimental equipment, rather than an observer. For convenience, both $h$ and $\Lambda$ are treated as properties of the observer in Figure 5.2.

If modelling a discrete rating scale experiment, the quantising function could be positioned before the transfer function instead of after it, so that $\Lambda$ represents a step function transform of $Y$ onto $Q$, and $h$ represents an s.m.i. rescaling of $Q$. This does not affect the results that follow.

Whether the rating scale is continuous or discrete does not fundamentally affect the theory of GOC analysis, and so the theory is initially presented without regard to the role of a quantising function. Discrete ratings in the theory of GOC analysis are covered in Section 5.4.3.

**Assumptions in the theory of GOC analysis**

The assumptions underlying the theory of GOC analysis are represented by Figure 5.2. The theory assumes an ESO of the particular form in Figure 5.2, with particular components, functions, samples and random variables. Other assumptions are given where needed (primarily about the existence of expected values). Otherwise, the theory of GOC analysis requires few assumptions, and it is easier to point out what is *not required*. The theory makes no assumptions about:

1. the type of random variables that describe either unique or common noise (i.e. discrete, continuous or mixed),
2. the distributional forms involved (e.g. normal, chi-squared, uniform, or unimodal or multimodal),
3. how the common noise mixes with unique noise (e.g. additively, multiplicatively, or otherwise),
4. whether unique noise and common noise ($U_j$ and $x_j$) are dependent or independent of each other,
5. whether or not the $U$-variables are independent across stimuli,
6. whether or not the $Y$-variables are independent across stimuli, and
7. whether or not the common noise decision axis, $X$ (from which $x_j$ derives), is either a likelihood ratio axis, or is s.m.i. with likelihood ratio.[4]

## 5.3 The theory of GOC analysis

GOC analysis operates by ordering a stimulus set according to mean rating per stimulus.[5] As more replications are added, the variance of the mean rating tends to zero, and the sample mean ratings tend towards expected values of their associated random variables. The order of a stimulus set therefore tends toward some asymptotic ordering, and the GOC curve based on the mean rating tends toward an *asymptotic GOC curve*. For the $j^{th}$ stimulus, the expected unique-noise-affected evidence value is $E(Y_j)$, the expected rating on $R$ is $E(R_j)$, and the expected rating on $Q$ is $E(Q_j)$. In order for unique noise removal from rating data to be equivalent to unique noise removal on a decision axis, the ordering of stimuli according to expected value must be the same if based on $E(Y_j)$, $E(R_j)$, or $E(Q_j)$. What is required is a statistical property that guarantees that the ordering

[4] GOC analysis works equally well with optimal and sub-optimal decision axes (Boven, 1976; Taylor, 1984; Taylor et al., 1991; Galvin et al., 1998).

[5] It is assumed throughout this chapter that the generalised GOC algorithm in Section 2.4.2 is used to calculate GOC curves, either based on sample mean values, or expected values of random variables.

of stimuli according to expected values is invariant with respect to the order-preserving transform of the decision axis onto the rating scale, and to the particular scaling of a rating scale. The key property that achieves this result is called *stochastic ordering*, also known as *stochastic dominance* (Findlay & Whitmore, 1978; Kroll & Levy, 1980; Whitt, 1988). It is important to note that stochastic ordering is a property of a *set* of random variables, rather than a property of any single random variable (in the same way that numerical ordering is a property of a set of numbers rather than any single number).

Mathematical results about stochastic ordering are developed in Appendix C. A number of definitions, theorems, and corollaries are presented there, along with their proofs, and commentary. The key results from Appendix C are restated here without proof,[6] and are applied to the ESO given in Figure 5.2. Within the context of the ESO, these results form the theory of GOC analysis.

Stochastic ordering may or may not hold for any given set of random variables. In Figure 5.2, stochastic ordering (if it occurs) is first encountered among the $Y$-variables on the unique-noise-affected decision axis, rather than among the $x$-values or $U$-variables. There are two types of stochastic ordering that are important for GOC analysis, *strict* and *non-strict* stochastic ordering. Strict ordering is relevant to s.m.i. transformations (e.g. from $Y$ to $R$), and non-strict ordering is relevant to step function transformations (e.g. from $R$ to $Q$).

Although the theory of GOC analysis is primarily concerned with sets of random variables, stochastic ordering also may apply to sets of sampled values. Ratings obtained in an experiment are viewed as values sampled from the random variable appropriate to each stimulus, such as $r_{ji}$ from $R_j$, or $q_{ji}$ from $Q_j$, on the $i^{th}$ replication. Whether two sample sets of values are stochastically ordered or not is determined by their sample cumulative distribution functions (c.d.f.'s), which would replace the c.d.f.'s of random variables in the definitions given in the following section.

The theory of GOC analysis does not make sense without an understanding of stochastic ordering. Prior to stating the theory, stochastic ordering is formally defined in Section 5.3.1, along with a description of patterns in the ROC space that correspond to stochastic ordering. This is followed by the central theorem of GOC analysis in Section 5.3.2. Section 5.4 then provides theorems and corollaries of stochastic ordering necessary for the theory of GOC analysis.

### 5.3.1 Definitions of stochastic ordering

The properties of strict and non-strict stochastic ordering are formally defined in Definitions 1 and 2, respectively, followed by a description of patterns in the ROC space that correspond to stochastic ordering.

[6] In the same order, and using the same numbering, as in the appendix.

**Definition 1** *Consider any two random variables, $Y_1$ and $Y_2$, which are either continuous, or discrete, or mixed, and which have respective cumulative distribution functions $F_{Y_1}$ and $F_{Y_2}$. $Y_1$ is stochastically less than $Y_2$ (denoted $Y_1 \overset{st}{<} Y_2$) if and only if $F_{Y_1}(t) \geqslant F_{Y_2}(t)\ \forall\, t \in \mathbb{R}$ and if $F_{Y_1}(t) > F_{Y_2}(t)$ for some non-zero interval on the real number line $\mathbb{R}$. The converse is $Y_1 \overset{st}{\not<} Y_2$, which means that $Y_1 \overset{st}{<} Y_2$ is not true.*

**Definition 2** *For two random variables, $Y_1$ and $Y_2$, $Y_1$ is stochastically less than or equal to $Y_2$ (denoted $Y_1 \overset{st}{\leqslant} Y_2$) if and only if $F_{Y_1}(t) \geqslant F_{Y_2}(t)\ \forall\, t \in \mathbb{R}$. The converse is $Y_1 \overset{st}{\not\leqslant} Y_2$, which means that $Y_1 \overset{st}{\leqslant} Y_2$ is not true.*

If $Y_1$ and $Y_2$ represent the $Y$-variables for any two stimuli, then Definition 1 states that $Y_1$ is *strictly* stochastically less than $Y_2$ ($Y_1 \overset{st}{<} Y_2$) if the c.d.f. of $Y_1$ is greater than or equal to the c.d.f. of $Y_2$, for all values of their argument, and if a strict inequality occurs over some portion of the real number line. An example showing the c.d.f.'s of two strictly ordered random variables is given in Figure C.1 in Appendix C. $Y_1$ and $Y_2$ are stochastically ordered, but not strictly ordered ($Y_1 \overset{st}{\leqslant} Y_2$) if the condition is dropped that $F_{Y_1}(t) > F_{Y_2}(t)$ necessarily holds over a range of $t \in \mathbb{R}$. For any particular pair of random variables, non-strict ordering implies that either $Y_1 = Y_2$, or $Y_1 \overset{st}{<} Y_2$, but not both.

Even though Definitions 1 and 2 are stated using the notation of $Y$-variables, the property is not restricted to $Y$-variables, and may apply (or not) to pairs of random variables either on $X$, $Y$, $R$, or $Q$.

For any two stimuli, stochastic ordering of their $Y$-variables, say $Y_1$ and $Y_2$, can be checked using a stimulus-pair ROC curve,[7] in which $Y_1$ takes the place of the $X_{\mathrm{N}}$ distribution, and $Y_2$ takes the place of the $X_{\mathrm{SN}}$ distribution. (Note that a stimulus-pair ROC curve is defined for only two individual stimuli out of a stimulus set. It is different from an experimental ROC curve, which is based on an entire stimulus set.) If $Y_1 \overset{st}{\leqslant} Y_2$, then the stimulus-pair ROC curve never lies below the chance line at any point. If $Y_1 \overset{st}{<} Y_2$, then the ROC curve never lies below the chance line, and it must lie above it at some point. Similarly, if $Y_1 \overset{st}{\geqslant} Y_2$, then the ROC curve never lies above the chance line, and if $Y_1 \overset{st}{>} Y_2$, then it must lie below the chance line at some point. $Y_1$ and $Y_2$ do not have to be s.m.i. with their likelihood ratio in order to be stochastically ordered, so there is no restriction about any monotonic decreasing slope of the stimulus-pair ROC curve.[8] If some portion of the stimulus-pair ROC curve lies above the chance line, *and* another portion lies below,

[7] Or, equivalently, an ordinal dominance graph (Bamber, 1975). Ordinal dominance graphs could also be called *stochastic dominance* curves, or *stochastic ordering* curves. Bamber (1975) showed how ordinal dominance graphs and ROC curves relate to one another—one is the rotation of the other by 180° in the ROC space around the central point, (0.5,0.5). An ordinal dominance graph may be converted into an ROC curve, and vice versa, without loss of information. ROC curves are preferred here, because of familiarity.

[8] If $Y_1$ and $Y_2$ were s.m.i. with their pairwise likelihood ratio, then the slope of the stimulus-pair ROC curve would have monotonic decreasing slope (Green & Swets, 1974; Egan, 1975).

then the random variables are not stochastically ordered. For example, if $Y_1$ and $Y_2$ are both Gaussian with *equal* variances, where the means, $\mu_1$ and $\mu_2$, are such that $\mu_1 < \mu_2$, then the stimulus-pair ROC curve lies entirely above the chance line, which implies that $Y_1 \overset{st}{<} Y_2$. If $\mu_1 > \mu_2$, then the stimulus-pair ROC curve lies entirely below the chance line, which implies that $Y_1 \overset{st}{>} Y_2$. If, on the other hand, $Y_1$ and $Y_2$ are both Gaussian with *unequal* variances (regardless of $\mu_1$ and $\mu_2$), then the stimulus-pair ROC curve crosses the chance line at some point (McNicol, 1972), in which case $Y_1 \overset{st}{\not<} Y_2$, $Y_1 \overset{st}{\not>} Y_2$ and $Y_1 \neq Y_2$.

Stochastic ordering of random variables sometimes follows the same form as the numerical ordering of quantities, but not always. If $Y_1 \overset{st}{<} Y_2$, then $Y_1 \overset{st}{\leqslant} Y_2$, for example, and if $Y_1 = Y_2$ then $Y_1 \overset{st}{\leqslant} Y_2$. However, $Y_1 \overset{st}{\not<} Y_2$ *does not* imply that $Y_1 \overset{st}{\geqslant} Y_2$. It is possible that $Y_1$ and $Y_2$ are such that $Y_1 \neq Y_2$, $Y_1 \overset{st}{\not<} Y_2$ and $Y_1 \overset{st}{\not>} Y_2$ all hold simultaneously. Further details of this and other rules of stochastic ordering are in Appendix C (following Definition 2).

Stochastic ordering is a transitive property that may apply to more than two random variables. Stochastically ordered sets of random variables are possible.

**Corollary 1** *(Transitivity) Let $Y_1$, $Y_2$ and $Y_3$ be any three random variables. If $Y_1 \overset{st}{<} Y_2$ and $Y_2 \overset{st}{<} Y_3$, then $Y_1 \overset{st}{<} Y_3$. If $Y_1$, $Y_2$ and $Y_3$ are such that either $Y_1 \overset{st}{<} Y_2 \overset{st}{\leqslant} Y_3$ or $Y_1 \overset{st}{\leqslant} Y_2 \overset{st}{<} Y_3$, then $Y_1 \overset{st}{<} Y_3$. If $Y_1 \overset{st}{\leqslant} Y_2 \overset{st}{\leqslant} Y_3$, then $Y_1 \overset{st}{\leqslant} Y_3$.*

### 5.3.2 The central theorem of GOC analysis

The theory of GOC analysis is a statement about the ESO outlined in Figure 5.2, and the random variables associated with it. The central theorem of GOC analysis is that

> If the $Y$-variables on a unique-noise-affected decision axis, $Y$, are stochastically ordered such that $Y_1 \overset{st}{<} Y_2 \overset{st}{<} Y_3 \overset{st}{<} \ldots$, then the continuous rating scale distributions are such that $R_1 \overset{st}{<} R_2 \overset{st}{<} R_3 \overset{st}{<} \ldots$ and the discrete rating scale distributions are such that $Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} Q_3 \overset{st}{\leqslant} \ldots$ Furthermore, if $Y_1 \overset{st}{<} Y_2 \overset{st}{<} Y_3 \overset{st}{<} \ldots$, then $E(Y_1) < E(Y_2) < E(Y_3) < \ldots$, $E(R_1) < E(R_2) < E(R_3) < \ldots$, and $E(Q_1) \leqslant E(Q_2) \leqslant E(Q_3) \leqslant \ldots$

The theorem implies that the ordering of a stimulus set according to expected rating value, $E(R_j)$ (for a continuous rating scale), or $E(Q_j)$ (for a discrete rating scale) are the same as the ordering according to expected values on $Y$, $E(Y_j)$, and that all of these orderings are determined by the stochastic ordering of the $Y$-variables. At the very least, there must be non-strict stochastic ordering between any successive pair of $Q$-variables ($Q_j$ and $Q_{j+1}$), and non-strict numerical ordering between their means ($E(Q_j)$ and $E(Q_{j+1})$). The ordering between successive pairs could be strict rather than non-strict for any particular instance, depending on how $\Lambda(r)$ partitions and transforms $R$,

that is, how $\Lambda(h(y))$ partitions and transforms $Y$. The central theorem holds *for all possible s.m.i. transfer functions, $h$, and all possible increasing step functions, $\Lambda$*, such that the expected values of the $R$-variables and $Q$-variables exist and are finite. The relationships among the $Y$, $R$ and $Q$-variables and their expectations may be summarised as

$$\begin{array}{ccccc} Y_1 \overset{st}{<} Y_2 \overset{st}{<} \dots & \Leftrightarrow & R_1 \overset{st}{<} R_2 \overset{st}{<} \dots & \Rightarrow & Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} \dots \\ \Downarrow & & \Downarrow & & \Downarrow \\ E(Y_1) < E(Y_2) < \dots & & E(R_1) < E(R_2) < \dots & & E(Q_1) \leqslant E(Q_2) \leqslant \dots \end{array}$$

(where arrows show the direction of implication).

The theoretical results that justify the central theorem of GOC analysis are given in the following sections. Note that the theorem is entirely predicated on the stochastic ordering of the $Y$-variables (i.e. *if* $Y_1 \overset{st}{<} Y_2 \overset{st}{<} Y_3 \overset{st}{<} \dots$) Stochastic ordering of $Y$-variables may or may not hold for any particular observer. The central theorem describes the consequences if it does hold. The implications when it does *not* hold are described in Section 5.4.2 and 5.6.

No mention has been made of the $x$-values or $U$-variables which are part of the ESO in Figure 5.2. This is deliberate, because the theory of GOC analysis can only recover the ordering—and hence ROC curve—based on a set of $Y$-variables. The consequences of this for existing models of unique noise are discussed in Section 5.6. Until then $x$-values and $U$-variables are not treated separately, but are explicitly incorporated within the set of $Y$-variables.

## 5.4 Stochastic ordering

There are three theorems in statistics about stochastic ordering that apply outside of the GOC context. These theorems (numbered 1 to 3), and their corollaries, contribute to the central theorem of GOC analysis. Proofs of the statistical theorems and corollaries are given in Appendix C.

Theorem 1 and its corollaries underlie the relationship between the $Y$-variables and the $R$-variables, and the link between strict stochastic ordering of a set of random variables and strict numerical ordering of their expected values. Theorem 2 shows that if two $Y$-variables are *not* stochastically ordered, then the expected values of the corresponding $R$-variables *depend* on the transfer function, $h$, and on the particular scaling of a rating scale. Theorem 1 shows that stochastic ordering is a *sufficient* condition for GOC analysis to be transform-invariant (for any transfer function, and any ordinal rescaling of a rating scale), whereas Theorem 2 shows that stochastic ordering is also a *necessary* condition. Theorem 3 and its corollaries describe the relationship between the $R$-variables and the $Q$-variables, and implications for discrete rating scales. The conditions under which the $Q$-variables are strictly ordered, or non-strictly ordered, are given in Section 5.4.3.

### 5.4.1 Stochastic ordering and continuous rating scales

A useful concept in the theory of GOC analysis is the *mutual domain*.

**Definition 3** *The domain of a random variable is the set of values of* $\mathbb{R}$ *for which the probability mass* or *density is non-zero. The mutual domain of a set of random variables is the smallest continuous interval on* $\mathbb{R}$ *containing all values of the union of the domains of the random variables.*

The mutual domain is not a standard mathematical concept. It is introduced for convenience, for situations when the same transform is applied to each of a large set of random variables (such as the set of $Y$-variables for an entire stimulus set). Rather than being concerned about bounds of the domain over which a transform must be defined (which could potentially consist of many disjoint intervals), it is simpler to define transforms over the smallest single continuous interval on which any of the random variables are defined. The mutual domain may be the entire real number line, $\mathbb{R}$, or only a small subset of $\mathbb{R}$.

**Theorem 1** *If* $Y_1$ *and* $Y_2$ *are any two continuous, mixed or discrete random variables whose expectations exist and are finite, then* $Y_1 \overset{st}{<} Y_2$ *implies that* $E(Y_1) < E(Y_2)$. *Furthermore, for any strictly monotonic increasing transform,* $h$, *defined over the mutual domain of* $Y_1$ *and* $Y_2$, *then* $Y_1 \overset{st}{<} Y_2$ *implies that* $h(Y_1) \overset{st}{<} h(Y_2)$ *and, consequently, that* $E(h(Y_1)) < E(h(Y_2))$, *if the expectations exist and are finite.*

The condition that the expectations exist and are finite is an important one, because there are random variables (e.g. Cauchy) for which no expectation exists, in which case, results involving expected values do not apply. Further details about the existence conditions of expectations are available in Appendix C (Definition 5). Note that although $Y_1 \overset{st}{<} Y_2$ implies $E(Y_1) < E(Y_2)$, the converse is not true. One counterexample is if $Y_1$ and $Y_2$ are Gaussian with unequal variances (from an illustration given in Section 5.3.1). In this counterexample, $E(Y_1) < E(Y_2)$, but $Y_1 \overset{st}{\not<} Y_2$.

Theorem 1 can be extended to cover multiple, nested transforms, which is shown in Corollary 2, and to a set of more than two random variables, which is shown in Corollaries 3 and 4.

**Corollary 2** *Let* $Y_1$ *and* $Y_2$ *be any two random variables whose expectations exist and are finite. Let* $h_1$, $h_2$, $h_3$ ... *be continuous, strictly monotonic increasing functions, where* $h_1$ *is defined on the mutual domain of* $Y_1$ *and* $Y_2$, $h_2$ *is defined on the mutual domain of* $h_1(Y_1)$ *and* $h_1(Y_2)$, $h_3$ *is defined on the mutual domain of* $h_2(h_1(Y_1))$ *and* $h_2(h_1(Y_2))$, *and so on. If* $Y_1 \overset{st}{<} Y_2$, *then*

$$E[\ldots h_3(h_2(h_1(Y_1)))] \quad < \quad E[\ldots h_3(h_2(h_1(Y_2)))],$$

*if the expectations exist and are finite.*

**Corollary 3** *If there is a stochastically ordered set of random variables,* $\{Y_1, Y_2, Y_3 \dots\}$ *such that* $Y_1 \overset{st}{<} Y_2 \overset{st}{<} Y_3 \overset{st}{<} \dots$, *then* $h(Y_1) \overset{st}{<} h(Y_2) \overset{st}{<} h(Y_3) \overset{st}{<} \dots$ *holds for any continuous s.m.i. transform* $h$ *defined over the mutual domain of all of the* $Y_j$.

**Corollary 4** *For any* $\{Y_1, Y_2, Y_3 \dots\}$, *defined and stochastically ordered as in Corollary 3, and for any continuous s.m.i. transform* $h$ *defined over the mutual domain of all the* $Y_j$, $E(Y_1) < E(Y_2) < E(Y_3) \dots$ *and* $E(h(Y_1)) < E(h(Y_2)) < E(h(Y_3)) \dots$ *The ordering of the expected values of both the untransformed and the transformed random variables follows the stochastic ordering of the Y-variables, if the expectations exist and are finite.*

**Corollary 5** *Corollaries 3 and 4 hold for any arbitrary combination of strict and non-strict inequalities in the ordering sequence, where the ordering is stochastic for the random variables and numerical for their expected values. For example, if* $Y_1 \overset{st}{<} Y_2 \overset{st}{\leqslant} Y_3 \overset{st}{<} \dots$, *then* $h(Y_1) \overset{st}{<} h(Y_2) \overset{st}{\leqslant} h(Y_3) \overset{st}{<} \dots$, *and consequently,* $E(Y_1) < E(Y_2) \leqslant E(Y_3) < \dots$ *and* $E(h(Y_1)) < E(h(Y_2)) \leqslant E(h(Y_3)) < \dots$

Corollary 3 shows that if the set of $Y$-variables is stochastically ordered, then the resulting set of $R$-variables ($R_j = h(Y_j)$) is also stochastically ordered, and the $R$-variables follow the same ordering as the $Y$-variables.[9] Corollary 4 shows that the same ordering extends to the expected values of the $Y$-variables, $E(Y_j)$, and to the expected values of the $R$-variables, $E(R_j)$. This is particularly relevant in the theory of GOC analysis, since GOC analysis is based on the ordering of a stimulus set based on mean ratings on $R$, rather than mean evidence values on $Y$.

Corollary 5 shows that non-strict ordering between a pair of $Y$-variables implies non-strict ordering between the expected values of the associated pair of $R$-variables. Non-strict ordering of $R$-variables implies that two stimuli could be tied according to expected rating value, but the fact of stochastic ordering (even if non-strict) implies the order of means on $R$ could not be reversed from the order of means on $Y$. The consequences of non-strict stochastic ordering are discussed in Section 5.6.

If the $Y$-variables are strictly stochastically ordered, then the asymptotic GOC curve based on $E(R_j)$, is identical to the theoretical curve based on $E(Y_j)$, which represents theoretical performance, once unique noise has been averaged out on $Y$. Other than the stipulations that the transfer function, $h$, is s.m.i., and that the expectations involved exist and are finite, these results hold regardless of the form of $h$. *The transfer function need not be known* in order for stochastic ordering to hold on $R$.

If the s.m.i. transfer function from $R_j$ to $Y_j$ is $R_j = h_1(Y_j)$ for all $j$, Corollary 2 shows that any subsequent s.m.i. transform of $R_j$ onto a new rating scale (defined by $h_2(R_j)$, for all $j$), still maintains the same stochastic ordering of the underlying $Y$-variables. Further

[9] The converse is also true, which can be shown by re-applying Corollary 3 to the $R$-variables, using the inverse function $h^{-1}$.

s.m.i. transforms of the rating scale (e.g. $h_3(h_2(R_j))\ldots$) also maintain the same stochastic ordering. This implies that if the set of $Y$-variables is stochastically ordered, then (from Corollary 4), *the ordering of a stimulus set according to expected rating value is the same regardless of the particular scaling of a rating scale*. This implies that transform-average GOC analysis results in the same asymptotic GOC curve.

**Summary**

With regards to the ESO in Figure 5.2, Theorem 1 and its corollaries showed that a strictly stochastically ordered family of $Y$-variables implies a strictly ordered family of $R$-variables (and vice versa). Strict stochastic ordering of random variables implies strict numerical ordering of expected values (but not vice versa). This demonstrates the first part of the central theorem of GOC analysis, and is summarised by

$$\begin{array}{ccc} Y_1 \overset{st}{<} Y_2 \overset{st}{<} \ldots & \Leftrightarrow & R_1 \overset{st}{<} R_2 \overset{st}{<} \ldots \\ \Downarrow & & \Downarrow \\ E(Y_1) < E(Y_2) < \ldots & & E(R_1) < E(R_2) < \ldots \end{array}$$

These results hold true, regardless of the specific transfer function, $h$, as long as $h$ is s.m.i. and defined on the mutual domain of the $Y$-variables, and given that the expectations involved exist and are finite. Furthermore, any number of s.m.i. rescalings of $R$, or of $Y$, do not affect these results.

### 5.4.2 Consequences if stochastic ordering does not hold

A major result of Theorem 1 and its corollaries is that the stochastic ordering of $Y$-variables is a *sufficient* condition for the numerical ordering of expected values, on $Y$ or on $R$, to follow the stochastic ordering under any s.m.i. transform. Theorem 2 shows that it is also a *necessary* condition.

**Theorem 2** *Let $h$, $Y_1$ and $Y_2$ be defined as for Theorem 1, such that $E(h(Y_1))$ and $E(h(Y_2))$ both exist and are finite. If $Y_1 \overset{st}{\not<} Y_2$, $Y_1 \overset{st}{\not>} Y_2$ and $Y_1 \neq Y_2$, then regardless of the order of $E(Y_1)$ and $E(Y_2)$, it is <u>always</u> possible to choose a strictly monotonic increasing transform, $h$, such that $E(h(Y_1))$ is either less than, greater than, or equal to $E(h(Y_2))$.*

Theorem 2 shows that if the $Y$-variables associated with two stimuli are not stochastically ordered, then the ordering of the stimuli according to mean rating, $E(R_1)$ and $E(R_2)$ (where $R_j = h(Y_j)$), is not independent of $h$. Not only that, but by a judicious choice of transform, it is always possible to order two stimuli as desired, or even to have them tied equal. Although the theorem is framed in terms of a transfer function, $h$, that rescales a decision axis, the same result also applies to the deliberate rescaling of a rating scale.

Theorem 2 implies that if stochastic ordering does not hold on $Y$, then the unique-noise-free, asymptotic GOC curve depends on the specific transfer function and scaling of a rating scale. If one of the two stimuli in the example was associated with the $SN$ event, and the other stimulus was associated with the $N$ event, then a reversal of means can alter the asymptotic GOC curve, because the hit and false alarm rate pairings that the stimuli contribute to can change.[10] Without stochastic ordering, expected unique-noise-free performance (such as area under the asymptotic GOC curve) could be manipulated and changed. The scaling of a rating scale can be changed once data has been collected, regardless of whether stochastic ordering applies or not. If stochastic ordering does not hold, then asymptotic performance depends, to some extent, on the details of data analysis rather than on the decisions that were made by the observer. In contrast, if stochastic ordering does hold, Theorem 1 shows that there is no possible s.m.i. transfer function or rescaling of a rating scale that could possibly change the order of expected values.

One example of the order reversal of means has already appeared (in Table 3.1 in Section 3.1.2). The numeric example showed that the order of two stimuli according to mean rating depended on the scaling of the rating scale. (Sample means were involved, and stochastic ordering did not hold between the sample c.d.f.'s). A more general example appears in the proof of Theorem 2 in Appendix C. The proof gives a particular transform that guarantees order-reversal for any pair of random variables that are not stochastically ordered.

### 5.4.3 Stochastic ordering and discrete rating scales

Section 5.4.1 provided key results that described a continuous s.m.i. transform of a unique-noise-affected decision axis, $Y$, onto a continuous rating scale, $R$, but more development is needed to understand GOC analysis of data on a discrete rating scale, $Q$ (which includes binary-decision data). In the ESO sketched in Figure 5.2, discrete ratings are produced by the quantising function, $Q = \Lambda(R)$, which is a monotonic increasing step function that converts a set of $R$-variables into a set of $Q$-variables. The $Q$-variables are always discrete, regardless of the nature of the $R$-variables, or the underlying $Y$-variables.

An emphasis is placed on the distinction between strict and non-strict stochastic ordering, when dealing with GOC curves based on discrete rating scales. This is because non-strict stochastic ordering allows the expected values of two successive members of the family $Q$-variables to be equal, that is, $Q_j \overset{st}{\leqslant} Q_{j+1}$ implies that $E(Q_j) = E(Q_{j+1})$ is possible, (although not guaranteed). If $Q_j \overset{st}{<} Q_{j+1}$, on the other hand, then $E(Q_j) < E(Q_{j+1})$

[10] Although whether or not hit and false alarm rates change for any specific pair of stimuli, and any specific transform, depends on the effect that the transform has on the expected ratings for the rest of the stimulus set. The same ordering of event-labels in the generalised GOC algorithm (Section 2.4.2) could occur based on a different ordering of mean ratings, which would not change the asymptotic GOC curve. It seems much more likely, though, that a change of one or many pairs of stimuli would alter the asymptotic GOC curve, because there are many more ways of disordering a particular event-label sequence than there are ways of maintaining it.

must hold (by Theorem 1, applied to $Q_j$ and $Q_{j+1}$), and so $E(Q_j) = E(Q_{j+1})$ is not possible. Preferably, tied expected values on $Q$ should be avoided if possible, because they lead to gaps between points in the GOC curve based on $Q$ that are not present in the GOC curve based on $R$ or on $Y$. If that is the case, then the curve based on $Q$ may only be an approximation to the curve based on $R$ or $Y$. On the other hand, if the $Q$-variables are all strictly ordered, following the ordering of $Y$-variables, then the GOC curve based on value of $E(Q_j)$ is identical to the curve based on values of $E(Y_j)$.

(An analogous effect to having tied expected values on $Q$ occurs in a unique-noise-free context based on a decision axis, $X$. A monotonic increasing step function transform may be used to transform $X$, in order to arrive at discrete ratings on $Q$. Bamber (1975, Figure 6), showed how an ROC curve based on a discrete rating scale is consistent with the theoretical ROC curve based on $X$, but that portions of the theoretical ROC curve may not be represented by the rating ROC curve, because the step function results in pooling or massing of probability.)

It was noted previously in Section 5.2 that the composite step function, $\Lambda_0(y) = \Lambda(h(y))$, could be applied directly to $Y$ without needing to work through an intermediate set of $R$-variables. Possible reasons for keeping $R$ and $Q$ separate were given in Section 5.2, mostly with regard to what is being modelled. The results that follow are based on a transform from $R$ to $Q$. If preferred, however, $Y$-variables can replace $R$-variables and $\Lambda_0$ can replace $\Lambda$ in the results that follow, without altering the relationship between $Y$ and $Q$.

The discrete rating scale, $Q$, may be achieved by partitioning $R$ into successive intervals. Let $\Psi_R = \{\dots, r_0, r_1, r_2, \dots\}$, $r_{\ell-1} < r_\ell$, be the set of cutoffs that forms the partition of $R$, and let $\Omega_Q = \{\dots, q_0, q_1, q_2, \dots\}$, $q_{\ell-1} < q_\ell$, be the discrete domain of $Q$, where where the number of values in $\Omega_Q$ is one more than the number of cutoffs in $\Psi_R$. Formally, the step function is defined as

$$\Lambda(r) \quad = \quad q_\ell, \quad \text{for } r: \; r_{\ell-1} < r \leqslant r_\ell, \tag{5.1}$$

For the $j^{th}$ stimulus, where $Q_j = \Lambda(R_j)$, the effect of $\Lambda$ is to mass the probability associated with $R_j$ within the $\ell^{th}$ interval, $r_{\ell-1} < r \leqslant r_\ell$, and assign it to the probability mass value of $Q_j$ at $q_\ell$, that is, $P(Q_j = q_\ell)$. The values of $\Omega_Q$ define the scaling of the discrete rating scale. Although $\Omega_Q$ is typically a set of integers, values in $\Omega_Q$ can be any real numbers. For example, in transform-average GOC analysis of Taylor et al.'s (1991) data set in Chapter 3, the values of $\Omega_Q$ were determined by the general transform, $g$ (Equation 3.1), applied to a set of integers. The transformed ratings that resulted were generally non-integer (e.g. if $g(r) = \sqrt{r}$).

Statistical results associated with stochastic ordering and $Q$-variables are developed and proved in Section C.3 in Appendix C. The main theorem on the topic (Theorem 3) and its corollaries are reproduced here.

**Theorem 3** *Let $R_j$ and $R_k$ be any two of the R-variables in a family of random variables $\{R_1, R_2, R_3 \dots\}$. Let the step function $\Lambda$, and its related partition $\Psi_R$, be defined as above, for Equation 5.1, and let $Q_j = \Lambda(R_j)$ and $Q_k = \Lambda(R_k)$. If $R_j \overset{st}{<} R_k$, then $Q_j \overset{st}{\leqslant} Q_k$. Whether $Q_j \overset{st}{<} Q_k$ holds, or $Q_j = Q_k$ holds, depends on how $\Psi_R$ partitions the mutual domain of $R_j$ and $R_k$. If $F_{R_j}(r_\ell) > F_{R_k}(r_\ell)$ for any $r_\ell \in \Psi_R$, then $Q_j \overset{st}{<} Q_k$, otherwise $Q_j = Q_k$.*

Theorem 3 states that if a pair of $R$-variables follows a strict stochastic ordering, then the resulting pair of $Q$-variables follows a stochastic ordering which is non-strict at the very least, and which could be strict for any particular pair, depending on certain conditions.

Since $R_j \overset{st}{<} R_k$, then $F_{R_j}(r) \geqslant F_{R_k}(r)$ for all $r$, and $F_{R_j}(r) > F_{R_k}(r)$ holds over some interval of $r$-values, and possibly more than one such interval. If *any* of the cutoffs, $r_\ell$, in the partition of $R$, $\Psi_R$, fall within an interval over which the c.d.f.'s of the $R$-variables are different, then the strict ordering $Q_j \overset{st}{<} Q_k$ holds. (The reason for this is covered in detail in the proof of Theorem 3 given in in Appendix C.) Conversely, if *none* of the cutoffs, $r_\ell$, in $\Psi_R$ fall within an interval over which the c.d.f.'s of the $R$-variables are different, then $Q_j = Q_k$. In other words, if the partition, $\Psi_R$, is such that it only partitions the mutual domain of $R_j$ and $R_k$ at places where their c.d.f.'s are equal, then $Q_j = Q_k$.

Where the partition $\Psi_R$ falls on the mutual domain of the $R$-variables is *independent* of the intervals over which $F_{R_j}(r) > F_{R_k}(r)$ holds true. This implies that it is possible to judiciously choose $\Psi_R$ such that $Q_j \overset{st}{<} Q_k$ results from the transform, $\Lambda$, or to choose $\Psi_R$ such that $Q_j = Q_k$.

In loose terms, the finer the partition of $R$, the less likely it is that $Q_j = Q_k$ and the more likely it is that $Q_j \overset{st}{<} Q_k$, because the more likely it is that at least one cutoff, $r_\ell$, will fall within an interval over which the c.d.f.'s of $R_j$ and $R_k$ are different. In practice, the greater the number of points on a discrete rating scale, the more likely it is that strict stochastic ordering is maintained on $Q$.

The conditions for the strict ordering of $Q_j$ and $Q_k$ are as general as possible, in that no distributional assumptions about $R_j$ and $R_k$ were made. Under certain conditions about the form of $R_j$ and $R_k$, $Q_j$ and $Q_k$ will *always* be strictly ordered, regardless of how $\Psi_R$ partitions the mutual domain of $R_j$ and $R_k$. Specifically, if the c.d.f.'s of $R_j$ and $R_k$ are always different over their entire mutual domain on $R$, then as long as there is at least one cutoff $r_\ell \in \Psi_R$ that falls within that mutual domain, then $Q_j \overset{st}{<} Q_k$ holds. This is always guaranteed if the mutual domain is the entire real number line, $\mathbb{R}$. One example is if $R_j$ and $R_k$ both Gaussian with equal variance but different means.

It does not matter whether $R_j \overset{st}{\leqslant} R_k$ holds, or $R_j \overset{st}{<} R_k$ holds, in the premise of Theorem 3. Either way, the same conclusion can be drawn. Hence,

**Corollary 6**

$$\begin{aligned} R_j \overset{st}{\leqslant} R_k \quad &\Rightarrow \quad Q_j \overset{st}{\leqslant} Q_k \\ &\Rightarrow \quad E(Q_j) \leqslant E(Q_k). \end{aligned}$$

A discrete rating scale may be further partitioned into a second rating scale with fewer categories. If so, Corollary 6 can be iteratively applied to the $Q$-variables themselves (in place of the $R$-variables). The corollary shows that further partitioning of a discrete rating scale maintains non-strict stochastic ordering (if an ordering it is present to begin with). This type of repartitioning is used later, in Section 8.1 for example, in which a 64-point rating scale is converted into a binary-decision scale by applying a single cutoff on the 64-point scale.

Corollary 6 shows that the converse of Theorem 3 is not true in that $Q_j \overset{st}{\leqslant} Q_k$ does not imply $R_j \overset{st}{<} R_k$. One counterexample is if $R_j = R_k$, because $Q_j \overset{st}{\leqslant} Q_k$ holds (by Corollary 6), but $R_j = R_k$ implies that $R_j \overset{st}{\not<} R_k$, so the converse to Theorem 3 cannot hold. Furthermore, although $Q_j \overset{st}{\leqslant} Q_k$ implies $E(Q_j) \leqslant E(Q_k)$, the converse is not true. A counterexample was described in the preceding section, in the examples of the order reversal of means based on Table 3.1 in Section 3.1.2. The mean ratings (on a discrete rating scale) followed any given numerical ordering, but the random variables that describe the distributions involved were not stochastically ordered.

Further corollaries that follow from Theorem 3 are:

**Corollary 7** *Let $R_j$ and $R_k$ be any two random variables whose expectations exist and are finite. Let $\Lambda_1$, $\Lambda_2$, $\Lambda_3$ ... be left-continuous, monotonic increasing step functions, where $\Lambda_1$ is defined on the mutual domain of $R_j$ and $R_k$, $\Lambda_2$ is defined on the mutual domain of $\Lambda_1(R_j)$ and $\Lambda_1(R_k)$, $\Lambda_3$ is defined on the mutual domain of $\Lambda_2(\Lambda_1(R_j))$ and $\Lambda_2(\Lambda_1(R_k))$, and so on. If either $R_j \overset{st}{<} R_k$ or $R_j \overset{st}{\leqslant} R_k$, then*

$$E[\ldots\Lambda_3(\Lambda_2(\Lambda_1(R_j)))] \quad \leqslant \quad E[\ldots\Lambda_3(\Lambda_2(\Lambda_1(R_k)))],$$

*if the expectations exist and are finite.*

**Corollary 8** *Assume there is a family of stochastically ordered $R$-variables, $R_1, R_2, R_3 \ldots$, in which any $R_j$ may be either continuous, discrete, or mixed. Let $\Lambda$, $\Psi_R$ and $\Omega_Q$ be defined as for Theorem 3, and let $Q_j = \Lambda(R_j)$ define a family of discrete $Q$-variables, $Q_1, Q_2, Q_3 \ldots$ If the $R$-variables are such that $R_1 \overset{st}{<} R_2 \overset{st}{<} R_3 \overset{st}{<} \ldots$, then the $Q$-variables are such that $Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} Q_3 \overset{st}{\leqslant} \ldots$ Strict stochastic ordering of the $Q$-variables is possible, but not guaranteed, in accordance with the conditions given in Theorem 3.*

**Corollary 9** *For a family of R-variables and Q-variables that are defined and stochastically ordered as in Corollary 8, the ordering of the expected values of the Q-variables follows the stochastic ordering of the R-variables, if the expectations exist and are finite, so* $E(Q_1) \leqslant E(Q_2) \leqslant E(Q_3) \ldots$ *Strict numerical ordering of these expectations is possible, but not guaranteed, in accordance with the conditions given in Theorem 3.*

Theorem 1 and Corollaries 2, 3 and 4, about strict stochastic ordering following continuous s.m.i. transforms of $Y$-variables have their respective parallels in Theorem 3 and Corollaries 7, 8 and 9 about non-strict stochastic ordering following monotonic increasing step function transforms of $R$-variables.

Corollaries 7, 8 and 9 complete the central theorem of GOC analysis by showing that if the $Y$-variables are stochastically ordered, then the expected values of the $Q$-variables—the expected mean rating values on the discrete rating scale—have a numerical ordering that follows the stochastic ordering of the $Y$-variables. The conditions in the proof of Theorem 3 that are referred to in Corollaries 8 and 9 are described following the statement of Theorem 3, namely that if any $r_\ell$ in the partition of $R$, $\Psi_R$, falls within any interval over which the c.d.f.'s of two $R$-variables are different, then the resulting pair of $Q$-variables are strictly ordered, otherwise the $Q$-variables are equal, and hence non-strictly ordered.

Corollary 7, about nested step function transforms, is relevant to transform-average GOC analysis. Each of the transforms, apart from the first one, may be used to implement an s.m.i. rescaling of prior s.m.i. rescalings of the initial discrete rating scale $\Lambda_1(R)$.[11] In all cases, stochastic ordering holds, although it is not necessarily strict. In general, Corollary 7 implies that once strict ordering has been lost between a pair of random variables, no further nesting of transforms (or rescalings) can recover it.

In the ESO in Figure 5.2, a continuous rating scale, $R$, is intermediate between the decision axis, $Y$, and the discrete rating scale, $Q$, but this is not necessary. It was stated earlier (p. 109) that $\Lambda_0(y) = \Lambda(h(y))$ could be applied directly to $Y$ without requiring either $h$ or $R$ to be separate. This does not alter the results of Theorem 3 and its corollaries, but only changes how they are applied (using a step function denoted $\Lambda_0$ rather than $\Lambda$, and to random variables denoted "$Y$" rather than "$R$").

### Summary

With regards to the ESO in Figure 5.2, Theorem 3 and its corollaries showed that a strictly stochastically ordered family of $R$-variables implies a *non*-strictly ordered family of $Q$-variables (but not vice versa). Non-strict stochastic ordering of random variables implies non-strict numerical ordering of expected values (but not vice versa). This demonstrates

[11] This may be done by a careful choice of cutoffs, such that each successive transform does not pool neighbouring categories from any previous transform.

the second part of the central theorem of GOC analysis, which is summarised by

$$\begin{array}{ccc} R_1 \overset{st}{<} R_2 \overset{st}{<} \ldots & \Rightarrow & Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} \ldots \\ \Downarrow & & \Downarrow \\ E(R_1) < E(R_2) < \ldots & & E(Q_1) \leqslant E(Q_2) \leqslant \ldots \end{array}$$

These results hold true, regardless of the specific quantising function function, $\Lambda$, as long as $\Lambda$ is monotonic increasing and defined on the mutual domain of the $R$-variables (given that the expectations involved exist and are finite). Furthermore, any number of monotonic increasing step function transforms of $Q$, or of $R$, do not affect these results.

Strict stochastic ordering may occur between any particular pair of $Q$-variables, $Q_j$ and $Q_k$, if the partition of $R$, $\Psi_R$, contains any cutoff ($r_\ell$) that falls within an interval where the c.d.f.'s of $Q_j$ and $Q_k$ are different. $Q_j$ and $Q_k$ are equal if and only if no such cutoff exists. This must occur when the c.d.f.'s of $R_j$ and $R_k$ are always different over their mutual domain. An entire family of $Q$-variables would be strictly ordered, for example, if the c.d.f.'s of all $R$-variables were different over the mutual domain of all $R$-variables, and assuming that at least one partition cutoff fell within the mutual domain. This is a sufficient condition for strict ordering of $Q$-variables, but is not necessary. In the absence of knowledge about the underlying $R$-variables, then the greater the number of categories on a discrete rating scale, $Q$, the more likely it is that strict stochastic ordering is maintained throughout $Q$.

## 5.5 Generalisation of the theory of GOC analysis

The theory of GOC analysis accounts for the removal of unique noise as long as the statistical properties of unique and common noise remain constant across replications. When applied to multiple observers, the theory requires a strong assumption about the nature of the individuals in the group, namely that they can all be described by the same equivalent statistical observer. This assumption is hard to justify because of well known individual differences in performance. The central theorem of GOC analysis may account for any *single* observer in an experiment, but what about a *group* of observers, each of which has their own peculiar unique noise characteristics, common noise distributions, mixture process, decision axis, transfer function, rating scale, and even rescaling of the rating scale? What then?

Two generalisations of the theory of GOC analysis are developed here, which show that under certain conditions, (1) observers with different individual statistical characteristics may be combined without loss of stochastic ordering, and (2) weighted sums of ratings taken across observers also maintain stochastic ordering. Appendix D provides proofs of a theorem (Theorem 4) and its corollaries, that show the circumstances under

which stochastic ordering may extend to a group of observers, each of which can have very different unique noise characteristics, common noise distributions, mixture process, decision axes, transfer functions, and rating scales. If all observers in a group share the same statistical properties, the generalisation reduces to the central theorem of GOC analysis, already described. Theorem 4 also extends the theory of GOC analysis to arbitrarily weighted sums of ratings. Conventional GOC analysis based on unweighted sums of ratings, or on arithmetic mean ratings, are both special cases of weighted sums.

The key results proved in Appendix D, are restated here without proof.[12] The results in Appendix D follow from the results in Appendix C, and require extending the underlying model of the theory in order to cover multiple observers. Conceptually, individual observers in a group each have their own ESO, which means they each have their own common noise decision axis, $X$, sample set of $x$-values, unique noise $U$-variables, mixture process, $Y$-variables, transfer function, $R$-variables, quantising function, and their own set of $Q$-variables. In Appendix D, each set of variables and functions is collectively termed a *division* (rather than an *ESO* or *observer*) to keep mathematical results separate from their interpretation. In its broadest form, a division represents a model of a set of replications that share the same statistical details. At this point, assume each division represents an individual observer. The interpretation of a division is broadened later, after the extensions to the theory have been presented.

**Notation.** Let $\eta = 1, 2, 3 \ldots$ denote the observer number, and $\varepsilon = 1, 2, 3 \ldots$ denote the stimulus index number. For the $\varepsilon^{th}$ stimulus and the $\eta^{th}$ observer, the common noise value is $x_{\eta,\varepsilon}$; the random variables are $U_{\eta,\varepsilon}$, $Y_{\eta,\varepsilon}$, $R_{\eta,\varepsilon}$, and $Q_{\eta,\varepsilon}$; the mixture process is $\oplus_\eta$, the transfer function is $h_\eta$ and the quantising function is $\Lambda_\eta$. For the $\eta^{th}$ observer, the random variables are related by

$$Y_{\eta,\varepsilon} \quad = \quad x_{\eta,\varepsilon} \oplus_\eta U_{\eta,\varepsilon},$$

$$R_{\eta,\varepsilon} \quad = \quad h_\eta(Y_{\eta,\varepsilon}),$$

and

$$Q_{\eta,\varepsilon} \quad = \quad \Lambda_\eta(R_{\eta,\varepsilon}).$$

In Section 5.2, the model for a single observer was free from assumptions about distributional forms and about the dependence or independence of variables. The same freedom applies for each of the observers defined here, and also to any interactions across observers

[12] In the same order, and using the same numbering, as in the appendix.

(e.g. $Y_{1,\varepsilon}$ may or may not be independent of $Y_{2,\varepsilon}$, etc.). Furthermore, the mixture processes $\oplus_1, \oplus_2, \oplus_3 \ldots$, transfer functions $h_1, h_2, h_3 \ldots$, and quantising functions $\Lambda_1, \Lambda_2, \Lambda_3 \ldots$ may be either the same or different, across observers. Also, let $a_1, a_2, a_3 \ldots$ be a series of positive constants. These are used as weights when calculating a weighted sum of ratings, or weighted average rating.

Like the statistical theorems presented in Section 5.4, the statistical theorem about weighted sums of ratings deals with a minimal situation, in this case involving only two stimuli and two observers. Corollaries to the theorem extend the results to cover a group containing any number of observers, and stimulus sets containing any number of stimuli. The theorem and its corollaries are stated in terms of $Y$-variables, as if decision axes from various observers could be combined (which is only a temporary measure, for convenience). The results are later applied to rating variables on $R$ and on $Q$.

**Theorem 4** *Let $a_1$ and $a_2$ be any two positive constants. If $Y_{1,j} \overset{st}{<} Y_{1,k}$ and $Y_{2,j} \overset{st}{<} Y_{2,k}$, then $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{<} (a_1 Y_{1,k} + a_2 Y_{2,k})$.*

Theorem 4 is easy to state but hard to prove. It states that if the $Y$-variables for observers 1 and 2 are stochastically ordered for two different stimuli (numbers $j$ and $k$), where the stochastic ordering is the same for each observer, then the weighted sum of the $Y$-variables (summed across observers on a per-stimulus basis) is also stochastically ordered, in the same order as for each observer.

Corollary 10 extends Theorem 4 to cover an arbitrary number of observers.

**Corollary 10** *If $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ for all $\eta = 1, 2 \ldots m$, then*

$$\left( \sum_{\eta=1}^{m} a_\eta Y_{\eta,j} \right) \overset{st}{<} \left( \sum_{\eta=1}^{m} a_\eta Y_{\eta,k} \right) \tag{5.2}$$

*holds for any $m \geqslant 2$.*

Corollary 11 further extends Theorem 4 to cover an an arbitrary number of stimuli, and Corollary 12 states the implication for the expected values of the weighted sums.

**Corollary 11** *If the $Y$-variables for each of $m$ observers form a strictly stochastically ordered set, and the same ordering holds within each set, then the weighted sums across observers follow the same ordering. Without loss of generality, if $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \ldots$ holds for all observers, $\eta = 1, 2 \ldots m$, then*

$$\left( \sum_{\eta=1}^{m} a_\eta Y_{\eta,1} \right) \overset{st}{<} \left( \sum_{\eta=1}^{m} a_\eta Y_{\eta,2} \right) \overset{st}{<} \left( \sum_{\eta=1}^{m} a_\eta Y_{\eta,3} \right) \overset{st}{<} \ldots$$

**Corollary 12** *Without loss of generality, if* $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \ldots$ *holds for all observers,* $\eta = 1, 2 \ldots m$*, then*

$$E\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,1}\right) < E\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,2}\right) < E\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,3}\right) < \ldots$$

Corollary 11 shows that if each observer has a strictly stochastically ordered decision axis, $Y$, where all of the observers order an entire stimulus set in the same way (but each based on their own $Y$-variables), then the weighted sum of the $Y$-variables (summed across observers on a per-stimulus basis) will also be stochastically ordered, and will follow the same order as for each observer. Consequently, the expected values of weighted sums follow the same ordering (Corollary 12). Given the condition of stochastic ordering for each observer, then each individual's particular scaling on $Y$ does not matter. Furthermore, the $Y$-axis for one observer does not have to be an s.m.i. transform of the $Y$-axis of another observer in order for Corollaries 11 and Corollary 12 to hold. In contrast, if there is at least one observer for whom stochastic ordering of $Y$-variables does not apply, then it is not possible to state that the ordering of weighted sums in Corollaries 11 and 12 must hold—the ordering of weighted sums may or may not hold for any specific case.

The next two corollaries are concerned with *non-strict* stochastic ordering between the $j^{th}$ and the $k^{th}$ $Y$-variables for each of two observers.

**Corollary 13** *If* $Y_{1,j} \overset{st}{\leqslant} Y_{1,k}$ *and* $Y_{2,j} \overset{st}{\leqslant} Y_{2,k}$*, then* $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{\leqslant} (a_1 Y_{1,k} + a_2 Y_{2,k})$.

**Corollary 14** *If either* **(a)** $Y_{1,j} \overset{st}{\leqslant} Y_{1,k}$ *and* $Y_{2,j} \overset{st}{<} Y_{2,k}$*, or* **(b)** $Y_{1,j} \overset{st}{<} Y_{1,k}$ *and* $Y_{2,j} \overset{st}{\leqslant} Y_{2,k}$*, then* $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{<} (a_1 Y_{1,k} + a_2 Y_{2,k})$.

Corollary 13 shows that if the $Y$-variables for both observers are non-strictly ordered, then the weighted sums is also non-strictly ordered. In contrast, Corollary 14 shows that if there is strict ordering for either observer, then the weighted sums must be strictly ordered. Below, Corollary 15 is the extension of Corollary 14 to more than two observers, and Corollary 16 is the extension of Corollary 13 to more than two observers.

**Corollary 15** *Assume there are* $m \geqslant 2$ *observers, and that non-strict ordering* $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ *holds for all* $\eta = 1, 2 \ldots m$*. If* $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ *also holds for at least one value* $\eta \in \{1, 2, \ldots, m\}$ *(i.e. there is strict ordering for at least one observer), then* $\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right)$ *(i.e. Equation 5.2) holds.*

**Corollary 16** *Assume there are* $m \geqslant 2$ *observers, and* $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ *holds for all* $\eta = 1, 2 \ldots m$*. The weighted sums,* $\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right)$ *and* $\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right)$*, in Equation 5.2 only follow a non-strict stochastic ordering if the j-versus-k pairwise stochastic ordering is non-strict for all*

*of the contributing observers. Since* $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ *holds for all* $\eta$*, then the weighted sums follow the non-strict ordering*

$$\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right) \quad \overset{st}{\leqslant} \quad \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right)$$

*only if* $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ *does not hold for any of the observers,* $\eta = 1, 2 \ldots m$.

Corollaries 15 and 16 are complementary. Corollary 15 states that the weighted sums are strictly ordered if there is strict ordering for *any* of the observers, whereas Corollary 16 states that the weighted sums are only non-strictly ordered if each and every one of the contributing observers have non-strict ordering.

### Weighted sums of ratings

The results presented to this point deal with weighted sums of random variables defined on each observer's unique-noise-affected decision axis, $Y$. As they stand, the results are not practically useful because an observer's decision axis is not directly accessable. This section presents extensions of the previous results to cover weighted sums of *ratings*, taken across observers. It requires incorporating the s.m.i. transfer function of $Y$ onto $R$ for each observer, prior to calculating the weighted sums. Monotonic increasing step function transforms onto $Q$ are also dealt with. Fortunately, the bulk of mathematics needed for these extensions has already been presented in the preceding theorem and corollaries. The following equations are derived in Section D.3 in Appendix D, and the key results are restated here.

**Weighted sums of ratings on a continuous rating scale,** $R$**.** Suppose that there are $m \geqslant 2$ observers, each with an individual set of $Y$-variables, transfer function, and $R$-variables for the same experimental stimulus set, where $R_{\eta,\varepsilon} = h_\eta(Y_{\eta,\varepsilon})$ for the $\eta^{th}$ observer and $\varepsilon^{th}$ stimulus. Stochastic ordering of $Y$-variables may or may not exist for each observer, and even if it does, it may or may not be the same across observers for the same stimulus set. However, if such stochastic ordering does exist on $Y$, *and it is the same across all observers*, then the results in Section 5.4.1 (applied separately to each observer) show that the stochastic ordering of the $R$-variables for each observer is also the same across all observers, and follows the ordering of $Y$-variables. Assume that this is the case.

Theorem 4 and its corollaries were applied to weighted sums of $Y$-variables, but the same results can also be applied to weighted sums of $R$-variables—notation and interpretation are different (using "$R$" instead of "$Y$", and interpreting random variables as ratings instead of evidence variables), but the results still hold (on $R$). This works only because the $R$-variables are stochastically ordered, and are in the same order for each

observer. Explicitly, when there are $m \geqslant 2$ observers, and when $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \ldots$ holds for all observers, $\eta = 1, 2 \ldots m$, then $R_{\eta,1} \overset{st}{<} R_{\eta,2} \overset{st}{<} \ldots$ also holds for all observers, which implies that

$$\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,1}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta R_{\eta,2}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta R_{\eta,3}\right) \overset{st}{<} \ldots \tag{5.3}$$

and

$$E\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,1}\right) < E\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,2}\right) < E\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,3}\right) < \ldots \tag{5.4}$$

are guaranteed to hold, regardless of each observer's $Y$-variables, and regardless of their individual transfer functions.

**Weighted sums of ratings on a discrete rating scale, $Q$.** Following from the results about $R$-variables, suppose each of the $m \geqslant 2$ observers has their own quantising function and set of $Q$-variables, where $Q_{\eta,\varepsilon} = \Lambda_\eta(R_{\eta,\varepsilon})$ for the $\eta^{th}$ observer and $\varepsilon^{th}$ stimulus (in the same experimental stimulus set). Since the $R$-variables all follow the same stochastic ordering accross observers, then the results in Section 5.4.3 (applied separately to each observer) show that the stochastic ordering of the $Q$-variables for each observer is also the same across all observers, and follows the ordering of the underlying $Y$-variables. The ordering of $Q$-variables is not necessarily strict, but depends on how $R$ is partitioned for each observer (Section 5.4.3). The stochastic ordering of $Q$-variables for a given pair of stimuli could be strict for some observers, and be non-strict for other observers.

Corollaries 15 and 16 were applied to weighted sums of $Y$-variables, but the same results can also be applied to weighted sums of $Q$-variables (like the $R$-variables, the notation and interpretation are different, but the results still hold on $Q$). This works only because the $Q$-variables are stochastically ordered, and are in the same order for each observer. Explicitly, when there are $m \geqslant 2$ observers, and when $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \ldots$ holds for all observers, $\eta = 1, 2 \ldots m$, then $Q_{\eta,1} \overset{st}{\leqslant} Q_{\eta,2} \overset{st}{\leqslant} \ldots$ also holds for all observers, which implies that

$$\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,1}\right) \overset{st}{\leqslant} \left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,2}\right) \overset{st}{\leqslant} \left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,3}\right) \overset{st}{\leqslant} \ldots \tag{5.5}$$

and

$$E\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,1}\right) \leqslant E\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,2}\right) \leqslant E\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,3}\right) \leqslant \ldots \tag{5.6}$$

are guaranteed to hold, regardless of each observer's $Y$-variables, and regardless of their individual transfer functions *and* quantising functions. Strict ordering may or may not apply between each successive pair of weighted sums in Equations 5.5 and 5.6. Corollaries 15 and 16 may be applied to sets of $Q$-variables instead of $Y$-variables. Under these corollaries, the ordering of weighted sums in Equations 5.5 and 5.6 is strict (at a given position in the sequence) if the ordering of $Q$-variables is strict (at that position) for at least one of the contributing observers. If the ordering is non-strict (at that position) for all observers, then the weighted sums in Equations 5.5 and 5.6 follow a non-strict ordering. Whether the ordering is strict for each individual observer depends on their particular characteristics, and especially their individual $R$-axis and how it is partitioned by their quantising function (Section 5.4.3), that is, where their cutoffs are set.

### 5.5.1 Summary and Discussion

The results shown in the preceding section generalise the theory of GOC analysis to cover multiple observers with individual characteristics. The generalised theory is easier to see in sketch form. For brevity, let $\Sigma_{Y,j} = \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right)$, $\Sigma_{R,j} = \left(\sum_{\eta=1}^{m} a_\eta R_{\eta,j}\right)$, and $\Sigma_{Q,j} = \left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,j}\right)$. The relationships among all of the variables in the generalised theory may be summarised as

$$
\begin{array}{ccccc}
Y_{1,1} \overset{st}{<} Y_{1,2} \ldots & \Leftrightarrow & R_{1,1} \overset{st}{<} R_{1,2} \ldots & \Rightarrow & Q_{1,1} \overset{st}{\leqslant} Q_{1,2} \ldots \\
Y_{2,1} \overset{st}{<} Y_{2,2} \ldots & \Leftrightarrow & R_{2,1} \overset{st}{<} R_{2,2} \ldots & \Rightarrow & Q_{2,1} \overset{st}{\leqslant} Q_{2,2} \ldots \\
Y_{3,1} \overset{st}{<} Y_{3,2} \ldots & \Leftrightarrow & R_{3,1} \overset{st}{<} R_{3,2} \ldots & \Rightarrow & Q_{3,1} \overset{st}{\leqslant} Q_{3,2} \ldots \\
\vdots & & \vdots & & \vdots \\
\Downarrow & & \Downarrow & & \Downarrow \\
\Sigma_{Y,1} \overset{st}{<} \Sigma_{Y,2} \overset{st}{<} \ldots & & \Sigma_{R,1} \overset{st}{<} \Sigma_{R,2} \overset{st}{<} \ldots & & \Sigma_{Q,1} \overset{st}{\leqslant} \Sigma_{Q,2} \overset{st}{\leqslant} \ldots \\
\Downarrow & & \Downarrow & & \Downarrow \\
E(\Sigma_{Y,1}) < E(\Sigma_{Y,2}) < \ldots & & E(\Sigma_{R,1}) < E(\Sigma_{R,2}) < \ldots & & E(\Sigma_{Q,1}) \leqslant E(\Sigma_{Q,2}) \leqslant \ldots
\end{array}
$$

where arrows show the direction of implication. In this sketch, each row of the upper tier represents a different observer. Weighted summation takes place down each column of $Y$-variables that share the same second index (stimulus number). If there is only one observer, then the sketch reduces to the central theorem of GOC analysis.

The extended theory of GOC analysis shows that if stochastic ordering held among the $Y$-variables for each observer, and if the ordering was the same for a given stimulus set (i.e. the second subscripts all line up), then the ordering of stimuli by the expected value of a weighted sum of ratings, taken across individual observers, would be the same as the ordering based on the expected values of underlying $Y$-variables. Consequently, the asymptotic GOC curve based on any weighted sum across observers is the same as the

asymptotic GOC curve of each of the individual observers, which is the same for each observer.

Apart from the ordering property, very few assumptions are required. The $Y$-axes (across observers) may be the same, but need not be the same. The $Y$-axis for one observer does not have to be an s.m.i. transform of the $Y$-axis for another observer. Each observer may use a rating scale in different ways, and have a different transfer function and quantising function. The distributions on $R$ or on $Q$ may differ across observers, and they do not need to be approximations of distributions on $Y$ for each observer, as suggested by Metz and Shen (1992). If the same stochastic ordering applies across observers, then weighted sums of ratings, taken across observers on a per stimulus basis, are covered by the central theorem of GOC analysis (Section 5.3.2). However, if the stochastic ordering is different for different observers, then the asymptotic GOC curve for each observer may be different. There is no guarantee about what the asymptotic GOC curve based on a weighted sum across observers would look like, because there is no guarantee that stochastic ordering of weighted sums (Equations 5.3 and 5.5) would hold.

**Use of weightings.** In the results on weighted sums of ratings, the positive constants, $a_1$, $a_2$, $a_3$ ... were left unspecified so that arbitrary weightings could be applied in any particular model of a given experiment. The weightings could be equal or unequal. If the underlying assumptions hold (identical stochastic ordering on $Y$ for each observers), then the theory of GOC analysis shows that it does not matter *asymptotically* what the weights are. However, some weightings are more efficient than others for recovering unique-noise-free performance based on a fixed number of replications (Sorkin & Dai, 1994).

In any particular model, if the weighting constants are all equal, $a_1$, $a_2$, $a_3$ ... may all be set to any constant without affecting results. If the constant is unity, the weighted sum is a simple sum of ratings, and if the constant is $\frac{1}{m}$, the weighted sum is an arithmetic mean rating. When modelling a data set comprised of an unequal number of replications per observer, the weightings in the model may be set to equal the number of replications per observer. Weightings could also be used to reflect individual contributions to a group decision, in order to optimise group decision making (Sorkin & Dai, 1994), or to account for individual levels of unique noise (which was attempted in Chapter 3 for Taylor et al.'s, 1991, frequency discrimination data).

**Other uses for divisions.** The concept of a *division* was introduced on p. 124 to refer to a separate set of random variables on $X$, $U$, $Y$, $R$ and $Q$, all defined in the context of the ESO in Figure 5.2., along with a mixing function, transfer function, and quantising function. The concept of a division is used throughout Appendix D, rather than an individual observer, to keep the mathematical results and their interpretation separate. All of the results about weighted sums in the current section have been stated with the

view of describing GOC analysis of a group of individuals. Although each division can be used to represent an individual observer, under the broadest interpretation, each division represents a statistical model for a given set of replications based on a given stimulus set. It does not require each set of replications to originate from the same observer, or different sets to originate from different observers. Each division could represent the same observer associated with data collected under different experimental conditions (but still using the same stimuli). For example, consider the case of one observer who runs a series of replications in quick succession, takes a break for six months and then continues to run more replications on the same stimulus set. It may be appropriate to model each set of replications using the same ESO. Things may have changed over the break, however, and the observer may use the rating scale differently at the two different times. If so, the transfer function in the model could be adjusted accordingly.

The partitioning of an experimental data set is up to the experimenter. A data set may be partitioned along the lines of individual observers, or groups of observers or types of observers. A data set may also be categorised in terms of time, such as time of day, or whether or not there has been a break between replications. In general, a division represents a partition of an experimental data set, and there is nothing in the theory preventing each replication being described by a separate division. Theoretical results on weighted sums show that stochastic ordering may still hold, even if an observer's decision axis changes from replication to replication.

## 5.6 Discussion

### 5.6.1 Summary of the theory

The theory of GOC analysis assumes an equivalent statistical observer, or ESO, which is a general statistical model of an inconsistent observer. The ESO was outlined in Figure 5.2, and is comprised of a unique-noise-affected decision axis, $Y$, a continuous rating scale, $R$, and an optional discrete rating scale, $Q$. Ratings on $R$ derive from evidence values on $Y$, through an s.m.i. transfer function, $R = h(Y)$. Ratings on $Q$ derive from ratings on $R$ through a monotonic increasing step function, $Q = \Lambda(R)$. Both continuous and discrete rating scales are incorporated within the same model. The distinction between $R$ and $Q$ reflects the process of analysing a continuous rating scale by partitioning the rating continuum. The ESO also models discrete (e.g. push-button) rating scales, in which case the ESO is reformulated so that $Q$ derives from $Y$ without an intermediate $R$ scale. For this case, theoretical results between $R$ and $Q$ are reinterpreted in terms of $Y$ and $Q$.

The theory of GOC analysis is based on stochastic ordering on a decision axis, and on a rating scale. Each experimental stimulus is associated with a random variable on $Y$, and, consequently, a random variable on $R$, and a random variable on $Q$. A stimulus set is

associated with a *family* of random variables on each of $Y$, $R$, and $Q$. The main theoretical findings are that if the family on $Y$ is stochastically ordered, then:

- The families on $R$ and on $Q$ must also be stochastically ordered, in the same order as the family on $Y$. If the ordering on $Y$ is strict, then the ordering on $R$ is strict, and the ordering on $Q$ may be a combination of strict and non-strict (between successive pairs of $Q$-variables).
- Stochastic ordering holds on $R$, regardless of the specific s.m.i. transfer function between $Y$ and $R$, and holds on $Q$, regardless of the monotonic increasing step function between $R$ and $Q$.
- The *numerical* ordering of expected values on each of $Y$, $R$ and $Q$ must follow the *stochastic* ordering on $Y$. Numerical ties on $Q$ are possible, whenever there is non-strict stochastic ordering on $Q$, but this may be avoided if $Q$ results from a high-resolution partition of $R$.

As a result of stochastic ordering, the asymptotic GOC curve based on expected ratings on $R$ is identical to the ROC curve based on expected evidence values on $Y$. The asymptotic GOC curve based on expected ratings on $Q$, is at the very least, consistent with the ROC curve on $Y$, and the two curves may be identical, depending on possible tied values on $Q$.

The theory of GOC analysis is *very* general with respect to distributional forms. The theory shows that rating distributions do not need to approximate distributions on a decision axis. Examples were given in Section 5.1 that showed that GOC analysis could work, even if rating distributions were highly distorted. The theory is also very general with respect to assumed transforms. The specific transfer function, and quantising function, do not need to be known in order for stochastic ordering to hold on $R$ or on $Q$.

The theory of GOC analysis was extended to multiple observers that have individual characteristics, and places constraints on existing models of them (e.g. Metz and Shen, 1992; Sorkin and Dai, 1994). Assume that, for the same set of stimuli, each observer in a group has their own unique-noise-affected decision axis, $Y$, continuous rating scale, $R$, and (optionally) discrete rating scale, $Q$, all of which are based on individual transfer functions and quantising functions. Furthermore, assume that stochastic ordering holds on $Y$ for each observer, and the stimulus set is ordered in the same way for each observer, based on the family of random variables on $Y$. In this case, the weighted sum-of-ratings, taken across observers and calculated on a per-stimulus basis, will also form a stochastically ordered family of random variables. The ordering of the weighted sums is the same ordering as for each individual observer. Consequently, the asymptotic GOC curve based on weighted sums of ratings is identical to that for any individual observer, which is also the same as the ROC curve based on expected values on $Y$ for each individual observer. If each observer shares the same statistical properties, then the extended theory of GOC analysis

reduces to the central theorem of GOC analysis. If the weightings are all equal, then the result is the same as for GOC analysis based on unweighted sums of ratings.

The theory of GOC analysis explains how GOC analysis can work. The theory does not state that GOC analysis must work. GOC analysis may be applied to any multiple-replication data set without any requirement of stochastic ordering. If stochastic ordering does not hold, then arbitrary results are possible, to some extent. Without stochastic ordering, the asymptotic GOC curve based on $R$ or on $Q$ depends on the scaling of the rating scale, and may be different from the ROC curve based on expected $Y$-values. Without stochastic ordering, it is *always* possible to choose a rating scale which places the mean ratings of a pair of stimuli in either order, as desired, or sets them to be tied equal. Discrepancies in unique-noise-free performance between $Y$ and $R$ (or $Q$) can arise because of scaling in the data analysis, and not because of the decisions that were originally made by the observer.

It is possible that stochastic ordering may apply to the $Y$-variables for some pairs of stimuli and not for others. At best, all $Y$-variables are ordered, whereas at worst, none are ordered. If there is a combination of both ordered and non-ordered $Y$-variables, then the asymptotic GOC curve may still depend on the scaling of the rating scale, because expected ratings for the non-ordered $Y$-variables can still be arbitrarily reordered.

### Decomposing $Y$ into common noise, $X$, and unique noise, $U$

Given an ordered set of $R$-variables or $Q$-variables, the theory of GOC analysis only reveals information about the ordered nature of $Y$-variables. The theory does not retrieve, and cannot retrieve stimulus ordering according to $x$-values *without further assumptions* beyond those stated for the ESO. This is because, given a set of $Y$-variables, there are an unlimited number of ways of decomposing of $Y$-variables into possible $x$-values and possible $U$-variables. Further assumptions may allow separation of separate common and unique noise from their combined mixture on $Y$. (For example, assuming that $E(U_j) = 0$ and that $Y_j = x_j + U_j$ for all stimuli, in which case $E(Y_j) = x_j$ for all stimuli.) In general, however, unique-noise-affected observers may be comprised of multiple processing stages, prior to the derivation of $Y$ (e.g. Taylor, 1984; Durlach et al., 1986), where each stage contributes to unique and common noise. In such cases, the derivation of $x_j$ and $U_j$ from $Y_j$ is not necessarily simple, and may not reflect how $Y_j$ came about. Given a set of random variables on $R$ or $Q$, the theory of GOC analysis cannot demonstrate with certainty how $Y$ was derived, but by the same token, nor can any other theory, without making extra assumptions.

### 5.6.2 Models incorporated within the theory of GOC analysis

Two types of assumptions that are used in models of unique noise are about the form of the unique noise and how it mixes with common noise. Simplifying assumptions are often made, for example that unique noise may be characterised by a single random variable, and that unique noise is additive with common noise (e.g. Tanner, 1961). This section shows the consequences of assuming additive mixing, and multiplicative mixing, for stochastic ordering in the theory of GOC analysis.

#### Identical additive unique noise

One of the simplest types of model of an inconsistent observer is where unique noise is additive with common noise, and unique noise characteristics do not change as a function of common noise value. In the theory, this translates into having the same unique noise distribution for every stimulus, that is, $U_j = U_k = U$ for all $j$ and $k$. In this case, $Y_j = x_j + U$ and $Y_k = x_k + U$ for all $j$ and $k$. Stochastic ordering is guaranteed to hold under this type of model, because the c.d.f. of $Y_j$ is the c.d.f. of $U$ shifted left or right along the decision axis by the amount $x_j$. Specifically, the c.d.f.'s of $Y_j$ and $Y_k$ are

$$F_{Y_j}(y) = F_U(y - x_j)$$

and

$$F_{Y_k}(y) = F_U(y - x_k)$$

respectively. If $x_j$ and $x_k$ are such that $x_j < x_k$, then $F_{Y_j}(y) \geqslant F_{Y_k}(y)$ for all $y$, and so $Y_j \overset{st}{\leqslant} Y_k$. Furthermore, $x_j + U \neq x_k + U$, since $x_j$ and $x_k$ are different, which is to say that $Y_j \neq Y_k$. Since $Y_j \overset{st}{\leqslant} Y_k$ and $Y_j \neq Y_k$, therefore $Y_j \overset{st}{<} Y_k$ holds for all $j$ and $k$ such that $x_j < x_k$. This shows that if unique and common noise are additive, and unique noise has the same form for all stimuli, then the resulting family of $Y$-variables is strictly stochastically ordered according to the set of $x$-values. The result holds regardless of the specific form of $U$, and regardless of the particular common noise distributions from which the $x$-values are sampled.

Additive unique and common noise models with unique noise distributions of a fixed form are the most common type of model of unique-noise-affected observers.[13] Most of these models explicitly assume that the unique noise random variable, $U$, is Gaussian,

[13](Swets et al., 1959; Tanner, 1961; Watson, 1963; Wickelgren, 1968; Wilcox, 1968; McNicol, 1972; Green & Swets, 1974; Boven, 1976; Siegel, 1979; Taylor, 1984; Berg, 1987, 1989, 1990; Siegel & Colburn, 1989; Taylor et al., 1991; Metz & Shen, 1992). Models of internal and external noise have been included here also.

and also that $E(U) = 0$, so that in effect $E(Y_j) = x_j$ for all stimuli. The theory of GOC analysis shows that stochastic ordering holds for any additive and fixed unique noise forms, including the Gaussian form (with fixed variance and zero mean).

**Non-identical additive unique noise without stochastic order**

The stochastic ordering of a family of $Y$-variables does not generally depend on the expected values of distributions on $X$, $U$, $Y$, $R$, or $Q$. It can be the case that $E(Y_j) = x_j$ for all $j$, and yet stochastic ordering does not hold. For example let $Y_j = x_j + U_j$ for all $j$, where the $U$-variables are all Gaussian with zero mean ($E(U_j) = 0$), but all have different variances. For this case, none of the possible pairs of $Y$-variables will be stochastically ordered. Each stimulus-pair ROC curve crosses the chance line, since it is based on a pair of Gaussian distributions of unequal variance (Section 5.3.1). The ordering of stimuli according to $E(Y_j)$ is the same as the ordering according to $x_j$-values, since $E(U_j) = 0$ and so $E(Y_j) = x_j$. The ordering according to $E(R_j)$ or $E(Q_j)$ is not necessarily that based on $E(Y_j)$ or $x_j$. Hence, it is possible to re-order stimuli by applying an s.m.i. transform to such a decision axis. If Sorkin and Dai's (1994) additive Gaussian model (described in Chapter 3) is extended to incorporate transfer functions on to a rating scale, then specific cases of their model would be affected by this result.

**Identical multiplicative unique noise**

Another type of model that is described in the theory of GOC analysis is where unique noise is multiplicative with common noise, and unique noise characteristics are the same across $x$-values. The following results hold if the set of $x$-values are either all positive, or are all negative. The positive case is assumed below.

Let the unique noise distribution, $U$, be the same for every stimulus, so $U_j = U_k = U$ for all $j$ and $k$. Under multiplicative mixing, the $Y$-variables are $Y_j = x_j\,U$ and $Y_k = x_k\,U$. The c.d.f. for the $j^{th}$ stimulus is

$$\begin{aligned} F_{Y_j}(y) &= P\left(x_j U \leqslant y\right) \\ &= P\left(U \leqslant \frac{y}{x_j}\right) \\ &= F_U\left(\frac{y}{x_j}\right), \end{aligned}$$

and the c.d.f. for the $k^{th}$ stimulus is

$$F_{Y_k}(y) = F_U\left(\frac{y}{x_k}\right).$$

If $x_j$ and $x_k$ are such that $x_j < x_k$, where $x_j, x_k > 0$, then $\frac{y}{x_j} > \frac{y}{x_k}$. This implies that

$$F_U\left(\frac{y}{x_j}\right) \geqslant F_U\left(\frac{y}{x_k}\right) \tag{5.7}$$

holds for all $y$, that is,

$$F_{Y_j}(y) \geqslant F_{Y_k}(y).$$

holds for all $y$, and so $Y_j \overset{st}{\leqslant} Y_k$. Furthermore, $x_j\, U \neq x_k\, U$, since $x_j$ and $x_k$ are different, which is to say that $Y_j \neq Y_k$. This implies that $Y_j \overset{st}{<} Y_k$ holds for all $j$ and $k$ such that $x_j < x_k$. Like in the additive case, strict stochastic ordering holds, and it is determined by the values of $x_j$ and $x_k$.

Stochastic ordering also holds when the common noise is all negative (i.e. $x_j, x_k < 0$). If $x_j$ and $x_k$ are such that $x_j < x_k$, where $x_j, x_k < 0$, then $\frac{y}{x_j} > \frac{y}{x_k}$, and through a similar argument to the development above, $Y_j \overset{st}{>} Y_k$. The $Y$-variables are stochastically ordered according to $x$-values, but the ordering of $Y$-variables is in the direction opposite to the ordering of $x$-values.

If the common noise could be either positive or negative (e.g. sampled from Gaussian distributions), then a general conclusion about the stochastic ordering of $Y$-variables is not possible. For a given pair of stimuli, the direction of the inequality in Equation 5.7 then depends on the absolute values of $x_j$ and $x_k$, as well as on their signs, but not on whether $x_j < x_k$.

In summary, if unique and common noise are multiplicative, where common noise is entirely positive and where unique noise has the same distributional form for all stimuli, then the resulting family of $Y$-variables is stochastically ordered according to the set of $x$-values. If common noise is entirely negative, then the set of $Y$-variables is stochastically ordered in a direction opposite to set of $x$-values. Like the earlier results for additive noise, the result for multiplicative noise holds regardless of the distributional forms of unique and common noise.

### Additive unique noise compared to multiplicative unique noise

The results for additive and multiplicative mixing have implications for the special case where the common noise is entirely positive in value.[14] This may occur, for example, if common noise results from a rectified process, as is the case in models of aural amplitude discrimination (Jeffress, 1964; Green & McGill, 1970). If common noise is entirely positive *and* unique noise is identical across stimuli (i.e. $x_j > 0$, and $U_j = U$, for all stimuli, $j$),

[14] Examples of possible common noise distributions include (central or non-central) gamma or $F$ distributions, including special cases such as chi, chi-squared, Rayleigh, Rayleigh-Rice, exponential, Maxwell, and half-normal distributions.

then the stochastic ordering based on additive mixing is identical to the stochastic ordering based on multiplicative mixing. This is because the order of $Y$-variables is determined by the order of $x_j$ and $x_k$ under both types of mixing. Consequently, the asymptotic GOC curves based on the two types of mixing are identical. Furthermore, it is impossible to determine, based on the rating distributions or on the asymptotic GOC curve, which type of mixing occurred. If, on the other hand, the form of unique noise differs across stimuli (e.g. $U_j \neq U_k$ for all $j$ and $k$), or the common noise is not all positive or all negative, then additive and multiplicative mixing can result in families of distributions that either differ in their stochastic ordering, or may not be ordered at all.

### 5.6.3 Transforms of decision axes when stochastic ordering does not hold

The theory of GOC analysis is aimed at explaining the removal of unique noise by averaging ratings on a rating scale. Results about stochastic ordering, and Theorem 2 in particular, are also relevant to models in which unique noise is removed by averaging values on a *decision axis*. This covers most existing models of unique-noise-affected observers, including models of multiple-presentation experiments (Swets et al., 1959; Berg, 1987; McKinley & Weber, 1994), and models of multiple-replication experiments (Watson, 1963; Taylor et al., 1991; Metz & Shen, 1992; Sorkin & Dai, 1994).

In Theorem 2, the function, $h$, was interpreted in terms of an s.m.i. transform of $Y$ onto $R$. Instead, let $h$ represent an s.m.i. transform of one decision axis onto a second decision axis, rather than onto a rating scale. In this case, Theorem 2 shows that if a family of $Y$-variables is not stochastically ordered on one decision axis, then the corresponding family of $Y$-variables on the second decision axis will not be stochastically ordered either. This implies that the ordering of a stimulus set according to expected values on one $Y$-axis is not necessarily the same as the ordering according to expected values on a second, s.m.i.-related $Y$-axis. In more general terms, when stochastic ordering does not hold, then averaging out unique noise on one $Y$-axis is not necessarily the same as averaging out unique noise on a second, s.m.i.-related $Y$-axis. Since stimulus order may change according to scaling, then the expected theoretical ROC curve (once unique noise has been averaged out) can differ for different s.m.i. transforms of a decision axis. This result runs contrary to the notion, applied in a unique-noise-*free* context, that the theoretical ROC curve is identical for all s.m.i. transforms of a decision axis. In contrast, Theorem 1 shows that if stochastic ordering holds on one $Y$-axis, then it holds on all s.m.i.-related $Y$-axes, and averaging out unique noise on one $Y$-axis results in the same stimulus ordering, and ROC curve, as averaging out unique noise some other s.m.i.-related $Y$-axis.

### 5.6.4 Quasi-molecular experiments

Quasi-molecular experiments are detection tasks that use a relatively small number of reproducible stimuli, each of which is presented many dozens of times.[15] Observer inconsistency is usually present in these experiments, and decisions change from trial to trial for the same stimulus. Often, the aim of these studies is to investigate the variability in signal detectability associated with individual samples of masking noise (Pfafflin, 1968; Gilkey, 1981). For example, a series of SIFC trials may be run using only a particular masker waveform, to which a signal is added on some of the trials. If the detection task uses binary-decision methodology, a hit and false alarm rate pair is calculated, which is plotted as a point in the ROC space. Each different masker results in a different stimulus-pair ROC point.

In the theory of GOC analysis, variability in decisions for the $j^{th}$ and $k^{th}$ stimuli are associated with random variables, $Y_j$ and $Y_k$, defined on $Y$. These are transformed by an s.m.i. transfer function into $R_j$ and $R_k$ defined on $R$, and $Q_j$ and $Q_k$ defined on $Q$. For a given masker waveform, say $Y_j$ is associated with the masker alone, that $Y_k$ is associated with the masker plus signal. In theory, a stimulus-pair ROC point based on the $j^{th}$ and $k^{th}$ stimuli (based on a finite number of presentations) results from samples from $Q_j$ and $Q_k$. For convenience, assume that enough presentations have been run so that $Q_j$ and $Q_k$ are known. In that case, the ROC point based on $Q_j$ and $Q_k$ is a single point on the stimulus-pair ROC curve based on $R_j$ and $R_k$, and the ROC curve based on $R_j$ and $R_k$ is identical to that based on $Y_j$ and $Y_k$. This is because the transfer function between $Y$ and $R$ is s.m.i., and between $R$ and $Q$ is a monotonic increasing step function.

Stimulus-pair ROC curves based on $R$ can, in theory, demonstrate if a pair of $Y$-variables, such as $Y_j$ and $Y_k$, are stochastically ordered (as shown in Section 5.3.1). If they are ordered, then their associated stimulus-pair ROC curve on $R$ lies either on or above the chance line, but never below it. If the $Y$-variables are *strictly* ordered, then the curve must lie above the chance line over some some portion of the ROC space. Quasi-molecular SIFC experiments can be used to investigate stochastic ordering in a data set.

When binary-decision methodology is used, which is most often the case, then only a single ROC point may be obtained, based on $Q_j$ and $Q_k$. Most experimental stimulus-pair ROC points lie on or above the chance line,[16] although some occasionally lie below it. Without knowing where the rest of the curve is in the ROC space, it is not possible to say with certainty that stochastic ordering holds. If a continuous rating scale, $R$, is used instead of a binary-decision rating scale, $Q$, then a well defined stimulus-pair ROC curve is possible, which reflects the underlying curve based on $Y_j$ and $Y_k$. This could be done

[15] (Green, 1964; Pfafflin & Mathews, 1966; Pfafflin, 1968; Siegel, 1979; Gilkey, 1981; Gilkey et al., 1985; Siegel & Colburn, 1989; Isabelle & Colburn, 1991). Quasi-molecular experiments were described in the historical development of GOC analysis in Chapter 2.

[16] Assuming that the hit rate is conditional on an masker-plus-signal stimulus, and false alarm rate is conditional on a masker-alone stimulus.

in practice, and only requires a change in rating methodology, not in basic experimental design. Such experiments are still SIFC experiments, and would not take any longer to complete than existing studies. Quasi-molecular analyses based on well defined rating scales have not been reported to date, but such experiments are feasible.

For a set of $n$ stimuli, there are $n^2$ possible pairings, and stimuli may be paired regardless of which experimental event they are associated with. (If ordering holds over an entire stimulus set, then two *SN* stimuli should be ordered, as well one *N* stimulus and one *SN* stimulus.) Any systematic and obvious crossing of the chance line in a pair's ROC curve would indicate that the $Y$-variables for the stimuli were not ordered. A comparison among $n^2$ possible ROC curves allows much scope for testing, and the possibility of modelling the $Y$-variables for the entire stimulus set. Although it is not possible to work back from $R$ to uniquely derive $Y$, it is possible that analyses along these lines may help to eliminate classes of distributions from consideration.

Given some assumptions about a particular decision axis, it is possible to model sample sets associated with $Y$-variables. Chapter 4 described how transfer functions could be derived from an ROC or GOC curve. A transfer function was used to estimate unique-noise-affected evidence values on a decision axis. The estimated evidence value[17] for the $j^{th}$ stimulus and $i^{th}$ replication represents the sample value, $y_{ji}$. Assuming a decision axis, it was theoretically possible to estimate properties of $Y_j$ and $Y_k$, such as their variance, using sets of $y_{ji}$ values, taken across replications.

**Siegel and Colburn's model.** A theoretical model was proposed in a series of quasi-molecular studies by Siegel and Colburn (Siegel, 1979; Siegel & Colburn, 1983, 1989). Their model is based on additive Gaussian unique noise that has a fixed variance and zero mean for all common noise values. The model is an example of a specific case of the ESO in Figure 5.2, and it is possible to simply convert between the notation used in here and the notation used by Siegel and Colburn.[18] Their model describes a specific case of unique and common noise interaction, but is not in itself a theory of GOC analysis.

### 5.6.5 Sampling variablity of common and unique noise

The theory of GOC analysis is mostly concerned with families of random variables, and their properties in the context of the ESO, rather than with finite data sets. In practice, the number of stimuli used in an experiment, and the number of replications that can be run, are both finite. Finite data sets are associated with both common noise sampling variability, and unique noise sampling variability, both of which affect GOC results.

[17] This was called an "$x$-value" in Chapter 4, where similar notation was used differently from here.

[18] Equating notation across models, $L$, $m(x)$ and $y(L, x)$ in Siegel and Colburn's model would take the place of $Y$, $x_j$ and $U_j$ in Figure 5.2, respectively. Siegel and Colburn framed their model in terms of internal and external noise, rather than unique and common noise, respectively.

Common noise sampling variability occurs because the number of stimuli is finite, so only a finite number of $x_j$ values may be sampled from $X_{SN}$ and $X_N$. The ROC curve implied by a sample of $x$-values will vary from sample to sample, and is called the *sample-theoretical ROC curve* (Lapsley Miller, 1999), to distinguish it from the (population) theoretical ROC curve based on the random variables on $X$. In the case where the order of the $Y$-variables reflects the order of their associated $x$-values, then GOC curves tend towards the sample-theoretical ROC curve, and not the (population) theoretical ROC curve.

Common noise sampling variability has been investigated by simulation (Pollack & Hsieh, 1969; Lapsley Miller, 1996, 1999) and analytically (Bamber, 1975). A large part of common noise variability would relate to the sampling of stimulus waveforms, where a different set of experimental stimuli would result in a different value of $\mathcal{A}$, even if there was no unique noise. The smaller the stimulus set, the greater the possible variability in common noise, and conversely, the larger the stimulus set, the smaller the variability. To ensure that effects across experimental conditions, for instance, are not due to such variability, the stimulus sample size should be as large as possible.

While stimulus sampling (or its statistical equivalent) is a major source of potential variability or error in psychophysical performance, unique noise may have at least as great an effect. Unique noise sampling variability occurs because the number of replications that may be run is only finite, in practice. Assume that $m$ replications worth of data have been collected, and $r_{ji}$ is the rating made for the $j^{th}$ stimulus on the $i^{th}$ replication. The GOC curve is then based on $\frac{1}{m}\sum_{i=1}^{m} r_{ji}$, for each $j$, which tends to $E(R_j)$ as $m \to \infty$. Hence the ordering of the stimulus set according to sample mean rating tends toward the ordering of the stimulus set according to expected mean rating, and so as $m \to \infty$, the empirical GOC curve will tend toward the asymptotic GOC curve.

Stochastic ordering of sets of random variables (e.g. $R_1$,$R_2$, ... on a rating scale) does not ensure stochastic ordering of sample sets based on the random variables, or vice versa. Stochastic ordering of the sample sets of ratings for a pair of stimuli can be checked by comparing the sample c.d.f.'s across stimuli. By the definition of stochastic ordering, if the sample c.d.f. for one stimulus is always greater than or equal to the sample c.d.f. for another stimulus, then the two samples are stochastically ordered, otherwise not. In practice, *lack of sample stochastic ordering may often occur*, even if the underlying random variables are ordered.

Sampling variability of unique noise could account for the results of transform-average GOC analysis shown in Chapter 3. There, the series of transform-average GOC curves showed two opposing effects. The results clearly depended on the scaling of the rating scale, because GOC curves differed from transform to transform. At the same time, however, the GOC curves seemed independent of the scaling, because many of the curves were very similar across transforms. These results can be sensibly interpreted within the theory of

GOC analysis. It may have been that the rating random variables for the stimulus set were stochastically ordered, but that the sample sets of ratings were only partly ordered. Residual unique noise effects would have been present because the number of replications, although large, was still finite. The results in Chapter 3 suggest that although some scalings may have been more efficient at removing unique noise than other scalings, but possibly all scalings would result in the same asymptotic GOC curve if enough replications were run.

## 5.7 Conclusion

Stochastic ordering provides the key to GOC analysis of inconsistent observers. In a unique-noise-free context, the particular ordinal scale that underlies decision making does not matter, to the extent that an ROC curve remains unchanged by any s.m.i. transform of a decision axis. In situations where unique noise is present, however, this type of ordinal scale invariance *no longer holds* in general, either in the case of an s.m.i. transform from one rating scale to another (transform-average GOC analysis), or from a decision axis to a rating scale (a transfer function), or from one decision axis to another. Ordinal scale invariance of a sort will only hold if stochastic ordering holds.

If a stimulus set is associated with a family of random variables that lie on a decision axis or a rating scale, then stochastic ordering was shown to be both necessary and sufficient for the numerical order of their expected values to remain unchanged under arbitrary ordinal scaling. When applied to the relationship between a decision axis and a rating scale, stochastic ordering implies that averaging out unique noise on a rating scale is equivalent to averaging out unique noise on a decision axis, which explains how GOC analysis works.

# Part II

# Functions Of Replications Added

*For a multiple-replication data set, group operating characteristic (GOC) analysis may be used to minimise unique noise effects, and improve performance in a discrimination task. As more replications are combined, performance improves as a function of replications added (FORA). Stable empirical FORAs result from all combinations analysis (ACA), where average performance is calculated over all possible GOC curves for a given number of replications. A widely applicable FORA regression function is introduced. Extrapolation of this function to an infinite number of replications makes it possible to estimate asymptotic unique-noise-free performance, based on a finite data set. Chapter 6 introduces a FORA regression procedure, which is able to estimate known theoretical performance to better than two decimal places. Chapter 7 applies FORA regression to an amplitude discrimination experiment in which 100 replications were run. The very large data set makes it possible to not only estimate asymptotic performance, but to estimate sample statistics and error bounds of the asymptote. Chapter 8 shows FORA results for four sets of experiments on frequency discrimination and amplitude discrimination. FORA regression is shown to be very robust across experimental paradigms, observers, types of stimuli, stimulus parameters, performance levels and measures of sensitivity.*

# Chapter 6

# Functions of replications added

Observer inconsistency in a discrimination task can substantially decrease performance. Figure 2.5, for example, showed how typical single-replication performance was much worse than theoretical performance. GOC analysis can remove the effects of observer inconsistency, and substantially improve performance in the task. GOC analysis reduces error by averaging out unique noise. As more replications are added, GOC performance improves from the mean ROC level and may tend towards an asymptote, which may reflect theoretical performance. This chapter shows how to determine GOC performance as a function of replications added, and how to estimate *unique-noise-free performance* from a finite data set.

Let $m$ be the number of replications in a discrimination task experiment. The greater the number of replications combined in GOC analysis, the smaller the effect of any remaining unique noise until, in the limit as $m \to \infty$, all the unique noise is removed and only common noise remains. Since $m$ is always finite in practice, any estimate of unique-noise-free performance involves an extrapolation to infinity from performance based on a finite data set. The term "performance" primarily refers to a measure of sensitivity, such as area under the GOC curve, rather than the GOC curve itself.

Taylor (1984) pointed out that there are many ways of considering how performance changes as more replications are combined in GOC analysis. One way could be to start with just one replication and calculate its ROC curve. Then, combine the data from the first replication with that of a second replication and derive the two-replication GOC curve. Next, add in a third replication, and so on. If a measure of task performance is calculated for the GOC curve at each step, then sensitivity should improve as more replications are added. The function of sensitivity versus number of replications is called a *sample function of replications added* (sample-FORA).

It may be possible, *in principle*, to work out how GOC performance changes in a sample-FORA as each replication is added. If so, extrapolation to an infinite number of replications would provide a way of estimating unique-noise-free performance. There is a

problem with this approach *in practice*, and that is the choice of what replication to add at each step. There may or may not be an inherent order in the data set, for example a temporal order, and even if there is, the order may not be important. The order in which replications are added becomes a question of experimental design and experimental purpose. Unless order effects are specifically under investigation, the ordering of replications could be quite arbitrary (e.g. if $m$ observers simultaneously run one replication each).

Taylor (1984) noted that in the absence of a useful order of replications, there are $m!$ possible different sequences of $m$ replications. All $m!$ sequences share the same, final, $m$-replication GOC curve. That aside, each sequence is different from every other sequence, both in terms of the set of $m$ GOC curves resulting from the sequence and of the $m$ performance measures derived from those curves. This makes overall interpolation and extrapolation difficult, even if all of the sequences themselves give well-behaved results (which is extremely unlikely).

### All combinations analysis (ACA)

Taylor (1984) proposed a solution to the problem of arbitrary sequences. Instead of calculating multiple sequences of GOC curves, he estimated expected performance when a given number of replications are combined. For an initial data set consisting of $m$ replications, there are ${}^{m}C_{\xi}$ possible subsets of $\xi$ replications, $1 \leqslant \xi \leqslant m$, where $\xi$ is called the *combination-size*. Combinations are appropriate rather than permutations because any sum-of-ratings per stimulus, taken across replications, does not depend on the order of replications. A GOC curve and its associated performance measures can be calculated for each subset of size $\xi$. Performance measure values from all GOC curves based on the same combination-size, $\xi$, are averaged to give an estimate of the expected GOC performance for $\xi$ replications. This procedure can be repeated for all combination-sizes from 1 to $m$, giving a *function of replications added* (FORA). The process of calculating all possible GOC curves from a data set is called *all combinations analysis* (ACA) (Taylor, 1984). The type of FORA obtained by ACA is an average FORA (or just "FORA"). It is equal (pointwise) to the average of the $m!$ sample-FORAs described previously.

For a data set of $m$ replications, there are $\sum_{\xi=1}^{m} {}^{m}C_{\xi} = 2^m - 1$ possible GOC curves, and full ACA involves the calculation of all of them, or at least a measure of sensitivity for all of them. Different measures of sensitivity result in different FORAs for the same set of data, and any measure of sensitivity may be used. A computer program for performing ACA is described in Appendix F, along with details of some computational efficiencies that can be implemented when calculating a FORA.

**Resampling effects in ACA.** ACA is somewhat like bootstrapping, which is used to estimate sample statistics by resampling subsets from a set of values. The analogy is incomplete because the values that are averaged in ACA are measures of performance that

result from GOC analysis. The calculations involved are not as simple as those for the sample mean or standard deviation that are often the focus of bootstrapping. The closest procedure in form to ACA is bootstrapped ROC analysis (Moise, Clement, Ducimetiere, & Bourassa, 1985), but these two analyses are not that similar. ACA involves repeated GOC analysis of combinatorially generated sets of replications (systematically sampled with replacement from the total set of replications), whereas bootstrapped ROC analysis involves repeated ROC analysis of sets of ratings (randomly sampled with replacement) from the data for a single replication.

The effect of resampling in ACA is unknown. Ideally, each set of performance measure values that are averaged at each combination-size would be independent, so that the average FORA is an unbiased estimate of the expected FORA. It is not possible, however, to run the very large number of replications that are needed to obtain independent sample sets.[1] The only reasonable way to do this is by a statistical simulation using Monte Carlo methods, which has not been done.

### FORAs in the literature

Experimental FORAs have appeared in the literature for multiple-presentation tasks as well as for multiple-replication tasks. FORAs, or FORA-like functions, may be obtained using different experimental methodologies, and measures of sensitivity. FORAs from multiple-presentation experiments show how performance changes as a function of the number of observation intervals per trial (Swets et al., 1959; Berg, 1987; McKinley & Weber, 1994), whereas FORAs from multiple-replication experiments show how performance changes as a function of the number of replications added (Boven, 1976; McAulay, 1978; Taylor, 1984; Lapsley Miller, 1999). In the former case, unique noise is averaged out internally whereas in the latter case, unique noise is averaged out externally. Generally, performance improves as the number of stimulus presentations increases, although the rate of improvement depends very much on the parameters of unique and common noise. Theoretical FORAs have been derived, primarily as a consequence of observer models that incorporate additive Gaussian unique noise, and average out unique noise on a decision axis (Swets et al., 1959; Berg, 1987; Metz and Shen, 1992; Sorkin and Dai, 1994; Taylor, 1994, personal communication).

[1] For example, say that only 30 sample values were needed per combination-size, $\xi$, in order to estimate mean FORA points for combination-sizes from one to five. Even for these modest numbers, *independent* sampling would require $30 \times \xi$ replications at each value of $\xi$, or $(1 + 2 + 3 + 4 + 5) \times 30 = 450$ independent replications. Each replication may consist of several hundred trials per event, implying hundreds of thousands of trials in total. For Taylor et al.'s (1991) experiment from earlier chapters, with a 24 replication and with 416 trials per replication, the same reasoning implies that independent sampling would require 3.7 million trials in total.

**Overview of the rest of the chapter**

The FORAs presented in this chapter are not based on any previous theoretical FORA, or any model of a unique-noise-affected observer, including the equivalent statistical observer described in the previous chapter. Rather, the FORAs in this and the following chapters demonstrate a distribution-free and measure-free data model that is applicable to many different FORAs from many different experiments.

An empirical FORA for Taylor et al.'s (1991) data set based on the measure $\mathcal{A}$, is presented in Section 6.1. The FORA shows how GOC performance tends toward known theoretical performance as replications are added, and also shows the variability associated with sample-FORAs. Exploratory data analysis in Section 6.2 results in non-parametric regression of empirical FORAs. Two regression methods based on the same mathematical form are compared and contrasted, and a non-linear FORA regression is shown to be a good fit to data. The regression-FORA can be extrapolated to an infinite number of replications to provide an estimate of unique-noise-free, asymptotic performance, which can be a good approximation to theoretical performance. FORAs based on $d'$, $\mathcal{D}_2$ and $P(C)$ are also presented, and are compared to known theoretical performance.

## 6.1 Experimental example

All combinations analysis was performed on data from Taylor et al.'s (1991) continuous rating scale experiment, and the resulting FORA based on the area under the GOC curve, $\mathcal{A}$, is shown in Figure 6.1. Since there were 24 replications, the FORA is based on about $2^{24}$ ($\simeq$ 17 million) GOC curves, with ${}^{24}C_{\xi}$ GOC curves contributing to the data point at each combination-size, $\xi$. The central point at each combination-size is the arithmetic mean value of $\mathcal{A}$ for that combination-size, and the error bars indicate plus or minus one standard deviation from the mean.

The mean value of $\mathcal{A}$ increases as a function of combination-size. Performance improves as more replications are combined, reflecting a removal of unique noise. As the combination-size tends to infinity, the mean value of $\mathcal{A}$ should tend towards the theoretical value of $\mathcal{A}$, which was 0.8550 for this experiment. The mean value of $\mathcal{A}$ tends toward theoretical performance because the expected $m$-replication GOC curve tends toward the theoretical ROC curve as $m \to \infty$. The initial FORA value, at a combination-size of 1, is just the mean value of $\mathcal{A}$ for all 24 single-replication ROC curves. (Each ROC curve can be seen as a 1-replication GOC curve). The first point on the FORA in Figure 6.1 ($\mathcal{A} = 0.7394$) was very close to the area under the arcsin-averaged mean ROC curve ($\mathcal{A} = 0.7400$). Not surprisingly, such a pattern is typically the case for most of the FORAs presented in this thesis.

If the individual data points (values of $\mathcal{A}$) averaged at each combination-size were to

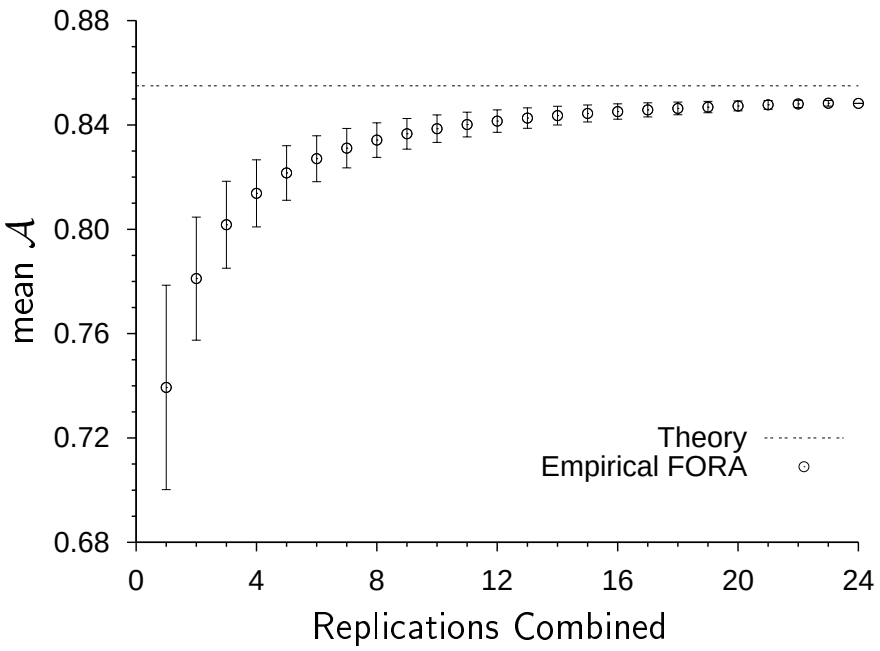


FIGURE 6.1: The function of replications added for Taylor et al.'s (1991) continuous rating scale experiment, showing mean-$\mathcal{A}$ as a function of the number of replications combined. Error bars show plus or minus one standard deviation from the mean. The horizontal line indicates the theoretical value of $\mathcal{A}$ at 0.8550.

be plotted instead of the mean value, they would line up in columns at each value of $\xi$. There would be ${}^{24}C_{\xi}$ points in each column, and the $m!$ sample-FORAs result from all possible connections between single points taken at successive combination-sizes. Since the error bars in Figure 6.1 only indicate plus or minus one standard deviation, the points would be spread over over a much larger range than is shown. The graph indicates that the potential variability in sample-FORAs is huge, hence using an average FORA is a practical solution for showing how performance improves with replications.

## 6.2 FORA regression

This section deals with fitting a regression function to a FORA. The development is presented based on the measure $\mathcal{A}$, with the understanding that similar regression procedures can be applied to FORAs based on other measures of sensitivity.

The FORA in Figure 6.1 approaches the theoretical level asymptotically from below. This suggests diminishing returns in GOC analysis, because the performance increment gained by running each additional replication becomes smaller and smaller. Theoretical performance in an experiment is generally unknown, but it can be estimated from a finite data set. The asymptote of the FORA indicates how well an observer *could* perform if *all* of the unique noise could be removed (i.e. if the observer made entirely consistent decisions

across replications).

Estimating an asymptote requires fitting a regression function to a finite, empirical FORA and extrapolating the function to infinity. Hyperbolic and exponential functions were initially examined, since they are asymptotic in form, but they did not prove useful. A regression of FORA-*increments* was attempted instead, which led to a successful form of FORA regression.

The FORA in Figure 6.1 is reproduced in Figure 6.2(a), without error bars. The FORA increases rapidly initially, but then flattens out. The first increment in performance is the largest, the second increment is the next largest, and so on. Let $y_j$ denote[2] the mean value of $\mathcal{A}$ at combination-size $j$ $(1 \leqslant j \leqslant m)$, and let

$$\delta_j \quad = \quad y_j - y_{j-1}, \qquad 2 \leqslant j \leqslant m \tag{6.1}$$

denote the $(j-1)^{th}$ *increment* in mean value of $\mathcal{A}$. Figure 6.2(b) shows the increment, $\delta_j$, as a function of $j$, which decreases to zero as a function of replications added. Figure 6.2(c) shows $\log_e(\delta_j)$ versus $j$, which is also a decreasing function, and Figure 6.2(d) shows $\log_e(\delta_j)$ versus $\log_e(j)$. This plot of the log of the increment versus the log of the index are referred to as a *log-log plot*.[3] A log-log plot can be presented using logarithms to any arbitrary base, and natural logarithms (base $e$) have been chosen as the convention here.

The log-log plot in Figure 6.2(d) was fairly linear. Pearson's product-moment correlation coefficient, denoted $r$, was $-0.9988$ for this data set, and its square was $r^2 = 0.9977$. In the present context, $r^2$ can be interpreted as the proportion of the variability in the logarithm of the increments that was dependent on the variability in the logarithm of the combination-size. A linear function was used at this point as a first approximation to the data series. Although the function in Figure 6.2(d) curved slightly downwards, log-log plots from other experiments, presented in Chapters 7 and 8 are even straighter.

Assume a linear approximation to the log-log plot. If the slope of the line is $\mu$, the intercept is $c$, and $\kappa = exp(c)$, then

$$\begin{aligned} \log_e(\delta_j) &\simeq \mu \log_e(j) + c \\ &= \log_e(j^\mu) + \log_e(\kappa) \\ &= \log_e(\kappa j^\mu), \qquad j \geqslant 2. \end{aligned} \tag{6.2}$$

This implies that

$$\delta_j \quad \simeq \quad \kappa j^\mu, \tag{6.3}$$

[2]The notation used in this chapter is independent of the notation in the preceding chapter.

[3]Although there are 24 points on the FORA in Figure 6.2(a), there are only 22 points on the log-log plot in Figure 6.2(d). This is because the very last FORA increment for this data set is actually a very small decrement, so no logarithm is calculated for it.

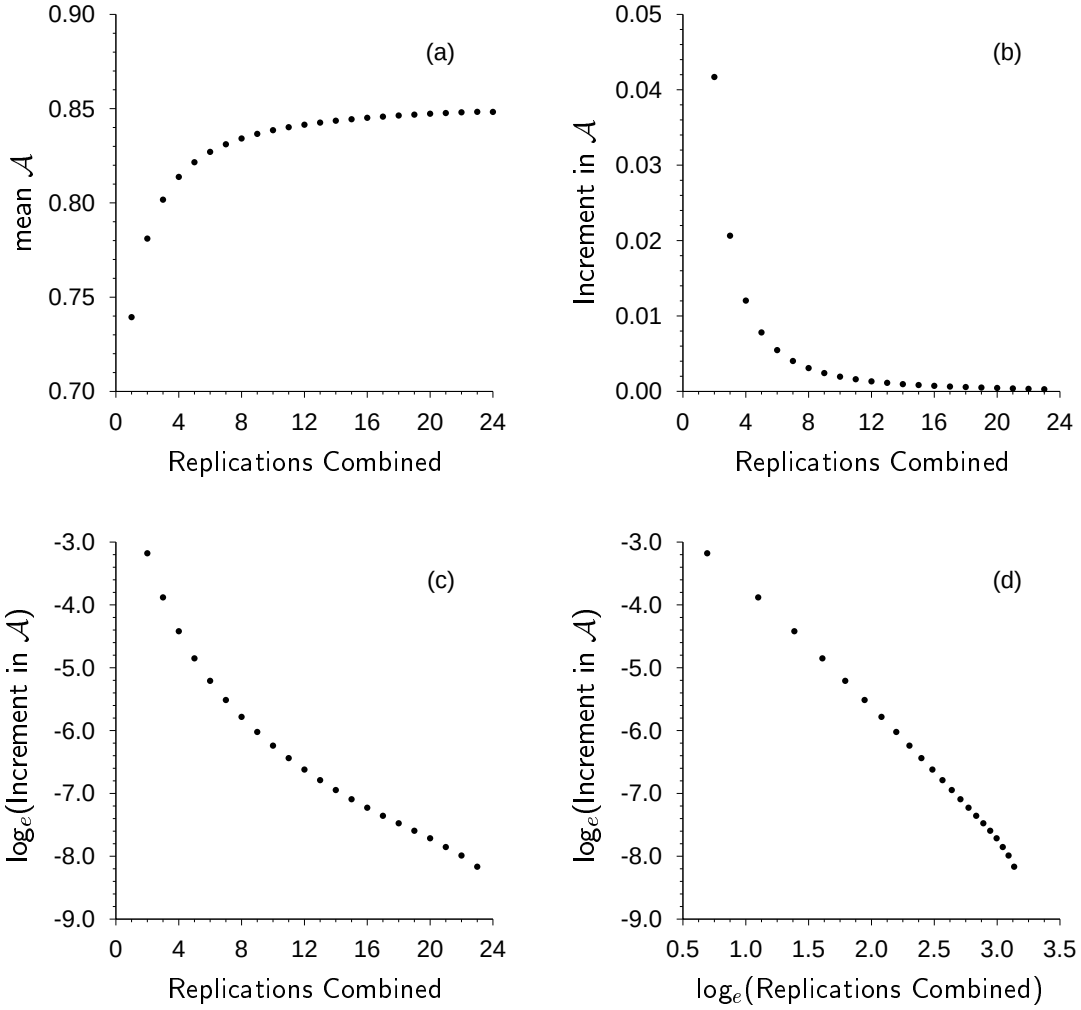


FIGURE 6.2: Transformation of the FORA presented in Figure 6.1: (a) mean value of $\mathcal{A}$ versus combination-size; (b) increments in mean $\mathcal{A}$ versus combination-size; (c) log-increments in mean $\mathcal{A}$ versus combination-size; (d) log-increments in $\mathcal{A}$ versus the logarithm of the combination-size.

which forms the basis of a FORA regression function. Since the data series is of the form

$$y_i = y_1 + \sum_{j=2}^{i} \delta_j, \quad i \geqslant 2, \tag{6.4}$$

then the regression-FORA is of the form

$$\begin{aligned} A_i &= A_1 + \sum_{j=2}^{i} \kappa\, j^{\mu} \\ &= A_1 + \kappa \sum_{j=2}^{i} j^{\mu}, \quad i \geqslant 2, \end{aligned} \tag{6.5}$$

where $A_1$ is the first *predicted* value of $\mathcal{A}$. Equation 6.5 may be extrapolated to infinity to give an estimate of the asymptotic value of $\mathcal{A}$, namely

$$A_\infty = A_1 + \kappa \sum_{j=2}^{\infty} j^{\mu} \tag{6.6}$$

$$= A_1 + \kappa\,(\zeta(-\mu) - 1) \tag{6.7}$$

for $\mu < -1$ (Gradshteyn & Ryzhik, 1965, Equation 9.522 (1)), where $\zeta(x) = \sum_{j=1}^{\infty} j^{-x}$ is the Riemann zeta function.

The purpose of the regression is to find a function of best fit to the empirical FORA rather than to the log-log plot, so Equation 6.5 is the regression function of interest rather than Equation 6.2. Equation 6.5 describes a three-parameter data model with parameters $A_1$, $\kappa$ and $\mu$. The log-log plot is useful nevertheless, because it gives a simple, graphical way of checking goodness-of-fit, and its curvature can reveal subtle trends that are not obvious in a FORA plot.

The derivation of Equations 6.3 and 6.5 from Equation 6.2, and the calculation of $\mu$ and $\kappa$, are invariant with respect to the base of the logarithm used in the plot. The slope, $\mu$, is independent of an arbitrary logarithm base $b$, because both the abscissa and ordinate are equally scaled in double-logarithmic coordinates. The intercept, $c = b^{\kappa}$, is not independent of $b$, but since $\kappa$ is the FORA regression parameter of interest, rather than $c$, then $c$ may be defined in terms of $\kappa$ ($c \stackrel{def}{=} \log_b(\kappa)$), rather than vice versa.

FORA regression originates from a linear log-log plot, but can be achieved without deriving any log-log plot (Section 6.2.3). There should be an interpretation of each parameter with respect to a FORA which is independent of any log-log plot. $A_1$ denotes the initial value of the regression-FORA (i.e. expected ROC performance), and $\mu$ determines the relative curvature of a FORA (see below), but what $\kappa$ by itself means is not clear. It is clear that $\kappa$ and $\mu$ jointly determine the *total improvement* of a FORA, which is

equal to $\kappa \sum_{j=2}^{\infty} j^{\mu}$ (from Equation 6.6), or $\kappa(\zeta(-\mu) - 1)$ (from Equation 6.7). The total improvement is the absolute value of the difference between the first point on the FORA and the predicted asymptote. It could also be called the *potential improvement from ROC performance.*

As well as being the slope of a log-log plot, $\mu$ also solely determines the *relative curvature* of a FORA (which is different from the curvature of a log-log plot). The relative curvature of a FORA determines the rate at which a FORA approaches its asymptote, relative to its total improvement. If $\mu$ is close to -1 (e.g. $\mu = -1.1$), then the log-log plot is very shallow, meaning that successive FORA increments only decrease slowly, and hence the FORA only approaches its asymptote slowly. On the other hand, if $\mu$ is more negative (e.g. $\mu = -3.0$), then the log-log plot is much steeper, so successive FORA increments decrease more rapidly. In that case, the FORA increases rapidly at first and then flattens out. Relative curvature can be used to compare performance across observers, experimental conditions or signal-to-noise ratios. The larger the absolute value of $\mu$, the greater the relative curvature and so the smaller the number of replications needed in order to attain any given proportion of the potential improvement.

### 6.2.1 Factors affecting FORA regression

It difficult to specify the *best* FORA regression procedure because of the complexity of ACA calculation. In ACA, interrelated combinations of subsets of data are taken from a single experimental data set and then GOC analysis is applied to each combination. These steps make it hard to mathematically derive statistics that describe *experimental FORAs*, including the variability associated with each FORA point.

There are at least three factors that *should*, in principle, be taken into account by any FORA regression procedure, although it is difficult to do so in practice. First, points on an empirical FORA are highly interdependent rather than independent, because ACA requires combinatorial subsampling of a data set, and because GOC analysis is done on each subsample (combination) of replications. Second, the variability of individual values of $\mathcal{A}$ clearly changes with combination-size, as shown in Figure 6.1, so the points on the sample-FORAs contributing to the average FORA are heteroscedastic, not homoscedastic. Third, the number of combinations contributing to the mean value is different for different combination-sizes, sometimes vastly so. The last two factors might possibly be taken into account of by applying a weighting scheme to FORA regression. How (or even if) the first factor can be accomodated is not known. Of the three factors, the last one may be the easiest to deal with because it is the easiest to quantify.

Although there may be only a relatively small number of points on a FORA, this hides the fact that the number of contributing GOC curves per point could be small or could be huge. For example, with Figure 6.1, there are ${}^{24}C_1 = {}^{24}C_{23} = 24$ GOC curves of size 1 or of size 23, ${}^{24}C_2 = {}^{24}C_{22} = 276$ GOC curves of sizes 2 or 22, but ${}^{24}C_{12} = 2.7$ million GOC

curves of size 12. This suggests that the most stable points on the FORA are those in the middle and the least stable points are those at the ends. This is presumably because the greater the number of GOC curves involved, the smaller the variability associated with the *mean* value of $\mathcal{A}$ (as opposed to the variability of the raw values of $\mathcal{A}$, shown by errorbars in Figure 6.1). The last data point on a FORA always consists of the value of $\mathcal{A}$ from just the single $m$-replication GOC curve. Although the standard deviation associated with the last data point is zero, the fact that only a single GOC curve contributed to the value of $\mathcal{A}$ makes the last point on a FORA perhaps the least stable point of all.

In general, the middle points of a log-log plot are the most stable and the end points are the least stable. The number of GOC curves contributing to a data point on a log-log plot is complicated by the fact the data point is the difference between neighbouring points on the FORA. The $j^{th}$ point on the log-log plot ($2 \leqslant j \leqslant m$) is based on ${}^{m}C_{j} + {}^{m}C_{j-1}$ GOC curves, a number which is smallest for the last point on the right. It is often the case that the last point on the log-log plot appears to be the least stable, and may deviate from the trend shown by the rest of the data points in a log-log plot.

The three factors, interdependence, heteroscedasticity and number of combinations per FORA point, must affect the statistics associated with the mean performance value at each combination-size. In spite of this, all three factors have been ignored in FORA regression, because it is very difficult to specify their influence and to know how to take them into account. *The basic datum* of FORA regression is the *mean* performance value at each combination-size. All data points have equal weight, and no correction is made for variability or number of values contributing to the mean.

There are many potential ways to of fitting Equation 6.5 to an empirical FORA, but the essential difference among them is how they assign values to the parameter triplet $(A_1, \kappa, \mu)$. Two methods are presented. The first is described in Section 6.2.2 and uses a *linear* least-squares fit *to the log-log plot*. The second method is described in Section 6.2.3 and uses a *non-linear* least-squares fit *to the FORA*. Each method arrives at a different result.

### 6.2.2 Linear regression of the log-log plot

Regression of a log-log plot can provide a quick and easy way of fitting the three-parameter data model to the FORA. A linear least-squares fit to the log-log plot may be *useful*, but it does not provide the most accurate regression of a FORA. The linear log-log plot regression provides the parameters $\kappa$ and $\mu$, while the $A_1$ parameter is set to equal the initial FORA value, $y_1$ (the mean value of $\mathcal{A}$ at a combination-size of 1). This regression procedure applied to the data in Figure 6.2 provided a parameter triplet $(A_1, \kappa, \mu)$ of

$(0.739382, 0.193971, -2.016418)$.[4] Figure 6.3(a) shows the empirical FORA, the regression-FORA using these parameters and the estimated asymptote, while Figure 6.4(a) shows the accompanying log-log plot. The sum of the squares of the 24 residuals between the regression-FORA of Equation 6.5 and the empirical FORA was $7.48 \times 10^{-4}$. This was a small total, but the function that fitted to the data points in Figure 6.3(a) is clearly not a function of best fit *to the FORA*, even though the line through the data points in Figure 6.4(a) *is* a line of best fit *to the log-log plot*. Figure 6.3(a) suggests that the asymptote would be over-estimated, and it was. The asymptotic value of $\mathcal{A}$ calculated via Equation 6.7 using the above parameter triplet was 0.8615, compared to the theoretical value of $\mathcal{A}$ of 0.8550.

The correlation of the data points in Figure 6.2(d) was $r = -0.9988$, and so a linear fit *seems* reasonable. If the empirical log-log plot is *highly* linear, with $r$ at $-0.9999$ or better, then this method *can* provide a reasonable fit to the FORA, because the regression function would be very close to the data. If the empirical log-log plot is slightly curvilinear, as it is for this data set, then a linear least-squares fit *to a log-log plot* is not the best way of fitting the three-parameter data model *to a FORA*. Regression of a log-log plot provides a quick and easy regression, but it should only be used as a first approximation.

### 6.2.3 Non-linear regression of the FORA

It is possible to apply a non-linear least-squares regression of the form of Equation 6.5 directly to an empirical FORA. The essential mathematics of the regression are provided in Appendix E. Non-linear FORA regression requires the simultaneous solution of three equations, Equations E.3, E.5 and E.7, which are derived in Appendix E. Each equation is a trivariate function of $A_1$, $\kappa$ and $\mu$, in which the data points of a given empirical FORA are treated as constants. When solved simultaneously, Equations E.3, E.5 and E.7 minimise the sum of squared residuals between the data points on a FORA and the regression function given by Equation 6.5. The regression procedure assumes an equal weighting of all data values (points on the empirical FORA). Appendix F describes a computer program that can numerically solve these equations and estimate asymptotic performance for a given parameter triplet.

Non-linear regression to the empirical FORA in Figure 6.1 provided a parameter triplet $(A_1, \kappa, \mu)$ of (0.738663, 0.166007, $-1.924461$). The $A_1$ parameter calculated this way was similar to that based on linear-log-log-plot regression, but $\kappa$ and $\mu$ were not so similar. Figure 6.3(b) shows the regression-FORA based on non-linear FORA regression and its estimated asymptote, while Figure 6.4(b) shows the accompanying log-log plot. The straight line in Figure 6.4(b) represents the linear function implied by the parameters of

[4] Values of $A_1$, $\kappa$ and $\mu$ are given to 6 decimal places to make it easier to check calculated asymptotes based on the values given. Compiler-specific error, described in Appendix F, means that the parameter values will depend (to a small extent) on the data analysis program that is used.

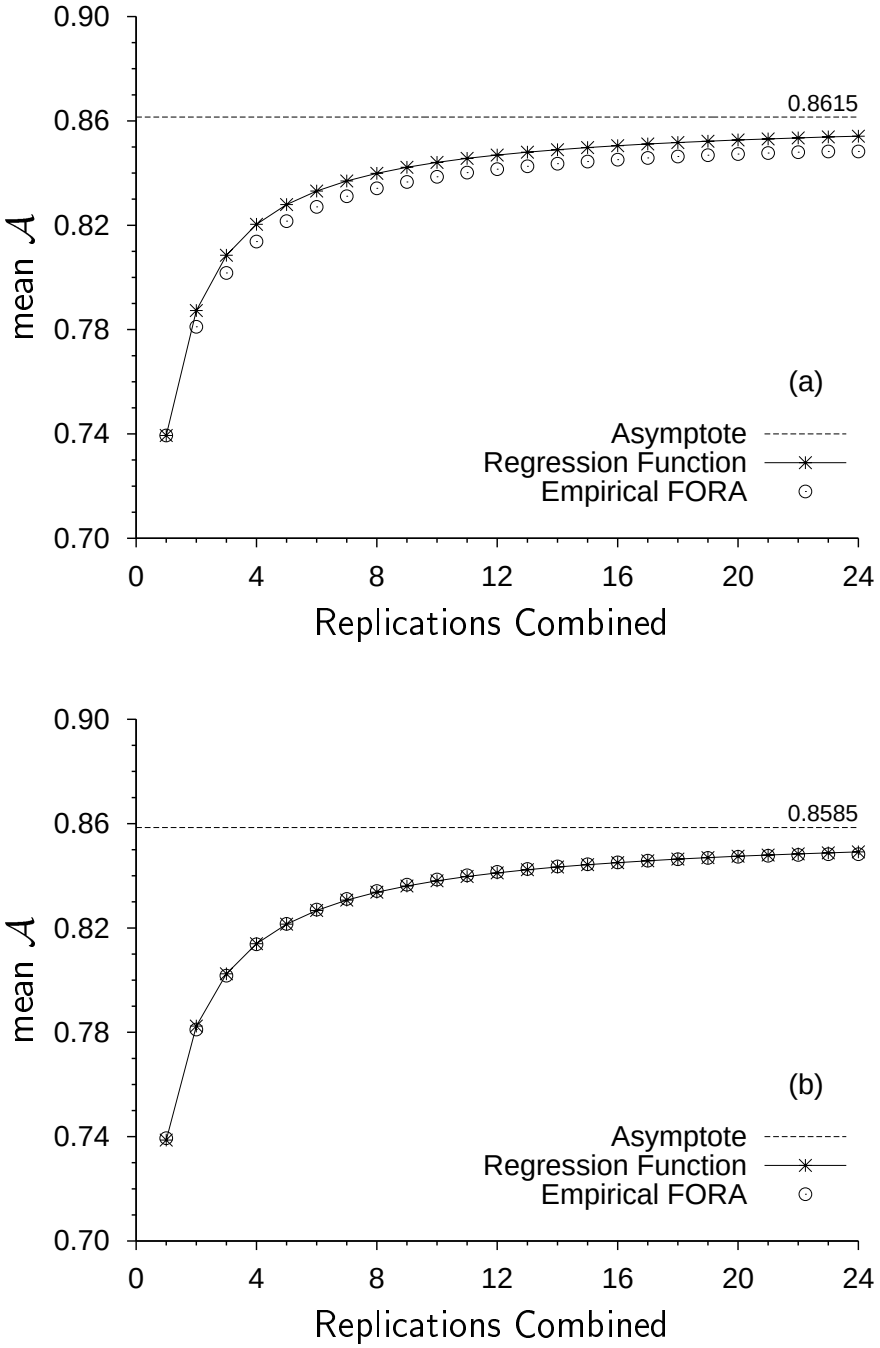


FIGURE 6.3: Different regression procedures applied to the 24-replication FORA presented in Figure 6.1. The horizontal line represents the estimated asymptote for each procedure. (For comparison, the theoretical value of $\mathcal{A}$ was 0.8550.) (a) Regression based on a linear fit to the log-log plot (presented in Figure 6.4(a)). (b) Non-linear regression using the procedure described in Section 6.2.3 and Appendix E. The data points are the same in both of panels (a) and (b).

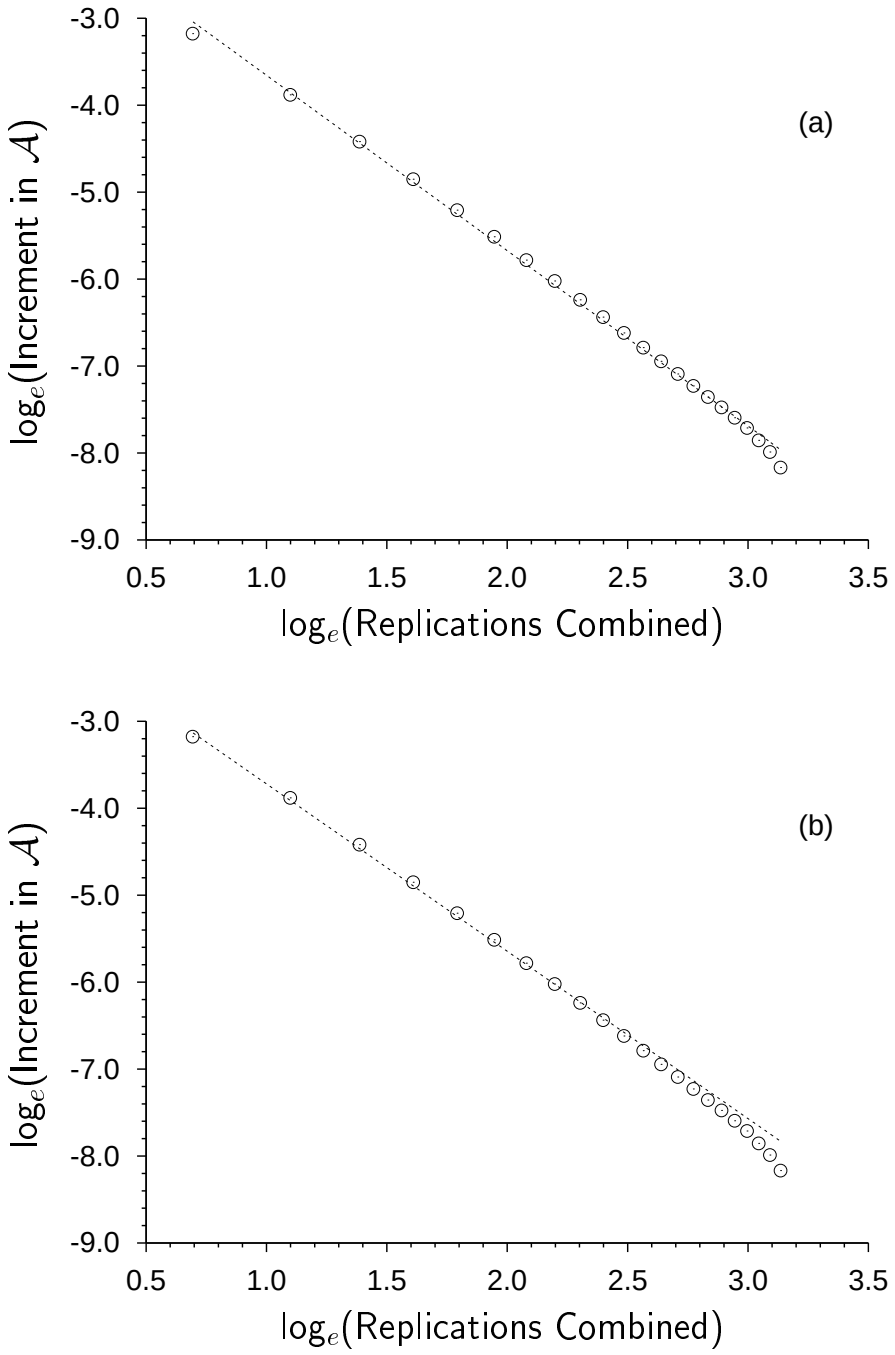


FIGURE 6.4: Log-log plots associated with the different regression-FORAs presented in Figure 6.3. Straight lines indicate the log-log relationship based on parameters from each FORA regression procedure. (a) From a linear least-squares fit to the log-log plot. (b) From a nonlinear FORA regression using the procedure described in Section 6.2.3 and Appendix E. The data points are the same in both of panels (a) and (b).

the regression-FORA in Figure 6.3(b). The implied function in Figure 6.4(b) is not a line of best fit to the log-log plot (which is shown in Figure 6.4(a)).

The non-linear regression-FORA in Figure 6.3(b) was a better fit to the data points than the regression-FORA in Figure 6.3(a). The sum of squared residuals for the non-linear regression was $5.7\times10^{-6}$, which was more than 130 times smaller than the equivalent sum for Figure 6.3(a). The asymptotic value of $\mathcal{A}$ calculated from non-linear regression was 0.8585, compared to the theoretical value of 0.8550. The non-linear regression provided a better estimate of the asymptote than the previous regression estimate (0.8615) in Figure 6.3(a).

Another drawback of basing FORA regression on a linear fit to the log-log plot is that any *decrements* in the FORA do not provide a point on the log-log plot, and so not all of the FORA points contribute to the regression. This happened, for example, between the 23rd and the 24th points in Figure 6.3, resulting in a 22-point log-log plot. Omissions such as this affect any estimation based primarily on a log-log plot, and could add to overestimation of the asymptote, because the FORA data could be flatter than the regression procedure can admit. In contrast, the non-linear least-squares regression is not affected by decrements in a FORA and uses all available FORA points.

The discrepancy between the data series and implied linear relationship in the log-log plot in Figure 6.4(b) is due to the cumulative nature of the non-linear FORA regression procedure. Consider Equation 6.4, $y_i = y_1 + \sum_{j=2}^{i} \delta_j, \quad i \geqslant 2$, which describes the empirical FORA. Each point beyond the first one is comprised of a starting value, $y_1$, and a series of increments, $\delta_1$, $\delta_2$ ... The first increment, $\delta_1$, contributes to all FORA points beyond the first point. The second increment, $\delta_2$, contributes to all FORA points beyond the second point, and so on. In terms of the FORA regression, it is generally more important to get a good approximation to $\delta_{j-1}$ than it is to achieve a good approximation to $\delta_j$, and it is most important to achieve a good approximation to $\delta_1$ (and to the starting value, $y_1$). The result is that the straight line, derived from the parameters that give the best fit of Equation 6.5 to a FORA, tends to be a good fit to the first few points of an empirical log-log plot, but not necessarily to the last few points. This can be seen in Figure 6.4(b). As well as that, the increments in a FORA tend to zero as more replications are added, and so the logarithm of the increments becomes very small. What seems like a large discrepancy in the log-log space, like that in Figure 6.4(b) between the straight line and the empirical data, is in fact only a small discrepancy at the high end of the FORA, as seen in Figure 6.3(b).

*Only non-linear FORA regression* based on Equation 6.5 is used in the rest of this thesis, rather than the regression based on a linear fit to the log-log plot. *The straight lines shown in all subsequent log-log plots derive from the parameters that best fit Equation 6.5 to the FORA, and are <u>not</u> the line of best fit to the log-log plot.*

**Curvature in a log-log plot.** The curvature of an empirical log-log plot indicates where the estimated asymptote may lie, relative to underlying, theoretical performance. The data points in Figure 6.4(b) curved downwards relative to the straight line. A downwards curving log-log plot indicates that as the number of replications increases, the increments in the regression-FORA are slightly larger than those in the empirical FORA. Consequently, the regression-FORA tends to a larger asymptote than suggested solely by the data and so theoretical performance is overestimated. (The asymptotic value of $\mathcal{A}$ estimated in Figure 6.3(b) was 0.8585, compared to a theoretical value of $\mathcal{A}$ of 0.8550.) In contrast, an upwards curving log-log plot indicates that increments in the regression-FORA are slightly smaller than those in the empirical FORA, meaning theoretical performance would be underestimated. Exactly how the bias of an asymptote quantitatively relates to the curvature of a log-log plot is unknown.

### 6.2.4 FORAs based on various measures of sensitivity

Most of the FORAs shown in this thesis are based on $\mathcal{A}$ as the measure of sensitivity or performance. FORAs based on other measures ($d'$, $\mathcal{D}_2$ and $P(C)$) were also calculated for Taylor et al.'s (1991) 24-replication frequency discrimination experiment. Figures 6.5, 6.6 and 6.7 show FORAs and log-log plots for $d'$, $\mathcal{D}_2$ and $P(C)$ respectively, along with the regression-FORA and estimated asymptote for each measure. These results are summarised in Table G.1 in Appendix G. Asymptotes for various measures, including $\mathcal{A}$, generally overestimated theoretical performance, whereas the 24-replication GOC curve (i.e. the last point on each FORA) generally underestimated theoretical performance.

Like in Figure 6.1, the central point at each combination-size in Figures 6.5 to 6.7 is the arithmetic mean sensitivity value for that combination-size, and the error bars indicate plus or minus one standard deviation from the mean. No account has been taken here of the symmetry or skewness of the distribution of sensitivity values at each combination-size. The error bars are presented symmetrically around the mean to indicate the amount of variability rather than its nature.

The FORAs and log-log plots based on $d'$ and $\mathcal{D}_2$ were very similar to those based on $\mathcal{A}$ (Figures 6.1, and 6.2(a) and (d)). The regression-FORAs tended to lie under the empirical FORA for the middle combination-sizes and over at the higher combination-sizes. This was also reflected in the downwards curve of the log-log plots in Figures 6.5 and 6.6. The under-over pattern of the regression-FORAs and curve of the log-log plots were more prominent for $d'$ and $\mathcal{D}_2$ than for $\mathcal{A}$. The log-log plots suggested that the asymptotes for $d'$ and $\mathcal{D}_2$ would also be overestimated, like for $\mathcal{A}$. The estimated asymptotic $d'$ was 1.5310 compared to a theoretical value of 1.4966, and the estimated asymptotic $\mathcal{D}_2$ was 0.4222 bits compared to a theoretical value of 0.4029 bits. The overestimation was about 2.3% for $d'$, and about 4.5% for $\mathcal{D}_2$.

Two FORAs based on two related measures of sensitivity, such as $\mathcal{A}$ and $d'$, are *not*

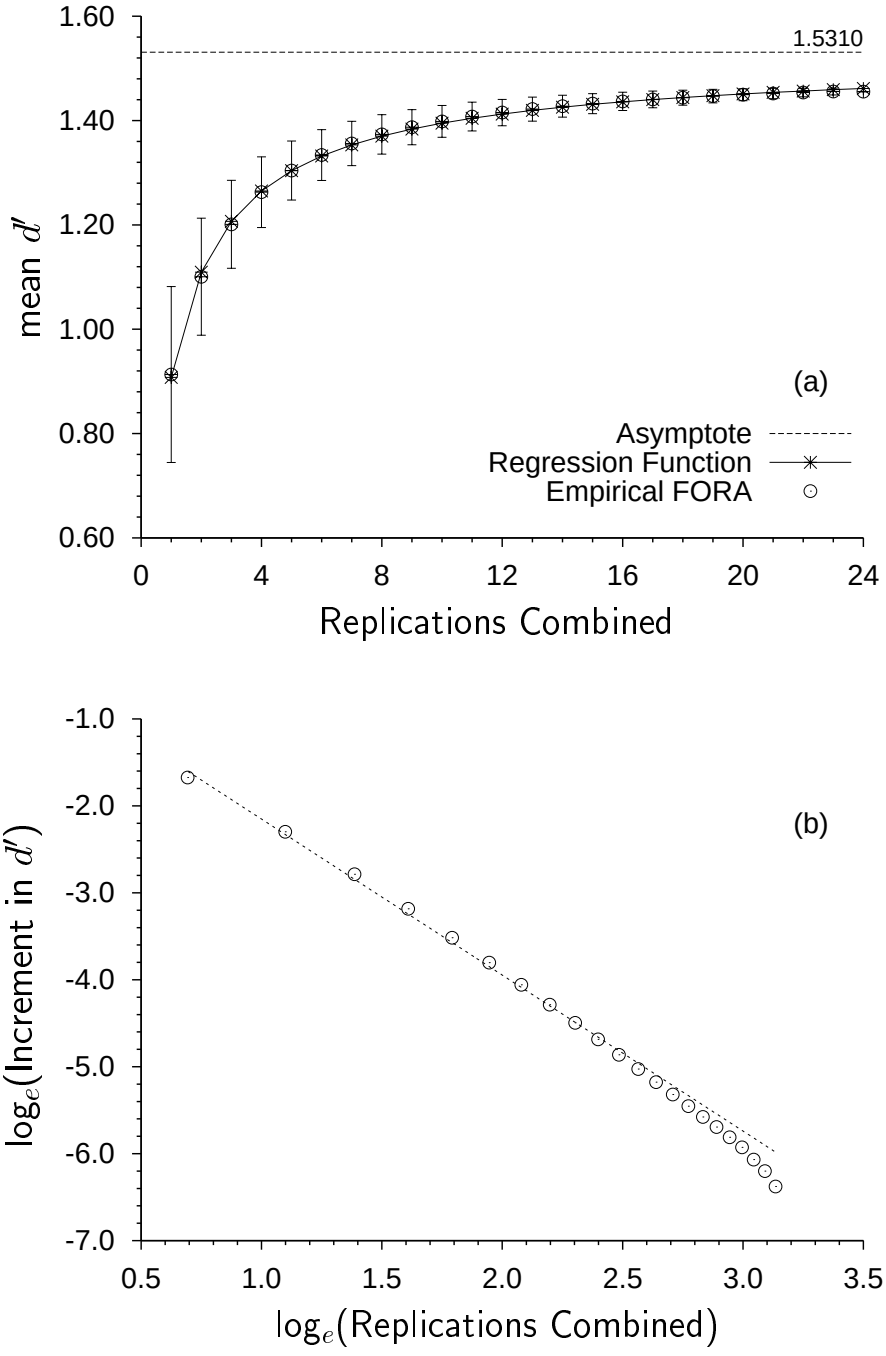


FIGURE 6.5: (a) Function of replications added, showing mean $d'$ as a function of replications combined, error bars of plus or minus one standard deviation, the fitted FORA for $d'$ and the estimated asymptote (horizontal line). The theoretical value of $d'$ was 1.4966. (b) The accompanying log-increment in $d'$ versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

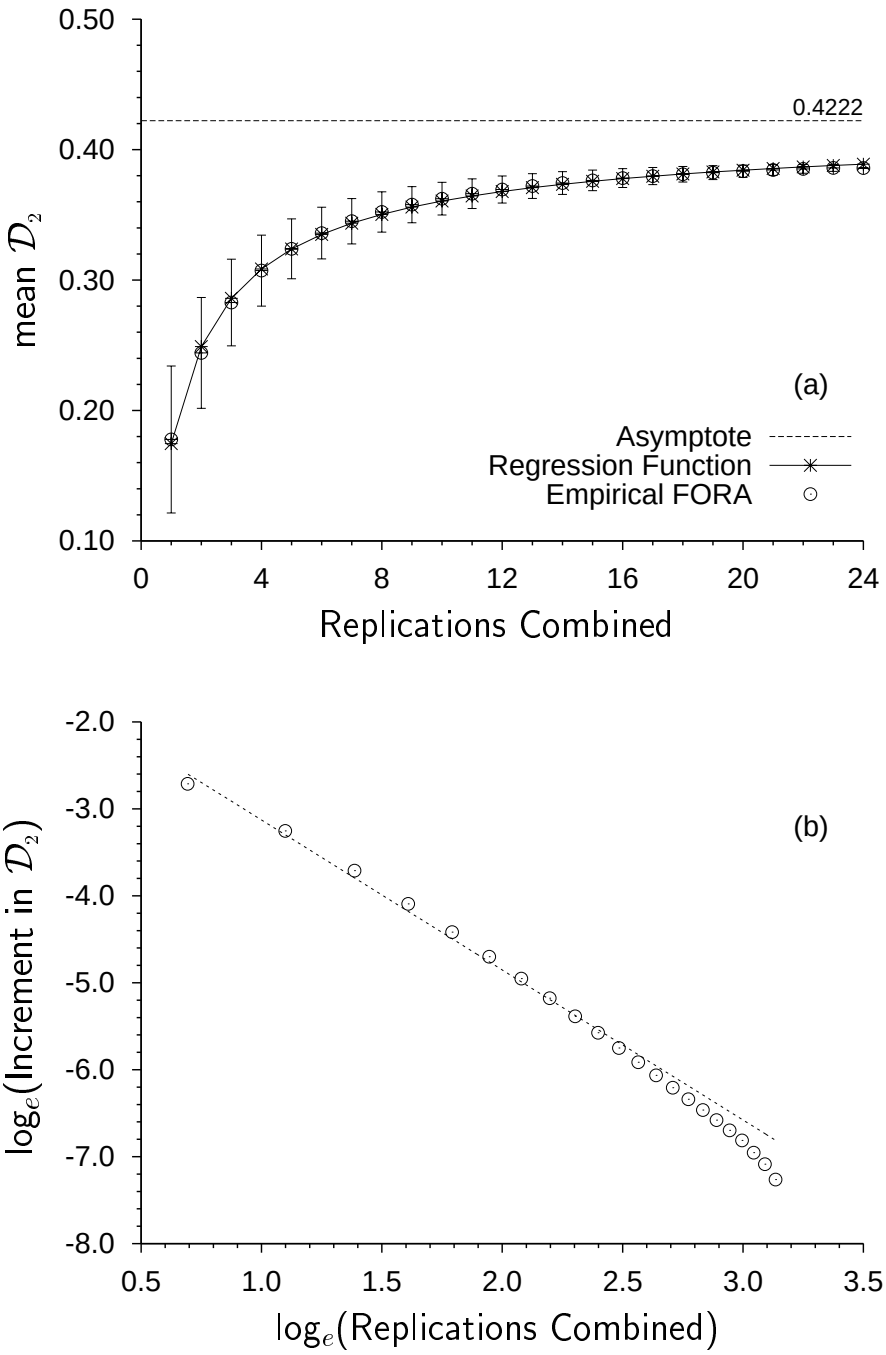


FIGURE 6.6: (a) Function of replications added, showing mean $\mathcal{D}_2$ as a function of replications combined, error bars of plus or minus one standard deviation, the fitted FORA for $\mathcal{D}_2$ and the estimated asymptote (horizontal line). The theoretical value of $\mathcal{D}_2$ was 0.4029 bits. (b) The accompanying log-increment in $\mathcal{D}_2$ versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

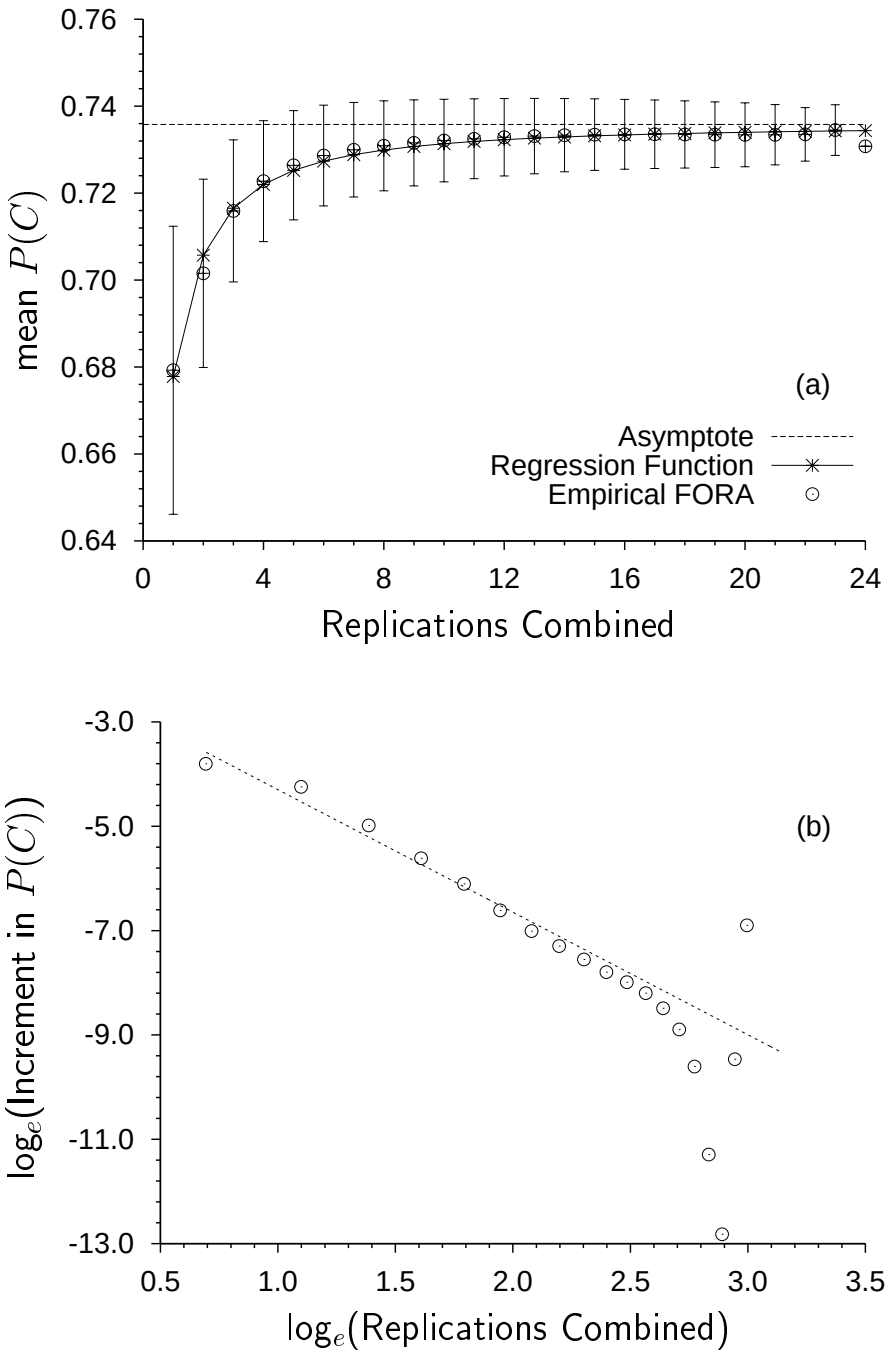


FIGURE 6.7: (a) Function of replications added, showing mean $P(C)$ as a function of replications combined, error bars of plus or minus one standard deviation, the fitted FORA for $P(C)$ and the estimated asymptote (horizontal line, at 0.7358). The theoretical value of $P(C)$ was 0.7308. (b) The accompanying log-increment in $P(C)$ versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

simple transformations of one to the other via the function that relates the two measures. Since points on a FORA are an average value, the situation is analogous to transform-averaging.[5] Consider for example the relationship between $d'$ and $\mathcal{A}$, which is given by $d' = \sqrt{2}\,\Phi^{-1}(\mathcal{A})$ (Equation 1.2). If the $n = {}^{m}C_{\xi}$ values of $\mathcal{A}$ for a given combination-size, $\xi$, are indexed as $\mathcal{A}_j$, then the mean value of $\mathcal{A}$ is $\overline{\mathcal{A}} = \frac{1}{n}\sum_j \mathcal{A}_j$. Similarly, the mean $d'$ value is $\overline{d'} = \frac{1}{n}\sum_j d'_j = \frac{1}{n}\sum_j \left(\sqrt{2}\,\Phi^{-1}(\mathcal{A}_j)\right)$, but $\overline{d'}$ is not necessarily equal to $\sqrt{2}\,\Phi^{-1}(\overline{\mathcal{A}})$, although they could be similar. For example, in Taylor et al.'s (1991) data set, $\overline{\mathcal{A}} = 0.7394$ and $\overline{d'} = 0.9132$ at $\xi = 1$, but $\sqrt{2}\,\Phi^{-1}(\overline{\mathcal{A}}) = 0.9072$. This result is typical of most of the FORAs presented in this chapter and the next, in that related measures are similar, in this sense, to 2 decimal places at $\xi = 1$, and become more so (to 4 decimal places) as $\xi$ increases to the number of replications in a data set.

The FORA based on $P(C)$ is shown in Figure 6.7. The FORA generally increased with added replications, but not as smoothly as the FORAs based on other measures. Consequently, the FORA regression function for $P(C)$ only provided a fair to middling fit to the data when compared to the regression fits for $\mathcal{A}$, $d'$ or $\mathcal{D}_2$ (Figures 6.3(b), 6.5 and 6.6 respectively). The variability in Figure 6.7(a) was reflected in the data points of the log-log plot in Figure 6.7(b), which were highly scattered ($r^2 = 0.8612$) compared to the log-log plots based on other measures. Ironically, the value of $P(C)$ for the 24-replication GOC curve was *identical* to the theoretical $P(C)$ value of 0.7308, because the 24-replication GOC curve happened to intersect the negative diagonal at exactly the same point as the theoretical ROC curve. The FORA was also unusual because the final point in Figure 6.7(a) stood out from the FORA trend.

Figures 6.1, 6.5(a) and 6.6(a) show that the standard deviation decreased as the number of replications increased. In the FORAs based on $\mathcal{A}$, $d'$ and $\mathcal{D}_2$, the standard deviation decreased sytematically towards zero at the larger combination-sizes, which reflected interdependence of GOC curves at larger combination-sizes. From a 24-replication data set, all GOC curves based on 23 replications are very similar because they are based on much the same data (subsets of 23 out of 24 replications). Hence the standard deviation of $\mathcal{A}$, $d'$ or $\mathcal{D}_2$, is small. The same is true, to a lesser extent, of GOC curves based on 22 replications, or 21 replications, and so on.

### Variability in the FORA based on $P(C)$

The FORA based on $P(C)$ in Figure 6.7 is notable because the standard deviation of $P(C)$ did not decrease as sharply as the FORAs based on $\mathcal{A}$, $d'$ and $\mathcal{D}_2$. The variability of $P(C)$ has to do with the way in which the measure is calculated. $P(C)$ is based on only a single point on an ROC or GOC curve, whereas $\mathcal{A}$ is calculated by using the entire GOC curve and, by extension, so are $d'$ and $\mathcal{D}_2$. As a result, $P(C)$ is more vunerable to

[5] Like the transform-averaging described in Chapter 3, except the quantity being averaged here is different.

the effects of sampling variability of both common and unique noise. Small fluctuations in a GOC curve may have minimal effect on the area under the entire curve, $\mathcal{A}$, but have a much larger effect on $P(C)$. If $d'$ were calculated from a single point in the ROC space, instead of from $\mathcal{A}$, it too would suffer the same type of extra variability that is associated with $P(C)$.

**Implications for $P(C)_{2IFC}$.** Although the data in Figure 6.7 is from an SIFC experiment, the same point holds true for 2IFC experiments. The proportion of correct decisions in a 2IFC experiment, $P(\mathrm{C})_{\mathrm{2IFC}}$, can be derived from a single point on a 2IFC ROC curve, and because of that, it suffers from the same drawbacks as $P(C)$ derived from a single point on an SIFC ROC curve. In any 2IFC experiment affected by observer inconsistency, $P(\mathrm{C})_{\mathrm{2IFC}}$ is a much more variable measure than $\mathcal{A}_{\mathrm{2IFC}}$.

**Possible solutions.** Two solutions to this problem are suggested, although neither has been implemented (since FORAs based on $P(C)$ are not often used in this thesis). One solution would be to run more replications, so that the number of combinations per combination-size becomes much larger. This would decrease the variability associated with each mean-$P(C)$ value and lead to a smoother average-FORA, mainly in the middle points of the FORA. Although running more replications is fine in principle, it may not be so useful in practice. Increasing the number of replications much beyond 24 replications will lead to computation-time limitations, because the number of GOC curves increases exponentially with the number of replications.

A second solution would be to fit an ROC curve to each of the GOC curves calculated in ACA, and to calculate measures and parameters from the fitted curve (Taylor, 1984), including $P(C)$, or $\mathcal{A}$. This solution may complicate and slow down FORA calculation, depending on the fitting procedure, form of the fitted curve, and the number of points on each GOC curve. Fitting ROC curves could potentially bias results if the form of the fitted curve is not similar to that of the GOC curves. At a given combination-size, GOC curves are often as variable in shape and location as single-replication ROC curves. Any particular form of ROC curve may successfully fit some GOC curves but not others. Since ACA may involve many thousands of GOC curves, the assessment of bias and error due to ROC regression would be difficult, to say the least.

### FORA patterns based on various measures

Many of the patterns shown in the FORA results for Taylor et al.'s (1991) experiment also hold throughout the various experiments presented in the next two chapters. The similarity in form between the FORAs and log-log plots based on $\mathcal{A}$ and those based on $d'$ and on $\mathcal{D}_2$ is very robust across experiments. This is not surprising since both $d'$ (as used

here) and $\mathcal{D}_2$ are functions of $\mathcal{A}$. Any curvature in a log-log plot based on $\mathcal{A}$ is usually accentuated slightly in the log-log plots based on $d'$ and on $\mathcal{D}_2$.

The patterns of ACA standard deviations found in Taylor et al.'s (1991) data are also very robust across experiments. Standard deviations based on $\mathcal{A}$, $d'$ and $\mathcal{D}_2$ always decrease smoothly as a function of combination-size because the contributing performance values are so constrained and interdependent (due to the resampling scheme in ACA). Standard deviations based on $P(C)$ also decrease (to a lesser degree), but are always quite large because of the way the measure is calculated. The variability in $P(C)$ means that $\mathcal{A}$, $d'$ and $\mathcal{D}_2$ is used in preference to $P(C)$, even for 2IFC experiments.

None of the standard deviations associated with FORA points are useful in determining where the asymptote may lie—the asymptote could lie within a wide range of $P(C)$ values in Figure 6.7, and the asymptote is generally several standard deviations away from the FORA points for $\mathcal{A}$, $d'$ and $\mathcal{D}_2$ in Figures 6.1, 6.5(a) and 6.6(a). The standard deviations are not useful for placing sensible error bounds *on the asymptote*, since their value varies with the number of replications combined. For these reasons, standard deviations associated with FORA points are not considered beyond here.

## 6.3 Summary

Group operating characteristic curves may be calculated for all combinations of replications from a multiple-replication data set. This process is called *all combinations analysis* (ACA). The average sensitivity per combination-size defines a *function of replications added* (FORA), which generally increases as more replications are added. A FORA begins at the average ROC performance level and tends asymptotically towards theoretical performance.

Experimental FORAs were approximated quite well by a three-parameter data model of the form

$$A_i \quad = \quad A_1 + \kappa \sum_{j=2}^{i} j^{\mu}, \qquad i \geqslant 2$$

(Equation 6.5), where $A_1$, $\kappa$ and $\mu$ are parameters whose values derive from a data set, and where $A_i$ represents the mean sensitivity value for GOC curves based on $i$ replications. It is possible to fit this function to data by using non-linear least-squares regression. The fit of the *regression-FORA* may be visually assessed by means of a *log-log plot*, which shows FORA-increments versus number of replications, plotted in double-logarithmic coordinates. Computational details about ACA and FORA regression are given in Appendices E and F.

Extrapolation of the data model, based on a finite data set, makes it possible to estimate asymptotic, unique-noise-free performance. The asymptote can be very close to

theoretical performance, both of which are considerably better than single-replication ROC performance. A small discrepancy between the asymptotic and theoretical performance is explained in terms of the curvature of the log-log plot data series, relative to a straight line that is derived from the data model.

For the same data set, each different measure of sensitivity results in its own FORA. The amount of variability associated with each point on a FORA, and how smoothly the FORA increases, depends on the measure and how it is calculated. FORAs based on measures that use an entire GOC curve are much less variable than those based on measures based on only a single point on the GOC curve.

# Chapter 7

# Sampling statistics of asymptotic performance

Asymptotic discriminability based on FORA regression only provides an *estimate* of unique noise-free performance. If two or more independent sets of replications are analysed from the same experiment, estimates of the asymptote will vary across sets. Hence, there is sampling variability and error associated with empirical asymptotes. Asymptotes may be estimated from as few as three replications but there is no upper limit to how many replications can be combined. When the number of replications is small, the error associated with the estimate is relatively large. As more replications are combined in ACA, the estimate of the asymptote becomes stable, and the error associated with it decreases to zero. A very large data set is used to show how these sampling statistics can be estimated. Practical problems arising in ACA of large data sets are described, and solutions are given.

### Overview of chapter

The data set that is analysed in this chapter is from an SIFC amplitude discrimination experiment by Whitmore et al. (1993), obtained through the courtesy of the authors. One observer ran 75 replications and a second observer ran 25 replications. Experimental methodology is described in Section 7.1, and ROC and GOC results are given in Section 7.2. ROC curves varied from replication to replication, and GOC curves showed better performance than mean ROC curves. FORA results are presented in Section 7.3. Combinatorial explosion made it impossible to compute ACA on such a large data set, but two workable solutions to this problem are described. The FORA regression equation (Equation 6.5) is shown to provide an excellent description of an empirical FORA out to 75 replications. It is also shown that the observer with the best average ROC performance does not necessarily have the best asymptotic performance. A FORA is presented based on data from both observers, which shows that FORA regression works even when each observer has a different level of common noise, and potentially different transfer functions

and scaling of the rating scale. Section 7.4 shows how to estimate the sampling statistics of asymptotes, and limitations of the estimation process are discussed.

## 7.1 Method

The experimental task was to detect whether or not a narrowband Gaussian noise signal had been added to a wideband Gaussian noise masker. Decisions were made using a continuous rating scale slider. Unlike Taylor et al.'s (1991) experiment in the preceding chapter, the theoretical ROC curve was unknown.

**Observers.** The observers were two adult males, each of whom had recently completed over 10 000 trials in a similar SIFC amplitude discrimination experiment. Both men were also observers in Taylor et al.'s (1991) continuous rating scale experiment.

**Stimuli.** The maskers for the experiment were short-duration, low-pass Gaussian noise transients with a 3dB bandwidth of 1980 Hz (nominally 2000 Hz). The signals were short-duration, band-pass Gaussian noise transients with a 3 dB bandwidth of 100 Hz, centred at 500 Hz. Both the signals and maskers were of the same duration and were windowed using a Kaiser window (Rabiner & Gold, 1975) with a shape parameter of 9. The absolute duration was 20 ms and the equivalent rectangular duration was 8.2 ms. The signal-to-noise ratio was 7.5 dB, with the gated masker presented at a spectrum level of 80 dB SPL. During experimental sessions, a 2 kHz low-pass analog Gaussian noise masker ran continuously at a spectrum level of 20 dB SPL. All stimuli were presented diotically.

Transients were computer-generated, and stored as digital code sequences on disk, so they could be reproduced across replications. The signal-generation program was written in HPL (Hewlett Packard Language) and implemented on an HP 9826 computer. The program used digital Butterworth filters (Stearns, 1975) to generate band-limited Gaussian noise. A two-section filter was used for the low-pass masker process, with a nominal cutoff of 2000 Hz. A 4-section band-pass filter was used for the signal-alone transients with nominal cutoffs of 475 Hz and 525 Hz. The input to each filter was 5 kHz low-pass, white Gaussian noise.[1] To generate signal-plus-noise transients, sequences of 200 consecutive points were copied from the output of each digital filter, additively mixed, windowed, and converted into 12-bit digital code sequences. Noise-alone transients were generated from the low-pass filter only. Each filter was run through 200 points before the next transients

[1] These nominal cutoffs, and the given data window, produced the measured 3 dB bandwidths of 1980 Hz for noise-alone transients and 100 Hz for signal-alone transients. The white Gaussian noise input for each filter was defined by a sequence of normally distributed pseudo-random numbers, presented at a nominal rate of 10 kHz, which was the clocking rate of the digital-to-analog converter used in the experiment. Normal variates were generated by applying Knuth's polar algorithm (Knuth, 1969; cited in Taylor, 1984) to the numerical output of Evans et al.'s (1967) uniform pseudo-random number generator.

were copied. Instantaneous values of the filter outputs were Gaussian out to 3 standard deviations from the mean.

The code sequences were processed by a 12-bit digital-to-analog converter card housed in a Hewlett-Packard 6940B multiprogrammer, under the control of the HP 9826 computer. The digital-to-analog converter was clocked at 10 kHz. Its output was smoothed using a passive 3 kHz low-pass filter before being attenuated, mixed with the continuous masker and passed to a headset amplifier. Stimuli were presented to observers using TDH-39 headphones mounted in Rudmose Tracor RA-125 headsets with MX-41/AR cushions. Observers sat in a booth in the same sound-attenuated chamber that was used for Taylor et al.'s (1991) experiment, described in previous chapters.

**Procedure.** The experiment involved multiple 1000-trial replications, with 500 trials per event. Stimuli were presented in a haphazard sequence, under the constraint that the same event could not occur more than four trials in a row. Each replication for each observer used a different trial sequence, so that any trial-by-trial order effects would contribute to the unique noise and not the common noise. Trials were run in sessions of 250 trials that lasted approximately 10 minutes. Only one observer ran at a time.

Decisions were indicated using a continuous 12 cm horizontal slider as a mechanical analog to a rating scale. Slider position was measured electronically at the end of the decision interval and converted into ratings on a 36-point rating scale (Watson et al., 1964). Each trial began with the slider reset to the extreme left. Observers used the slider so that its final position reflected their confidence that the *SN* event had occurred, with increasing position from left to right indicating increasing confidence. Observers were encouraged to try to use the whole scale evenly.

Each trial consisted of a warning interval of 100 ms, an observation interval of 20 ms, a decision interval of 1250 ms and a reset interval of 930 ms. The reset interval was a minimum duration. The next trial could not begin until the slider had been reset to its extreme left. A set of LED lights on the decision panel were switched on and off to mark the trial intervals. Trial-by-trial knowledge of results was provided during the reset interval using LEDs that indicated which event had occurred during the observation interval. The timing of intervals and data collection was under the control of an HP 9825 computer, but stimulus production and data storage onto disk was done by an HP 9826 computer.

Observer 1 completed 25 replications over the course of six weeks while Observer 2 completed 75 replications over the course of three months. Observers completed anywhere from zero to four replications per day. The initial plan was for observers to run 25 replications each, which they did, but the plan changed and Observer 2 continued and ran 75 replications. After they had completed 25 replications each, the observers repeated some of their poorer replications, which had been earmarked as being dodgy at the time they were run, for a variety of reasons. Observer 1 repeated nine of his first 25 replications,

and Observer 2 repeated four of his first 25 replications. After this, no further replications were repeated.

## 7.2 ROC and GOC results

The single-replication ROC curves for Observers 1 and 2 are shown in Figure 7.1. Both observers were comparable in their level of performance and the amount of variability across replications. The mean value of $\mathcal{A}$ for Observer 1 was 0.8017, with a standard deviation of 0.0133, while the mean value of $\mathcal{A}$ for Observer 2 was 0.7922, with a standard deviation of 0.0142.

Figure 7.2 shows the mean ROC and GOC curves for the ROC curves in Figure 7.1. Although Observer 1 had the higher mean ROC curve, Observer 2 had the higher GOC curve. According to Figure 7.2, which observer was better at the task depended on whether unique noise was removed or not.

Figure 7.3 shows the mean ROC and GOC curves for Observer 2 based on three successive 25-replication blocks of data. While the mean ROC curves are reasonably similar, the GOC curves are less so, which shows that unique noise was still present after 25 replications were combined. If all the unique noise had been removed, then all three GOC curves should be the same, because the same stimulus set was used for all replications. Observer 1 had a higher mean ROC curve than any of the 25-replication mean ROC curves for Observer 2, but a lower GOC curve than any of the 25-replication GOC curves for Observer 2. This indicates that the number of replications was not the reason why the 75-replication GOC curve for Observer 2 was better than the 25-replication GOC curve for Observer 1. Rather, the common noise performance was different across the two observers. This difference is also seen in the FORA results, which are given in the next section.

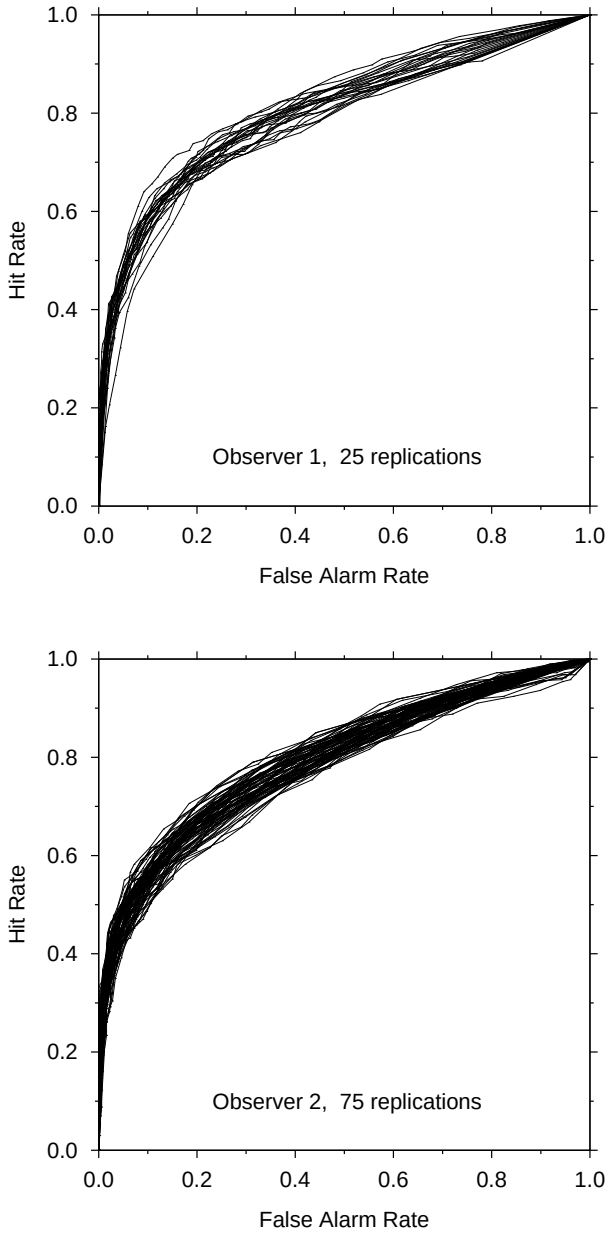


FIGURE 7.1: ROC curves from Whitmore et al.'s (1993) SIFC amplitude discrimination experiment. Upper panel: all 25 single-replication ROC curves for Observer 1, lower panel: all 75 single-replication ROC curves for Observer 2.

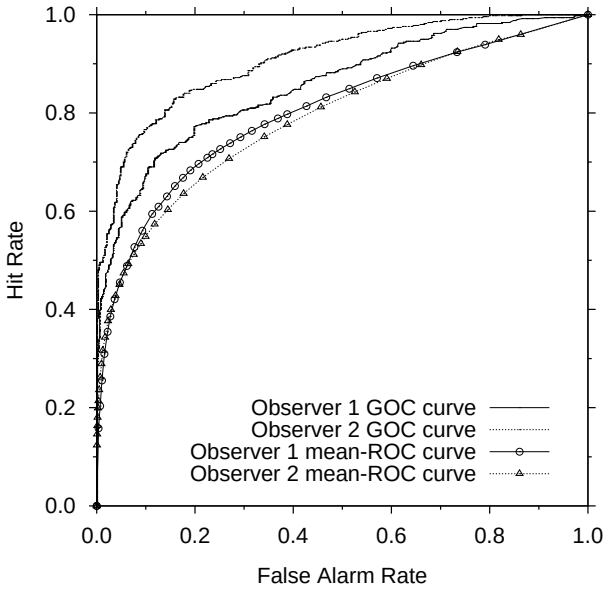


FIGURE 7.2: Mean ROC and GOC curves for each observer, based on 25 replications for Observer 1, and 75 replications for Observer 2.

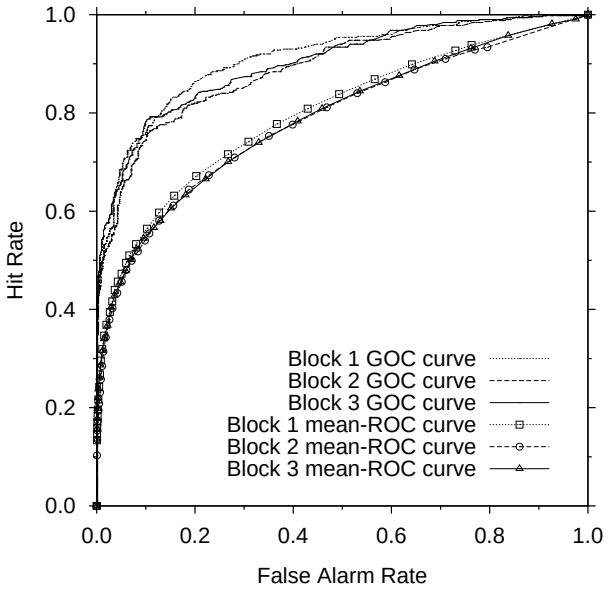


FIGURE 7.3: Mean ROC and GOC curves for Observer 2 based on successive 25-replication blocks.

## 7.3 FORA results

Various FORAs are presented in this section, starting with the 25-replication FORA for Observer 1 and three successive 25-replication FORAs for Observer 2. This is followed by estimates of the 75-replication FORA for Observer 2, which illustrate practical problems with ACA for large numbers of replications, and solutions to those problems. A 50-replication FORA based on 25 replications from each observer is then presented, which shows the result of combining data across observers. Finally, Section 7.4 deals with the sampling statistics of estimated asymptotes, including residual error, and addresses the question of how many replications should be run in GOC experiments. All of the FORAs and asymptotes presented in this chapter are based on $\mathcal{A}$. Data values and regression parameters for the various FORAs are given in Table G.2 in Appendix G.

### 7.3.1 FORAs based on 25 replications

**Observer 1.** The 25-replication FORA for Observer 1, and its accompanying log-log plot, are presented in Figure 7.4. The average value of $\mathcal{A}$ for single-replication data is 0.8017, which improves to 0.8567 after 25 replications have been combined. The regression-FORA in Figure 7.4(a) is given by Equation 6.5 using the parameter triplet $(A_1, \kappa, \mu) = (0.802134, 0.093299, -2.030904)$.

The regression-FORA fitted most of the data well. The regression-FORA was slightly higher than the empirical FORA at the middle combination-sizes, and slightly lower at the larger combination-sizes. This was reflected in the accompanying log-log plot, which curved upwards compared to the straight line implied by FORA regression. The log-log plot shows that the increments in the regression-FORA became smaller than the increments in the empirical FORA as the number of replications combined increased. This pattern suggests that the estimated asymptote, at 0.8596, underestimates the true asymptote, but probably not by much. The log-log plot was reasonably linear, with $r^2 = 0.9960$.

**Observer 2.** FORAs for Observer 2, based on three successive blocks of 25 replications, are shown in Figure 7.5. The three GOC curves based on the same three data blocks were shown in Figure 7.3. The GOC curves were different from each other and so were the FORAs and asymptotes. FORAs started off at 0.7995, 0.7882 and 0.7891 for blocks 1, 2 and 3, respectively. GOC analysis provided large improvements after 25-replications, with values of $\mathcal{A}$ of 0.9121, 0.8919 and 0.9018 respectively. Regression functions were fitted to the empirical FORAs (according to the parameters given in Table G.2 in Appendix G), providing estimated asymptotic values of 0.9270, 0.9047 and 0.9168 respectively.

Like the GOC curves in Figure 7.3, the variability across FORAs implies there is still unique noise present (even after 25-replications have been combined), and this affects the resulting asymptotes. These FORA results show that estimated asymptotic performance

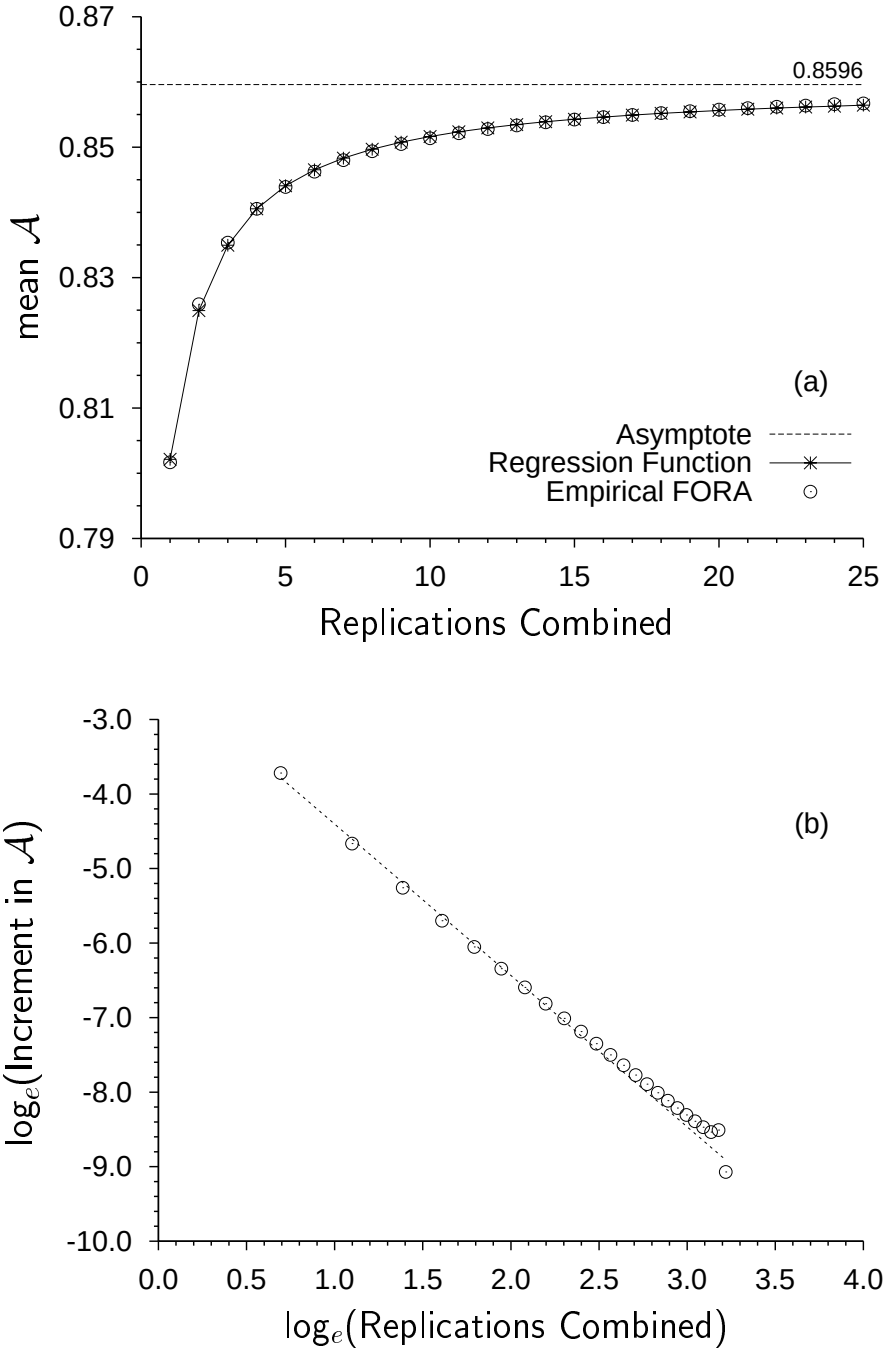


FIGURE 7.4: (a) The 25-replication FORA for Observer 1, showing mean area under the GOC curve as a function of replications combined. Endpoints of the solid line segments are the fitted FORA. (b) The accompanying log-increment in area versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

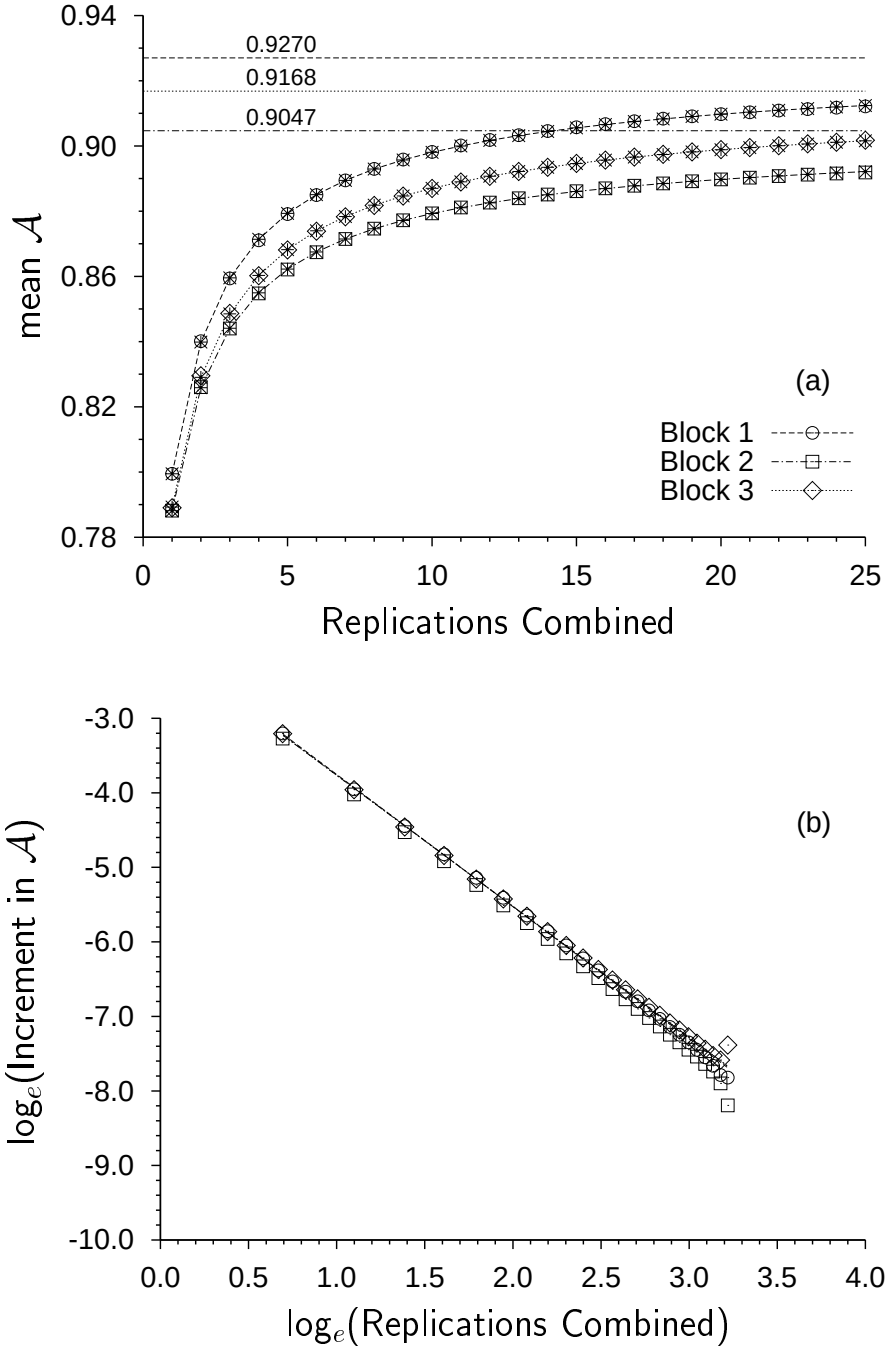


FIGURE 7.5: (a) FORAs for three successive blocks of 25 replications for Observer 2, showing mean area under the GOC curve as a function of replications combined. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (b) The accompanying log-increment in area versus log of replications combined for each of the three blocks, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each block. Symbols and line-types denote the same data blocks as for panel (a).

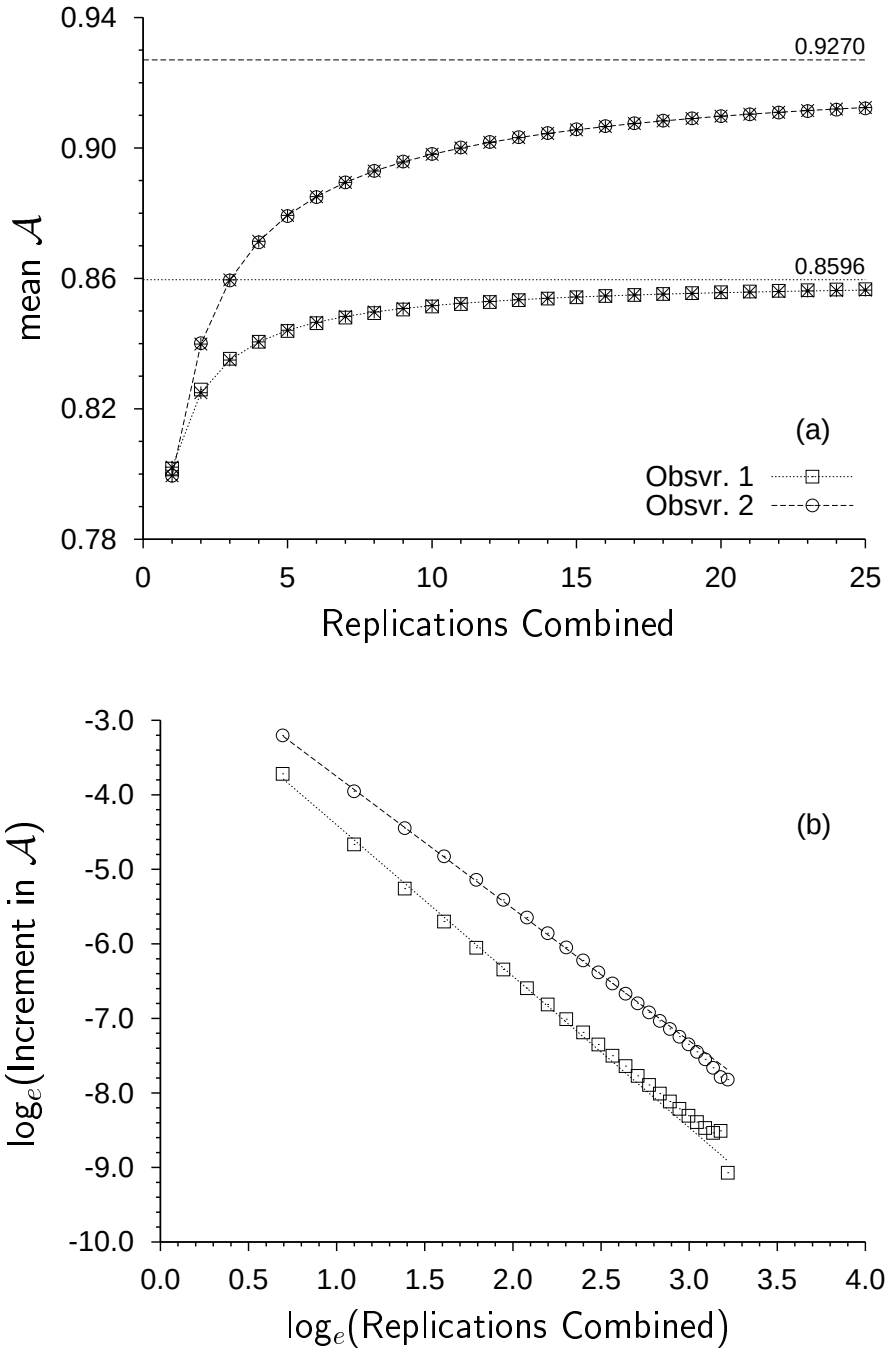


FIGURE 7.6: (a) FORAs based on the first 25 replications from each observer, showing mean area under the GOC curve as a function of replications combined. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (b) The accompanying log-increment in area versus log of replications combined for each observer, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each observer. Symbols and line-types denote the same observer as for panel (a).

does not *unambiguously* reflect unique-noise-free performance, no matter how good the regression may be for a given data block. The variability, or error, associated with estimated asymptotes is discussed in Section 7.4.

**A comparison between Observers 1 and 2.** The 25-replication FORA for Observer 1 (from Figure 7.4) and the 25-replication FORA for Observer 2's first data block (from Figure 7.5) are re-presented for comparison in Figure 7.6. The two FORAs cross between combination-sizes of 1 and 2, which shows again that Observer 1 had the better ROC performance, and Observer 2 had the better GOC performance. Each observer's FORA tended to quite different asymptotic values, which suggests that the discrepancy in GOC performance between GOC curves shown in Figure 7.2 was because the common noise is different for each observer, rather than because Observer 2 ran 50 more replications than Observer 1. The FORAs also suggest that Observer 1 would never catch up to Observer 2 because the unique-noise-free performance for Observer 1 was inherently worse than that of Observer 2.

Conclusions such as these may be hard to draw based on ROC or GOC curves alone, without calculating FORAs, regardless of whether the number of replications per observer are equal or not. If comparing observers on the basis of just one GOC curve per observer, any differences across observers could be due to differences in common noise, or due to differences in unique noise, or both. Without describing how performance changes with replications, it may *seem* possible that the observer with the worse GOC curve could run more replications and catch up to the observer with the better GOC curve. Calculating a FORA for each observer can help to show when this is possible and when it is not.

### 7.3.2 FORAs based on 75 replications

Combinatorial explosion puts a practical limit on the number of replications that can be analysed in full using ACA. The limit arises because there are $(2^m - 1)$ possible GOC curves from a data set of $m$ replications. Once $m$ becomes too big, even the fastest computer is unable to complete ACA in a reasonable amount of time.[2] Two possible solutions to this problem are suggested and applied to the 75-replication data set.

The first solution, called *partial-ACA*, is to only analyse data for combination-sizes from 1 up to the largest size computable in a reasonable amount of time, along with complementary combination-sizes. This results in a partial FORA, with outer points defined at each end, but with inner points missing from the middle. It is still possible to fit a regression-FORA to the partial FORA and to estimate an asymptote. The second

[2] Each of the 25-replication FORAs presented earlier required calculating the area under $(2^{25} - 1) \simeq 33.5$ million GOC curves. At the date of analysis, this was close to a practical limit, because of computation time. Full ACA on a 75-replication data set requires $(2^{75} - 1) \simeq 3.8 \times 10^{22}$ GOC curves, and would take a while.

solution, called *sampled-ACA*, is to use partial-ACA to derive as many outer points as possible, and to estimate values for the inner points by random sampling. ACA, when done in full on all combination-sizes, is called *complete-ACA*, to distinguish it from these other two variants. The term ACA without a qualifier implicitly refers to complete-ACA.

**Partial-ACA.** For each combination of $\xi$ replications taken from a set of size $m$, there is a complementary combination of $m - \xi$ replications. Calculating sums-of-ratings for the former combination provides an efficient way of calculating sums-of-ratings for the latter combination. It is more efficient to calculate GOC measures in ACA by pairing combinations with their complements than by treating all combinations separately. ACA results are identical, regardless of whether complementary combinations are used or not. Specific details about the computation are given in Appendix F.

The result of using complementary combinations is that FORA points are calculated in pairs, starting with $\xi = 1$ and $\xi = m - 1$, and working inwards from the outside. Values of $\mathcal{A}$ were only calculated using complementary combinations of sizes $\xi = 1$ to 6 and 69 to 74 (calculation time became prohibitive beyond $\xi = 6$). The last point on the FORA, at $\xi = m = 75$, was also calculated. It is a special case, based on a single GOC curve, because there is no combination that is complementary to all 75 replications.

The outer points of the 75-replication FORA based on $\mathcal{A}$ are presented in Figure 7.7, along with the accompanying log-log plot. The average area under the ROC curve is 0.7922, which improves to 0.9099 after 75 replications have been combined. Only a handful of outer points on a FORA are enough to fit Equation 6.5 to the data, estimate an asymptote and to also calculate some of the points on the a log-log plot. The regression-FORA in Figure 7.7(a) is given by Equation 6.5 using the parameter triplet $(A_1, \kappa, \mu) = (0.792340, 0.134380, -1.775623)$, which estimates the asymptote to be 0.9160. The accompanying log-log plot (Figure 7.7(b)) is consistent with the linear relationship implied by the FORA parameters (dashed line). The difference between the area under the 75-replication GOC curve and the estimated asymptotic area value is 0.006, which shows that most, but not all the unique noise has been removed after 75 replications.

**Sampled-ACA.** Sampled-ACA is another way of deriving FORA results without needing to work through *all* $(2^{75} - 1)$ combinations. Partial-ACA is first run to derive FORA values for outer combination-sizes, as shown in Figure 7.7, and random sampling is then used to *estimate* average values of $\mathcal{A}$ for inner combination-sizes (to fill in the gap in Figure 7.7). Sampled-ACA was applied to the data for Observer 2. For this data set, at each combination-size from $\xi = 7$ to 68 inclusive, $2^{20} = 1\,048\,576$ combinations of size $\xi$ were randomly sampled, with replacement, from the set of ${}^{75}C_{\xi}$ possible combinations, and the sample mean value of $\mathcal{A}$ was calculated.

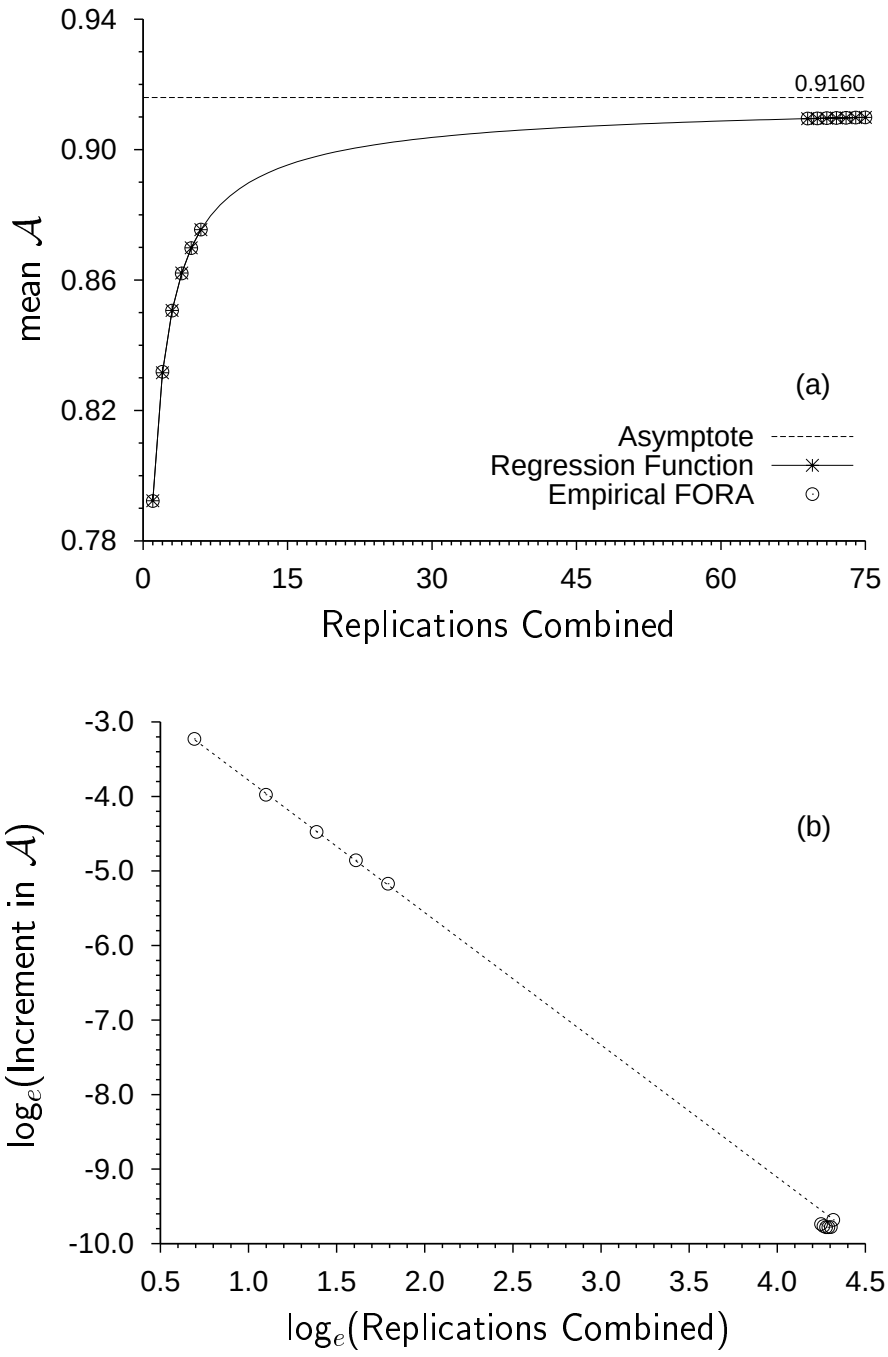


FIGURE 7.7: (a) The outer points on the 75-replication FORA for Observer 2, showing mean area under the GOC curve as a function of replications combined. All combinations were used for combination-sizes from 1 to 6 and from 69 to 75. (b) The accompanying log-increment in area versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

The estimated 75-replication FORA and its accompanying log-log plot are presented in Figure 7.8. Data points for combination-sizes 1 to 6 and 69 to 75 are the same as in Figure 7.7. The regression-FORA in Figure 7.8(a) is given by Equation 6.5 using the parameter triplet $(A_1, \kappa, \mu) = (0.791970, 0.136774, -1.783845)$. This provided an asymptote at 0.9160, agreeing with the earlier estimate to four decimal places. The empirical FORA in Figure 7.8(a) was remarkably smooth and the regression-FORA was almost a perfect fit to the data. It should be emphasised that *the solid line through the data points represents a regression function. It does not join the experimental data points.* (Very close scrutiny of Figure 7.8(a) will reveal tiny discrepancies between the small central dot in each data point and the regression function.) The empirical log-log plot in Figure 7.8(b) was almost linear, with $r^2 = 0.9992$. The data series in the log-log plot was not entirely smooth, however, which reflects the fact that combinations were sampled rather than calculated exhaustively.

The stability of the estimated points in the middle of the FORA is influenced by the number of random samples per point. If only 1000 combinations had been sampled, for example, instead of $2^{20}$, the estimated FORA and log-log plot would not be as smooth, because of greater sampling variability. The influence of the number of samples on data analysis was not investigated, but $2^{20}$ was used as a compromise between stability and practical computability.

The FORA based on partial-ACA of just the outer points and the FORA based also on sampled inner points are consistent with each other, and their estimated asymptotes agree to four decimal places. The agreement comes about because the empirical log-log plot was very linear for this data set. Had the log-log plot curved appreciably, then sampled-ACA would have provided a better estimate than partial-ACA, because the sampled inner points would have provided more data for the regression-FORA.

Complete-ACA is certainly practical for sets up to about 25 replications, but if a data set is too large for complete-ACA, then sampled-ACA is preferable to partial-ACA because it provides more certainty in FORA regression. The main disadvantage of sampled-ACA is that it takes longer to compute than partial-ACA. Sampled-ACA required about 70 million extra GOC curves for the 75-replication FORA compared to partial-ACA. Ultimately, computation time is a limiting factor in the choice of analysis.

### 7.3.3 Combining data across observers

GOC analysis can be performed on replications analysed within single observers, or across observers or both. Experiments involving more than one observer can be used to assess group performance, as long as the same stimulus set is used for each observer. The theory of GOC analysis shows that when the decision axes of two or more inconsistent observers are identical, or the axes share the same stochastic ordering, then all individual observers have the same asymptotic, unique-noise-free GOC curve. Furthermore, the asymptotic

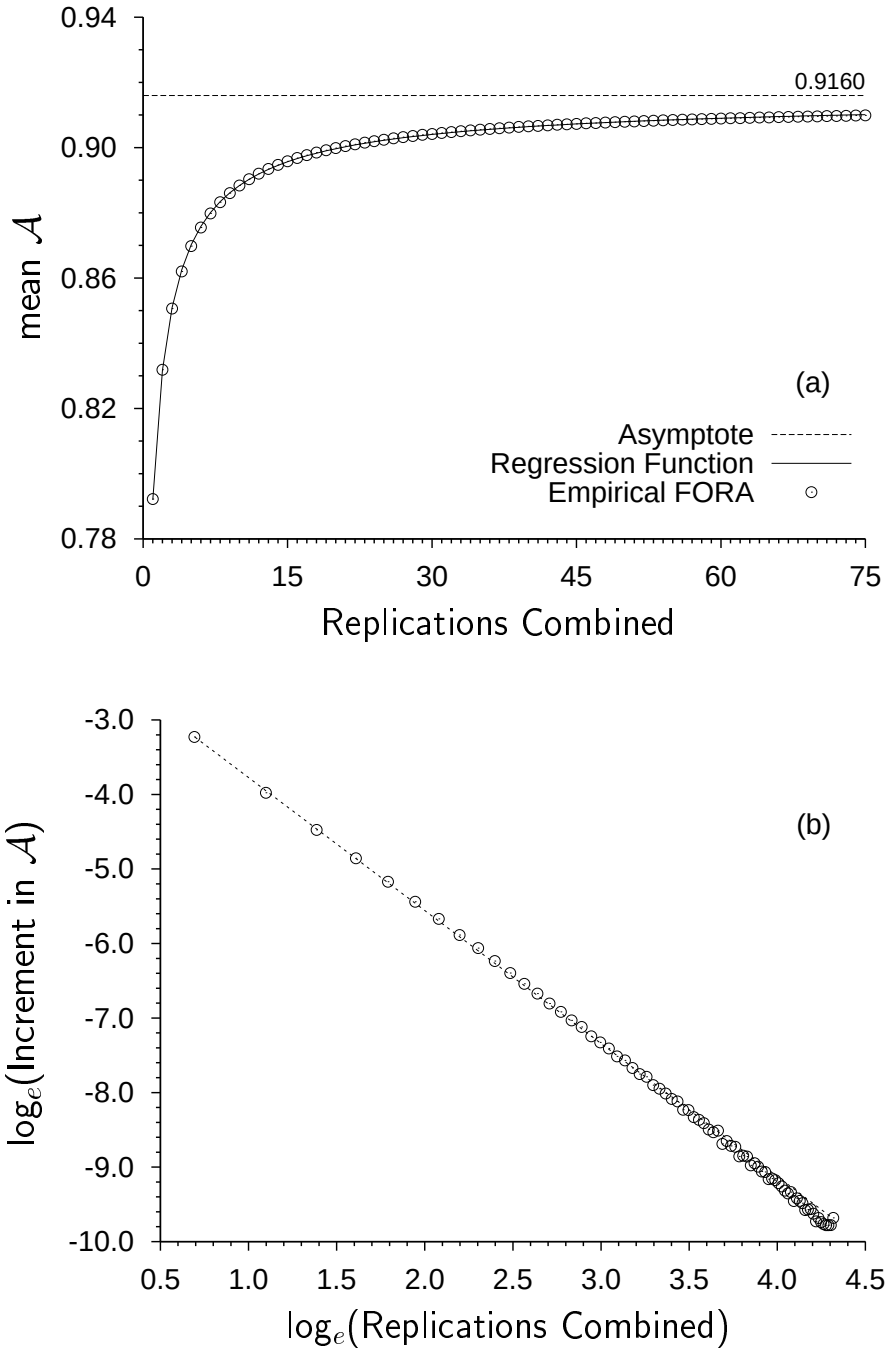


FIGURE 7.8: (a) Estimated 75-replication FORA for Observer 2, showing mean area under the GOC curve as a function of replications combined. Values for combination-sizes 1 to 6 and 69 to 75 were calculated exactly, and values for combination-sizes 7 to 68 were estimated by random sampling. (b) The accompanying log-increment in area versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

value of $\mathcal{A}$ will be the same across observers as within observers.

Figure 7.6 showed that the asymptotic value of $\mathcal{A}$ was different for each observer, which implies that the asymptotic GOC curves would be different also. This does not mean that GOC analysis and ACA cannot be done across observers, because GOC curves can be calculated. In such a situation, GOC and FORA results call for a cautious interpretation, because it is uncertain what is common across observers and what is unique.

The 25-replications from Observer 1 and the first 25-replications from Observer 2 were combined to estimate a 50-replication FORA. This required $(2^{50} - 1) \simeq 10^{15}$ values of $\mathcal{A}$, which is impractically large for complete-ACA, and so sampled-ACA was done instead. Partial-ACA was computed to provide outer points on the FORA for combination-sizes 1 to 4 and 46 to 50. Mean values of $\mathcal{A}$ were estimated for combination-sizes from 5 to 45 inclusive (based on $2^{20} = 1\,048\,576$ combinations sampled at each combination-size). The result was an estimated 50-replication FORA, which is given in Figure 7.9 with its accompanying log-log plot. The 50-replication regression-FORA provided an excellent fit to the data, apart from the first few points. The average value of $\mathcal{A}$ was 0.8006 initially, and reached 0.8851 after 50 replications, just short of the estimated asymptote at 0.8884.

In Figure 7.6(b), the log-log plot for the first 25 replications for Observer 2 curved slightly downwards whereas that for Observer 1 curved upwards. These trends seemed to cancel each other out, roughly speaking, once the two sets of data were combined. The log-log plot based on all 50 replications fell almost exactly on the linear relationship implied by the FORA parameters (dashed line in Figure 7.9(b)), with $r^2 = 0.9968$ for the data points. The relatively poor fit of the regression-FORA to the initial FORA points in Figure 7.6(a) showed up as only slight discrepancies in the log-log plot (between data points and the straight line in Figure 7.6(b)), and did not appear to affect regression of the rest of the FORA data points. The linear log-log pattern broke down for the last few data points in the log-log plot, which dropped off more steeply than the straight line. The effect of this small bend on the regression-FORA was minimal (with discrepancies of less than $e^{-9} \simeq 0.0001$).

Since FORA data points are average values, it might be expected that the FORA in Figure 7.9(a) is approximately the average of the two individual FORAs shown in Figure 7.6. This was certainly the case, even though no simple averaging was taking place in the estimation of either the 50-replication FORA or the asymptotic value. The asymptote estimated from the 50-replication FORA was 0.8884, which was close to the average of the two individual 25-replication asymptotes at 0.8933.

FORA regression using the three-parameter data model provided excellent descriptions of Whitmore et al.'s (1993) ACA results. Each observer showed different levels of unique noise and common noise. Group results across observers were intermediate to each individual's results, which seems intuitively correct.

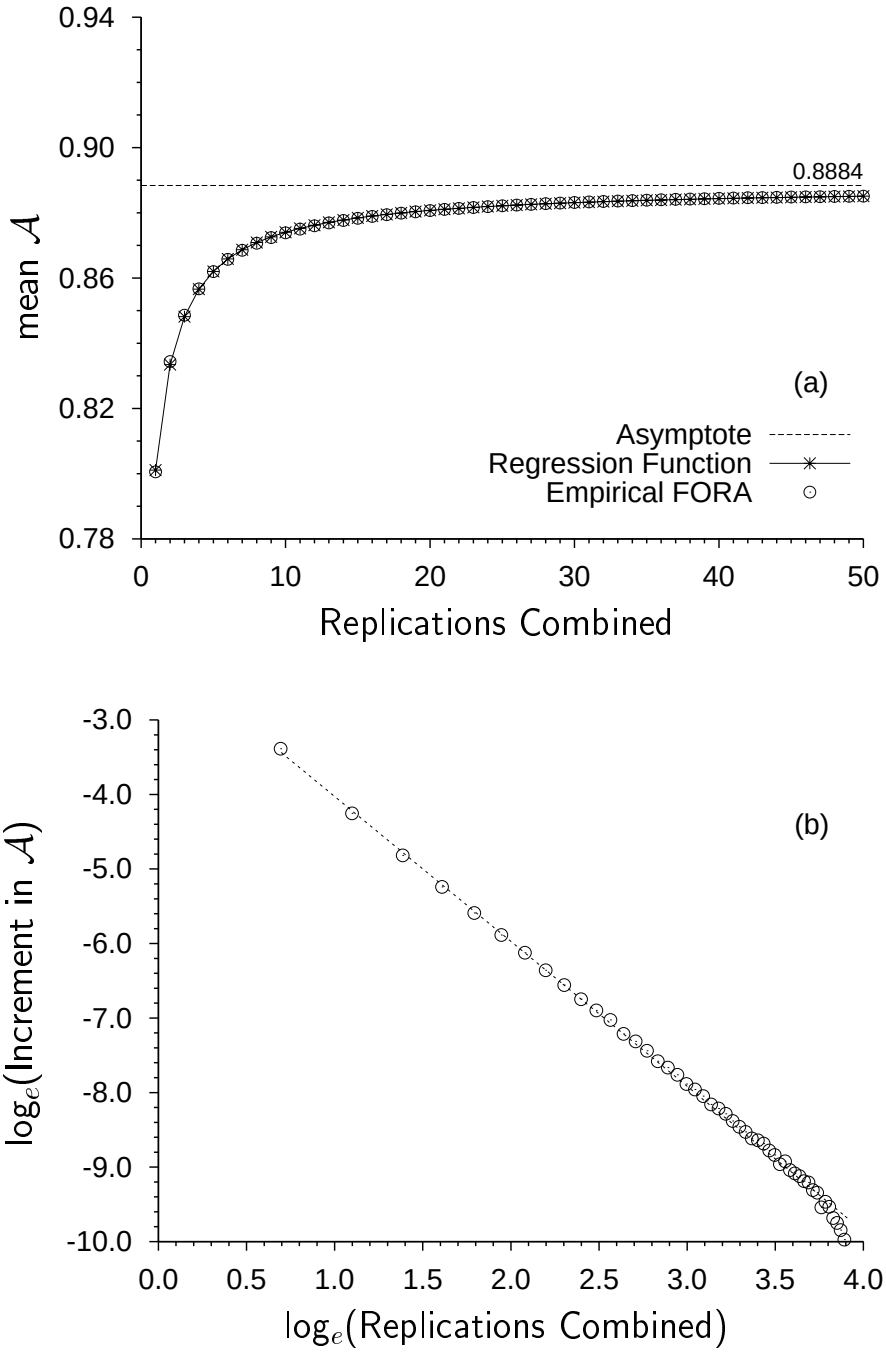


FIGURE 7.9: (a) Estimated 50-replication FORA based on 25 replications from each observer, showing mean area under the GOC curve as a function of replications combined. Values for combination-sizes 1 to 4 and 46 to 50 were calculated exactly, and values for combination-sizes 5 to 45 were estimated by random sampling. (b) The accompanying log-increment in area versus log of replications combined. The straight line indicates the log-log relationship based on parameters from the FORA regression function.

## 7.4 Sampling statistics of asymptotic values of $\mathcal{A}$

Group operating characteristic analysis reduces the effect of observer inconsistency, but cannot remove it entirely from any finite number of replications. FORAs tend asymptotically towards unique-noise-free performance, so there is residual error and variability associated with asymptotes from any finite data set. Two independent sets of replications from the same experiment are usually different, even if taken from the same observer. This section examines the variability of estimated asymptotes, rather than that of ROC or GOC performance, how to assess the variability, and how to estimate the residual error associated with any given number of replications.

Sampling statistics of the asymptotes were estimated by further analysis of data from Observer 2 only. (The data set for Observer 1 was not large enough for this purpose without conflicting with resampling problems.) Sets of a given size, $m$, were randomly sampled from the set of 75 replications, ACA and FORA regression was performed on each set, and the asymptotic value of $\mathcal{A}$ calculated. Having multiple ACA sets enabled the sample mean, standard deviation, skewness and kurtosis of asymptotes to be estimated for each ACA set size from 3 to 20 inclusive (asymptotes can only be estimated for 3 or more replications). The results show how these *statistics of asymptotes* change as a function of the number of replications used to estimate the asymptote.

### Sampling with and without replacement

Ideally, the sample statistics should be based on independent sets of replications, sampled without replacement, but this is impossible in practice, because it would require a huge number of trials. For example, assume that stable statistics may be estimated from a sample of 2000 asymptotic values. Each ACA set contributes only a single asymptote, so the statistics based on *independent* sets of size $m$ would require $2000 \times m$ replications. At 1000 trials per replication (for this experiment), this means 2 million trials per value of $m$, so estimates for $m = 3$ to 20, achieved below would require 414 million trials in total.[3] The only practical way of getting this quantity of data would be by computer simulation, and not experimentation. The 75 000-trial, 75-replication data set for Observer 2 seems small by comparison, yet it is the largest experimental set available.

Since independent ACA sets were not possible, sampling was done *with replacement* from the set of 75 replications. While the ACA sets were not independent, reasonably stable statistical estimates were possible at larger ACA set sizes (approximately $m > 8$). The effect of resampling with replacement, compared to sampling without replacement, is not known. Probable artifacts at smaller ACA set sizes are discussed in Section 7.4.3. The resampling procedure may be similar to statistical estimation of ROC parameters by jackknife techniques (Dorfman & Berbaum, 1986), but GOC analysis and ACA complicate matters to an unknown extent.

[3] Since $\sum_{m=3}^{20}(2 \times 10^6 \times m) = 414 \times 10^6$.

| ACA set size, $m$ | no. of resamplings | Sample statistics of asymptotic $\mathcal{A}$ | | | |
|---|---|---|---|---|---|
| | | mean | std. dev. | skewness | kurtosis |
| 3 | 16360 | 0.9113 | 0.0250 | 0.9261 | 4.9205 |
| 4 | 16360 | 0.9127 | 0.0187 | 0.5523 | 3.7616 |
| 5 | 16360 | 0.9143 | 0.0159 | 0.3595 | 3.3134 |
| 6 | 16360 | 0.9153 | 0.0141 | 0.3110 | 3.1126 |
| 7 | 16360 | 0.9161 | 0.0127 | 0.2733 | 3.0726 |
| 8 | 16360 | 0.9166 | 0.0115 | 0.2149 | 2.9804 |
| 9 | 16360 | 0.9169 | 0.0107 | 0.1762 | 3.0589 |
| 10 | 16360 | 0.9171 | 0.0100 | 0.1528 | 2.9934 |
| 11 | 4096 | 0.9173 | 0.0094 | 0.0949 | 2.9470 |
| 12 | 2048 | 0.9170 | 0.0089 | 0.1402 | 2.8139 |
| 13 | 2048 | 0.9176 | 0.0084 | 0.0981 | 2.8870 |
| 14 | 2048 | 0.9175 | 0.0080 | 0.1012 | 2.9418 |
| 15 | 2048 | 0.9177 | 0.0075 | 0.0441 | 3.0582 |
| 16 | 2048 | 0.9177 | 0.0072 | 0.0318 | 2.7880 |
| 17 | 2048 | 0.9173 | 0.0070 | 0.0074 | 2.7914 |
| 18 | 512 | 0.9177 | 0.0067 | 0.0440 | 2.9494 |
| 19 | 256 | 0.9174 | 0.0060 | -0.0077 | 2.7654 |
| 20 | 128 | 0.9167 | 0.0060 | -0.0831 | 2.9740 |

TABLE 7.1: Sample statistics of estimated asymptotic values of $\mathcal{A}$, for ACA set sizes from 3 to 20, based on ACA sets that were resampled from the 75 replications for Observer 2.

### 7.4.1 Results

The sampling statistics of estimated asymptotic values of $\mathcal{A}$, as a function of $m$, are given in Table 7.1. Compiler limitations meant that a maximum of $n_r = 16360$ (approximately $2^{14}$) resamplings could be run per ACA set size. This was done for ACA set sizes between 3 and 10 inclusive. Computation time became the limiting factor for $m > 10$. The number of resamplings was 4096 for $m = 11$, 2048 for $12 \leqslant m \leqslant 17$, and 512 (or fewer) for $18 \leqslant m \leqslant 20$. It is uncertain how large $n_r$ should be to achieve stable estimates. Preliminary investigations suggested, for this data set, that statistics were too variable when $n_r \leqslant 128$, but were generally stable when $n_r \geqslant 1024$. Also the smaller the ACA set size, the larger $n_r$ needed to be for the statistics to be stable.

Sample statistics as a function of ACA set size are plotted in Figures 7.10 and 7.11. The mean value of $\mathcal{A}$ is relatively constant, as can be seen in Figure 7.10(a), which plots $\mathcal{A}$ on a scale from 0.5 to 1.0. The mean asymptotic value of $\mathcal{A}$ changes only in the third decimal place (Table 7.1), so the more detailed scale given in Figure 7.10(b) is required

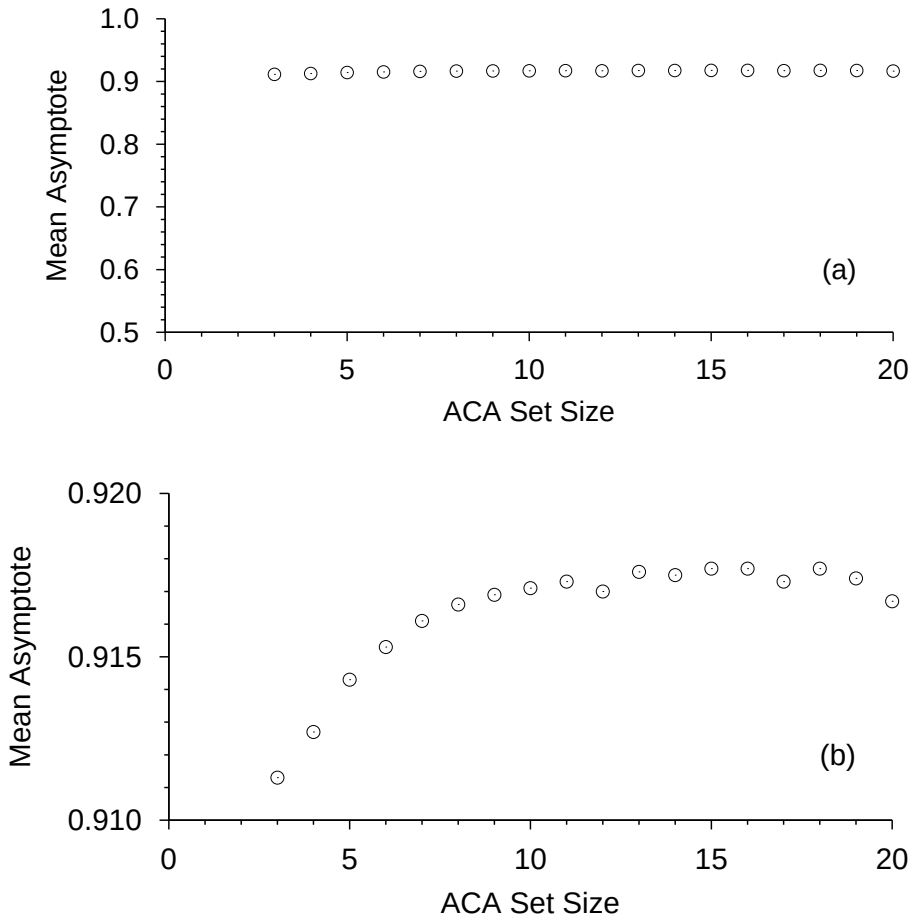


FIGURE 7.10: Mean estimated asymptotic value of $\mathcal{A}$ as a function of ACA set size, for subsets taken from the 75 replications for Observer 2. Both graphs show the same data, but the ordinate is scaled differently in each one. Panel (a) is scaled from 0.5 to 1.0, which is the usual range of $\mathcal{A}$, whereas panel (b) is scaled from 0.91 to 0.92 and shows small-scale details.

in order to see variation in the mean, which begins at 0.9113 for $m = 3$ replications and increases to around 0.9173 for $m \geqslant 10$. The mean seems stable for $m \geqslant 13$. (The decrease at $m = 20$ may have to do with the small number of resamplings (128) compared to the rest of the data.)

The initial increase in mean value from $m = 3$ to 10 suggests a bias in estimates based on small numbers of replications. The bias decreases as $m$ increases, suggesting that perhaps 9 or 10 replications are sufficient to obtain a reasonably unbiased estimate of the asymptote *for this data set*. This bias may be an artifact of the resampling procedure, and why it arises is described in Section 7.4.3.

It is difficult to put a value on any bias because theoretical performance is not known for this experiment. The data in Table 7.1 suggests the expected asymptotic $\mathcal{A}$ is 0.9176, although the estimate from the 75-replication FORA in Figure 7.8 was 0.9160. The latter

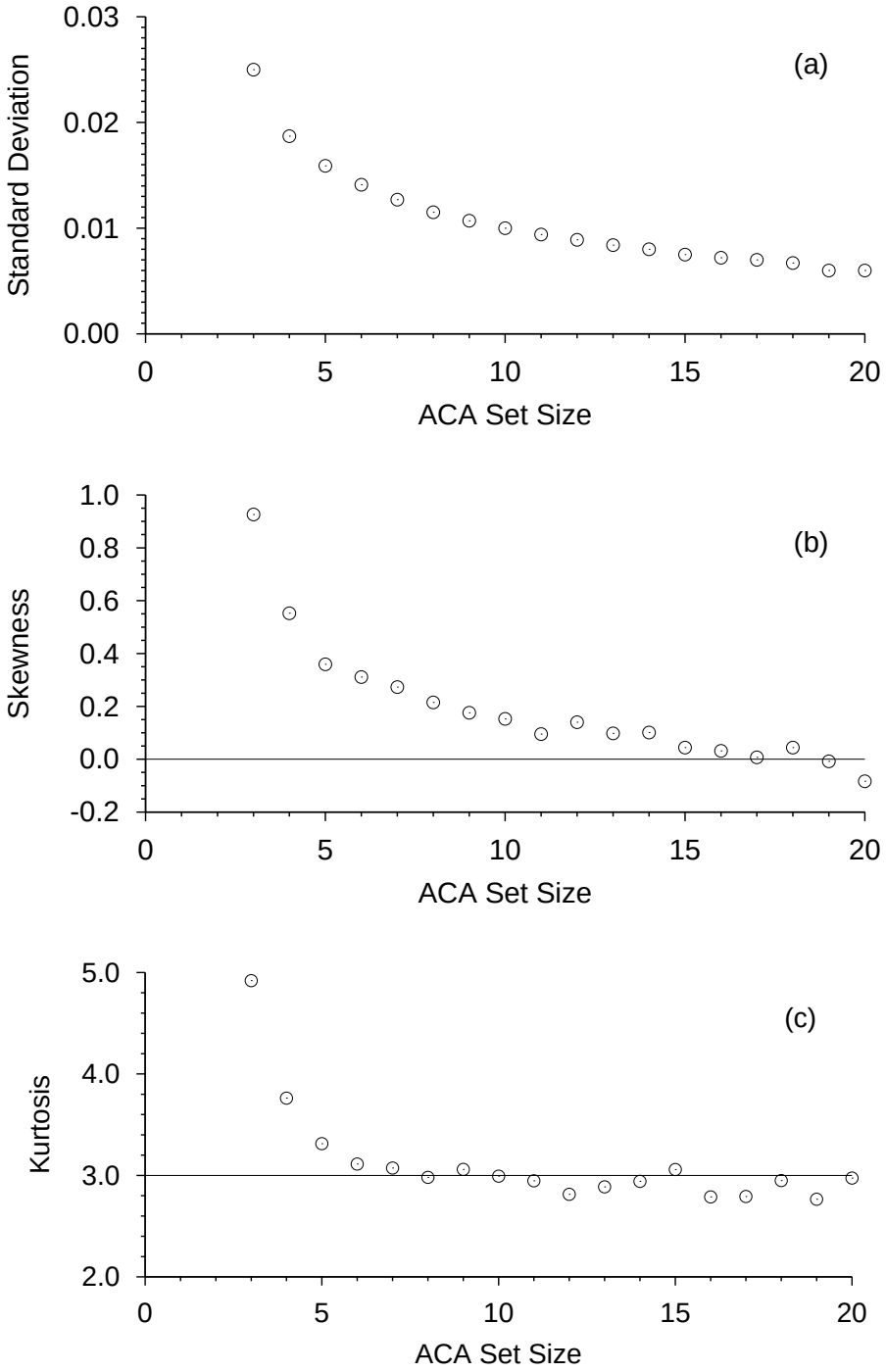


FIGURE 7.11: Sample statistics of estimated asymptotic values of $\mathcal{A}$ resulting from ACA of subsets of sizes 3 to 20 from the 75 replications for Observer 2. (a) standard deviation. (b) skewness—the horizontal line indicates the skewness of a symmetrical distribution. (c) kurtosis—the horizontal line indicates the kurtosis of a Gaussian distribution.

value is probably the best one to use because it is based on a much larger FORA (in Figure 7.8) than the FORAs that contributed to Table 7.1. Both estimates of the mean (0.9176 and 0.9160) fall well within one standard deviation of each other for each ACA set size given in Table 7.1.

Figure 7.11(a) shows that the standard deviation of the asymptotic $\mathcal{A}$ decreases steadily and smoothly as a function of $m$, although with diminishing returns as $m$ increases. The standard deviation is 0.01 at $m = 10$, and it would take more than twice as many replications in order to halve that. This suggests that there is still error in the second decimal place for the estimated mean $\mathcal{A}$, even after 20 replications. The function of standard deviation versus ACA set size in Figure 7.11(a) can be extrapolated, as is shown in Section 7.4.2.

Skewness is measured as the third moment around the mean divided by the cube of the standard deviation (McNemar, 1955). Skewness indicates the asymmetry of a distribution around its mean, and is zero for a symmetrical distribution. The skewness of estimated asymptotes, shown in Figure 7.11(b), is relatively large at the smaller ACA set sizes. It decreases towards zero, and remains near zero for $m \geqslant 11$. The initial positive skewness indicates that the distribution of asymptotes tails off to the *right* when $m$ is small.[4] The initial skewness may be another sampling artifact at small ACA set sizes.

Kurtosis is given by the fourth moment around the mean divided by the fourth power of the standard deviation (McNemar, 1955), which indicates the concentration of a distribution around its mean (roughly, how peaked a distribution is). The kurtosis of any distribution may be compared with that of a Gaussian distribution, whose kurtosis is equal to 3. Figure 7.11(c) shows that the distribution of asymptotes is peaked for small ACA set sizes—again, a potential artifact—but settles down to be in the vicinity of 3 for $m \geqslant 6$ which, together with the skewness statistics, suggests that the distribution of estimated asymptotes may well tend towards a Gaussian distribution for moderate to large ACA set sizes.

The data analysis leading to Table 7.1 and Figures 7.10 and 7.11 shows how sample statistics of asymptotic performance can be estimated, given a large but finite data set. Stable estimates were achievable for Observer 2 in this experiment only, using about 10 or more replications. There is probably no minimal number in general, because results and error will differ across observers and experiments. The minimum number of replications also depends on how much error is acceptable.

[4] Curiously, this suggests that the distribution of asymptotes is skewed in the *opposite direction* to that expected for a measure of sensitivity that is *bounded above* at 1.0, as suggested by McNicol (1972, Figure 5.6).

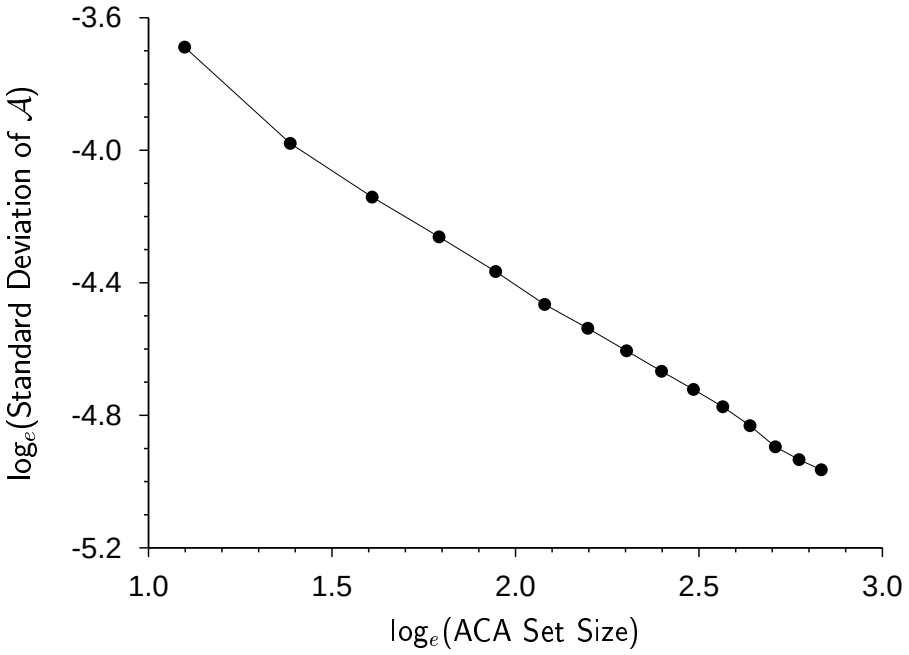


FIGURE 7.12: The standard deviation of estimated asymptotic values of $\mathcal{A}$ versus the ACA set size, presented on double-logarithmic co-ordinates, for ACA sets of size 3 to 17. Data points are joined by line segments only to highlight the linear relationship.

### 7.4.2 Estimating the standard deviation

The standard deviation of estimated asymptotes shown in Figure 7.11(a) shows a very smooth decrease as a function of ACA set size, $m$. Lapsley Miller (1998, personal communication) noted the similarity between Figure 7.11(a) and Figure 6.2(b), and suggested replotting the data in Figure 7.11(a) on double-logarithmic axes, as presented in Figure 7.12. Only ACA set sizes between 3 and 17 inclusive were used, because the number of resamplings was relatively small for $m \geqslant 18$ (Table 7.1).

The resulting function in Figure 7.12 is highly linear in double-logarithmic coordinates, suggesting an exponential regression to the data in Figure 7.11(a) (in linear coordinates). Such a regression function may be extrapolated, and provide an estimate of the standard deviation of asymptotic $\mathcal{A}$ for *any* ACA set size. A simple way of doing the regression is to fit a straight line to the data in double-logarithmic coordinates. The function in Figure 7.12 bends slightly at the left, and since the aim of regression was to extrapolate to the right, the first data point on the left was ignored. The remaining 14 points formed a highly linear series, with $r^2 = 0.9995$. If $s$ is the standard deviation, and $m$ is the ACA set size, then a linear least-squares regression of the remaining 14 points provided the regression equation

$$\log_e(s) \quad = \quad -0.6785 \ \log_e(m) - 3.0447,$$

which can be rearranged as

$$s = 0.0476\, m^{-0.6785} \tag{7.1}$$

where 0.0476 is equal to $\exp(-3.0447)$.

Estimated standard deviations based on Equation 7.1 agree with the empirical standard deviations in Table 7.1, with relative errors of 1% or less.[5] Table 7.2 lists estimated standard deviations for ACA set sizes from 5 up to 100 in steps of 5. Along with the estimated 75-replication asymptote of 0.9160, these results provide strong grounds for stating, with reasonable certainty, that unique-noise-free performance for Observer 2 in this experiment was at

$$\mathcal{A} = 0.916 \pm 0.005,$$

(within $\pm 2$ standard deviations). This is in contrast to average single-replication ROC performance at $0.792 \pm 0.028$. Not only does unique noise increase variability, but the decrease in performance can be viewed as a large bias, particularly if single-replication ROC performance is interpreted as a reflection of theoretical performance.

Table 7.2 shows that error could still affect the value of the asymptote in the second decimal place, even after 75 replications. The inverse of Equation 7.1,

$$m = 0.01125\, s^{-1.4738} \tag{7.2}$$

indicates the number of replications needed in order to achieve a given standard deviation. Table 7.3 lists the estimated number of replications for a set of given standard deviations. Equation 7.2 indicates that Observer 2 would need to run over a thousand replications in order for the asymptote to remain unchanged in the third decimal place, and tens of thousands of replications to achieve stability to the fourth decimal place. While this is possible in principal, it is certainly impractical.

### 7.4.3 Potential artifacts at small ACA set sizes

Potential artifacts arise when estimating sample statistics of asymptotes. The problem is that there is an unavoidable interdependence of results across (resampled) ACA sets, particularly at small ACA set sizes, $m$, which may affect sample statistics at small values of $m$. ACA sets were resampled *with replacement* from the set of 75 replications to approximate statistics of random samples taken *without replacement* from an unlimited number of replications. The number of resamplings, $n_r$, needed to obtain stable estimates

[5] Apart from when the ACA set size is 3, which Equation 7.1 is not intended to cover. Equation 7.1 should not be extrapolated to ACA set sizes of 1 or 2, because FORA regression requires at least 3 replications, and asymptotes cannot be estimated based on just 1 or 2 replications.

| ACA set size | estimated std. dev. | ACA set size | estimated std. dev. |
|---|---|---|---|
| 5 | 0.01598 | 55 | 0.00314 |
| 10 | 0.00998 | 60 | 0.00296 |
| 15 | 0.00758 | 65 | 0.00280 |
| 20 | 0.00624 | 70 | 0.00267 |
| 25 | 0.00536 | 75 | 0.00254 |
| 30 | 0.00474 | 80 | 0.00243 |
| 35 | 0.00427 | 85 | 0.00234 |
| 40 | 0.00390 | 90 | 0.00225 |
| 45 | 0.00360 | 95 | 0.00217 |
| 50 | 0.00335 | 100 | 0.00209 |

TABLE 7.2: Estimated standard deviation of the asymptotic value of $\mathcal{A}$ for Observer 2, for ACA set sizes from 5 up to 100 in steps of 5. Estimates are based on regression using Equation 7.1.

| standard deviation | estimated number of replications |
|---|---|
| 0.1 | 0.33 |
| 0.05 | 0.93 |
| 0.02 | 3.59 |
| 0.01 | 9.97 |
| 0.005 | 27.7 |
| 0.002 | 106.9 |
| 0.001 | 296.9 |
| 0.0005 | 824.7 |
| 0.0002 | 3182 |
| 0.0001 | 8839 |

TABLE 7.3: Estimated number of replications needed to achieve a given standard deviation of the asymptotic value of $\mathcal{A}$. Estimates are based on Equation 7.2. The estimated number of replications is a non-integer lower bound (given to as few decimal places as appropriate). These should be converted to the next largest integer to be realistic.

is not known, but was set to $n_r = 16360$ for $3 \leqslant m \leqslant 10$, and $n_r = 2048$ was used as a practical lower limit for $m > 10$. The total number of different ACA sets of size $m$ that are possible from a data set of 75 replications is ${}^{75}C_m$, and the ratio of this number to the number of resamplings is ${}^{75}C_m/n_r$. This ratio reflects the scope for variation in resampling ACA sets and asymptotes. Unfortunately, ${}^{75}C_m/n_r$ is relatively small at small ACA set sizes. The ratio is 4.1 for $m = 3$, and 74.3 for $m = 4$, but the ratio is some tens of thousands for $4 \leqslant m \leqslant 7$, and millions for $m \geqslant 8$. Resampling interdependence at small values of $m$ must affect the estimated sample statistics of asymptotes presented in Figures 7.10 and 7.11. Sample statistics at small values of $m$ probably reflect the particular sample of 75 replications more than at moderate to large values of $m$, because the scope for variability is reduced.

Once $m > 6$ (for a data set of 75 replications), interdependence is no longer a concern because ${}^{75}C_m/n_r$ becomes very large. Curiously, kurtosis seems to settle down to a value around 3 at just this point in Figure 7.11(c), and the mean asymptote tends to a stable level once $m$ is beyond 6. No clear cutoff for $m$ presents itself for either skewness or standard deviation, and Figure 7.12 hardly suggests any effect at all for the standard deviation (except perhaps at $m = 3$). If $m = 6$ is a reasonable cutoff, this suggests that resampling artifacts are minimal once ${}^{75}C_m/n_r$ is on the order of $10^5$ or more.

Sampling statistics of asymptotes were not calculated for Observer 1 because interdependence is much worse for a data set of only 25 replications. The ratio ${}^{25}C_m/n_r$ is never more than 318 (using $n_r = 16360$) for any value of $m$, and is generally less than 70.

In summary, it is likely that aspects of the sample statistics of asymptotes at small ACA set size are *an artifact of the resampling procedure* that was used, but the artifact is minimal once the ACA set size is large enough, and once the data set is large enough.

## 7.5 Summary

FORA regression of a large data set makes it possible to not only estimate asymptotic performance, but also to estimate sampling statistics and error bounds of the asymptote. Several new analyses and a new data set were introduced in this chapter. FORA regression provided excellent descriptions of the data, even out to 75 replications, and unique-noise-free performance values were obtained with unprecedented accuracy.

The data set from the amplitude discrimination experiment provided evidence that the data pattern seen in Chapter 6 was not a coincidence. Log-log plots for the current experiment were very linear, both within observers and across observers, and FORA regression could hardly have been better than it was. The regression-FORA in Figure 7.8, for example, was virtually identical to the data series, and yet the regression was based on the values of only three empirical parameters.

FORA results illustrated the large amount of improvement that was possible through

the use of GOC analysis. The mean value of $\mathcal{A}$ was initially around $\mathcal{A} = 0.79$ for Observer 2, but unique-noise-free performance was over $\mathcal{A} = 0.91$. There were definite individual differences between observers for the same stimulus set, both in the level of common noise implied by each observer's asymptote, and in the decrease in performance due to unique noise, which was much larger for Observer 2. Although Observer 1 was better than Observer 2, based on mean ROC performance, the roles were reversed when two or more replications were combined. Relative ROC performance did not, in this case, reflect relative GOC performance or relative asymptotic performance.

Combinatorial explosion puts practical limits on ACA of large data sets. Two variations on ACA were introduced, partial-ACA and sample-ACA. Partial-ACA involved the calculation of only a small number of data points near the ends of a FORA, whereas sample-ACA involved random sampling of a finite number of combinations to arrive at estimated mean performance for combination-sizes between the end points. For Observer 2 in this experiment, both variations on ACA arrived at the same estimated asymptote.

The large data set for Observer 2 made it possible to estimate sampling statistics of the asymptote. Repeated ACA over subsets of replications made it possible to build up a picture of the error associated with asymptotes. The minimum number of replications needed for FORA regression is three. The estimated asymptote based sets of only three replications, for example, was very close to estimates based on much larger sets. There was certainly an improvement over single-replication performance. Generally, the variability of asymptotic estimates was very large when only a small number of replications were used in ACA, but variability decreased regularly with increasing ACA set size. Possible artifacts in the estimation process were described. These were believed to have little impact, for this data set, once the ACA set size reached double figures.

An avenue of investigation not pursued here is that of simulation. Both statistical simulation, using Monte Carlo methods, and computational simulation of unique-noise-affected ideal observers, could offer ways around unavoidable practical limitations encountered in ACA. Theoretical FORAs, noted in Chapter 6, were not pursued either, and still need to be integrated with the current framework.

**In conclusion.** Given enough replications, it is possible to achieve very accurate descriptions of unique-noise-free psychophysical data. Reasonable estimates can be achieved based on only a small number of replications, but at the expense of greater potential error. The large number of replications from Observer 2 provided a way of evaluating how many replications may be required in an experiment to achieve given error bounds on estimated asymptotes. With regard to the asymptotic value of $\mathcal{A}$, for example, the results for Observer 2 suggested that 10 replications were enough to obtain a standard deviation of 0.01, but almost 300 replications would be needed to achieve a standard deviation of 0.001.

# Chapter 8

# Functions of replications added for various experiments

The previous two chapters introduced ACA and FORA analyses as extensions of GOC analysis. FORA regression was developed to estimate asymptotic, unique-noise-free performance. The previous chapter showed how the sampling statistics of asymptotes may be estimated. This chapter presents a collection of multiple-replication experimental results, which show how ubiquitous the FORA data pattern is. Together with the experiments in preceding chapters, these results show that ACA and FORA regression works over a wide variety of experimental paradigms (SIFC and 2IFC), experimental tasks (frequency discrimination and amplitude discrimination), decision methodologies (continuous rating scale and binary-decision tasks), individual observers (seven different people in all), stimulus parameters (bandwidth, duration, centre frequency, and signal-to-noise ratio), and measures of sensitivity. Each experiment investigates some, but not all, of these factors. Results show that the general FORA regression pattern is *extremely* robust across experiments and measures of sensitivity, with linear or near-linear log-log plots being the norm.[1]

These data sets did not come from one particular experimental project. Rather, they came from different projects used to investigate different topics at different times, and were obtained through the courtesy of the authors. The experiments took place over a period of nine years and are presented mostly in the order of data collection.

Section 8.1 presents results from two 2IFC frequency discrimination experiments from Lapsley Miller et al. (1998), which used known, discrete evidence distributions. The first experiment used a binary-decision methodology whereas the second experiment used a continuous rating scale methodology. FORAs based on 64 replications from each experiment are compared and contrasted to show the effect of decision methodology on estimated asymptotes.

[1] Similar findings were also found in other experiments run by Galvin et al. (1998), and by Lapsley Miller et al. (1998), which are not presented here.

Section 8.2 presents results from a 2IFC amplitude discrimination experiment, also from Lapsley Miller et al. (1998). The decision axis was not known for this experiment but was assumed to be continuous because of the nature of the stimuli. FORAs were calculated separately for each of four observers. The FORAs demonstrate a broad range of single-replication performance across observers, and also a broad range of asymptotic performance. Like Whitmore et al.'s (1993) results in Section 7.3, the ranking of observers according to asymptotic performance is not necessarily the same as the ranking based on single-replication performance.

Section 8.3 presents results from a previously unpublished 2IFC amplitude discrimination experiment which involved multiple signal-to-noise ratios. FORA regression is found to hold over a wide range of performance levels, from close to chance to near-perfect performance. Minimal ceiling effects are illustrated in FORA regression at the highest signal-to-noise ratio. Whereas previous experiments were mainly based on a large number of replications, this experiment shows that FORA regression is also practical when only eight replications are involved.

Finally, Section 8.4 summarises numerous FORA results reported by Lapsley Miller (1999). A series of multiple-signal-to-noise-ratio, SIFC, amplitude discrimination experiments were run to investigate the role of bandwidth and duration in the detection of Gaussian noise signals. These experiments covered 18 pairs of bandwidth and duration, and the whole experimental project was roughly equivalent to running 18 experiments like the one in Section 8.3, with 18 times as many FORAs.

Out of hundreds of empirical FORAs presented in Lapsley Miller (1999), (across conditions, observers, signal-to-noise ratios and measures of sensitivity), the vast majority were fitted by regression-FORAs that converged sensibly, having log-log plots with $r^2$ values greater than 0.99 (typically 0.999 or more). A handful of FORAs did not converge sensibly, or at all, and these are discussed in Section 8.4. As well as analysing two-event FORAs, Lapsley Miller analysed each data set using a six-event discrimination measure, $\mathcal{D}_6$. FORA regression based on $\mathcal{D}_6$ was extremely successful, with $r^2$ typically between 0.9995 and 1.0000.

## 8.1 FORAs based on binary-decision and continuous rating scale data

Two 2IFC multiple-replication, aural frequency discrimination experiments were run by Lapsley Miller et al. (1998), with the same four observers taking part in each experiment. A tonal transient was presented in each of the two observation intervals on each trial. One transient came from a set of higher-frequency[2] tones while the other came from a set of

[2] "Frequency" refers to tonal frequency in Hertz, and not frequency of occurrence.

lower-frequency tones. The task of an observer was to decide whether or not the higher-frequency tone had occurred in the first interval. The two experiments were essentially the same except for the way in way in which decisions were indicated—the first experiment involved conventional 2IFC binary-decision methodology (using two push-buttons), while the second experiment involved a continuous rating scale (using an electronic slider).

These were part of a series of experiments that Lapsley Miller et al. (1998) used to evaluate relationships between the SIFC and 2IFC tasks. The experiments were unusual in that the distributions of aural frequency were completely specified by the experimenters beforehand. This was so the theoretical ROC curve, which was the same for both experiments, was known *a priori* (assuming the decision rule given below). Like Taylor et al.'s (1991) SIFC experiment described in Chapter 2, these 2IFC experiments were aural equivalents of the dice game.

One possible decision rule in a 2IFC task is to use a *differencing strategy* (Tanner & Birdsall, 1958; Robinson & Watson, 1970; McNicol, 1972; Simpson & Fitter, 1973; Egan, 1975; Lapsley Miller et al., 1998). Observers could calculate evidence values for each observation interval separately (as they would for an SIFC task) and use the difference between the two values as the basis for 2IFC decisions. If the second evidence value is subtracted from the first, then the difference in frequency becomes a new decision axis for the 2IFC task. A systematic manipulation of criterion-based decision rule applied to this axis yields a 2IFC ROC curve.

The label *SN* refers to *higher* frequencies, and *N* refers to *lower* frequencies. The 2IFC events are labelled as $\langle \mathrm{SN}, \mathrm{N} \rangle$ and $\langle \mathrm{N}, \mathrm{SN} \rangle$, reflecting the possible orders of presentation of *SN* and *N* transients in each trial. For the purposes of 2IFC ROC analysis, the 2IFC hit rate is conditional on the $\langle \mathrm{SN}, \mathrm{N} \rangle$ event, while the 2IFC false alarm rate is conditional on the $\langle \mathrm{N}, \mathrm{SN} \rangle$ event.

### 8.1.1 Method

**Observers.** Four human observers took part in both experiments. Observers 1 and 4 were naive observers, whereas Observers 2 and 3 were experienced observers. Observers 2 and 3 were the same two individuals, respectively, as Observers 1 and 2 in Whitmore et al.'s (1993) experiment (Chapter 7). Both were also observers in Taylor et al.'s (1991) continuous rating scale, frequency discrimination experiment.

**Stimuli.** The *N* tones took on frequencies in 5 Hz steps from 595 Hz to 640 Hz inclusive. The *SN* tones took on frequencies in 5 Hz steps from 620 Hz to 665 Hz inclusive. There were 15 different frequencies in total, 10 each for the *SN* and *N* events with an overlap of 5 frequencies. The distribution of frequencies followed discrete, overlapping uniform distributions when considered for any single observation interval. Assuming 2IFC differencing strategy, this meant the theoretical 2IFC distributions followed the overlapping,

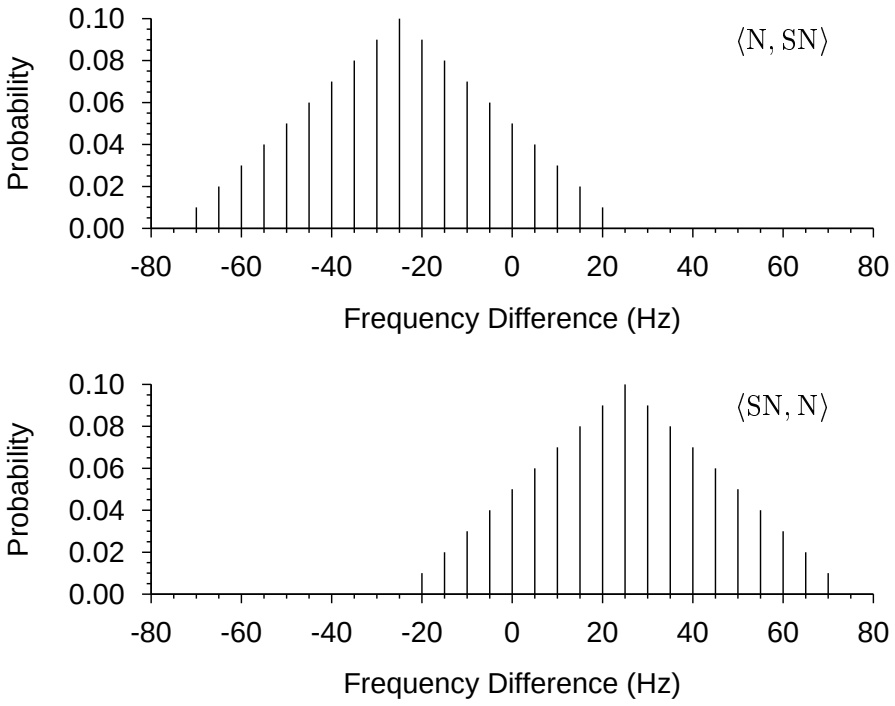


FIGURE 8.1: Probability mass functions for Lapsley Miller et al.'s (1998) 2IFC frequency discrimination experiment, assuming a differencing decision strategy (after Lapsley Miller et al., 1998, Figure 4).

discrete triangular distributions shown in Figure 8.1.[3] If $X_{\mathrm{SN}}$ and $X_{\mathrm{N}}$ denote the *SN* and *N* random variables within a single observation interval, then the random variable for the $\langle \mathrm{SN}, \mathrm{N} \rangle$ event is $X_{\mathrm{SN}} - X_{\mathrm{N}}$ while that for the $\langle \mathrm{N}, \mathrm{SN} \rangle$ event is $X_{\mathrm{N}} - X_{\mathrm{SN}}$. Hence the two distributions are mirror images of each other relative to a difference of 0, and the theoretical 2IFC ROC curve is symmetrical about the negative diagonal in the ROC space (Green & Swets, 1974).

The equipment used in these experiments was the same as described in Chapter 7 for Whitmore et al.'s (1993) experiment, in particular, the sound chamber, rating sliders, headphone and headset models, headset amplifier, 12-bit digital-to-analog converter (DAC), multiprogrammer and the controlling computers were the same.

Digital code sequences for the sinusoidal transients were generated on an HP 9826 computer using its internal sine function. Each code sequence, consisting of 720 points, was output to the DAC, which was clocked at 7.2 kHz. The output of the DAC was smoothed using a passive 1.25 kHz low-pass filter. The transients had an absolute duration of 100 ms and an equivalent rectangular duration of 81.2 ms. A Kaiser window with shape parameter of 11 was used to gate transients on from zero to maximum power over the first 15 ms and gate them off over the last 15 ms.

[3] The difference between two uniform distributions of equal width is a triangular distribution.

During experimental sessions, a 4 kHz low-pass analog Gaussian noise masker ran continuously at a spectrum level 40 dB SPL. Transients were presented diotically at a signal-to-masker level of 13.2 dB for Observers 1 and 2, and at 15.0 dB for Observers 3 and 4.[4] The signal-to-masker level only affected the level of unique noise, not the common noise or theoretical performance. Common noise was determined by the distributions of tonal frequencies, which were independent of the masker.

**Procedure.** Each observer ran 16 replications of each experiment, so there were 64 replications in total. Observers ran practice sessions until they demonstrated proficiency at the task.

Each replication consisted of 400 trials, with 200 trials per 2IFC event. Since there were 10 *SN* frequencies and 10 *N* frequencies, there were $10 \times 10 = 100$ possible pairings of frequencies across the *SN* and *N* sets. Each possible pairing was presented twice per 2IFC event in each replication. Pairings were presented using a different haphazard sequence across trials for each observer and each replication. The sequence of 2IFC events was run-limited so the same event could not occur more than a certain number of times in a row. This was to help avoid trial counting as a possible means of improving performance. Each replication had one run-limit value chosen at random from the values 4, 5 or 6. Observers did not know what the limit was on any replication.

Each trial consisted of a 50 ms warning interval, two 100 ms observation intervals separated by a 500 ms inter-stimulus interval, a 1000 ms decision interval, and a 750 ms reset interval. In the continuous rating scale experiment, the reset interval was a minimum duration. The next trial could not begin until the slider had been reset to its extreme left so the slider always started in the same position on each trial. A set of LED lights on the decision panel were switched on and off to mark the trial intervals. No trial-by-trial knowledge of results was given, but observers could later view their single-replication ROC curves at the conclusion of each replication.

The continuous rating scale was used to indicate confidence that the $\langle \mathrm{SN}, \mathrm{N} \rangle$ event order had occurred—the extreme left indicated zero confidence while the extreme right indicated 100% confidence, with increasing confidence indicated by an increasing slider position going from left to right. The 12 cm long continuous rating scale was partitioned evenly into 64 categories, and ratings were stored as integers from 1 to 64.

In the binary-decision experiment, decisions were made by pushing one of two 2.5 cm buttons which were on the same panel as the slider and LED lights. If neither button was pushed by the end of the reset interval, an incorrect decision was recorded, which would have added slightly to the unique noise.[5]

[4] Observers 1 and 2 had found the task noticably easier than Observers 3 and 4 in preliminary experimental sessions. The levels were different in order to even out single-replication performance across all observers, which was desirable for the purposes of Lapsley Miller et al.'s (1998) study.

[5] This happened only a handful of times for Observers 1 to 3, but almost 1% of the time for Observer 4.

### 8.1.2 Data analysis

For each experiment, the 16 replications from each observer were aggregated to form a data set of 64 replications. Since the same theoretical distributions applied to all observers, no distinction is made here across observers. ROC and GOC analyses were performed on both experimental data sets, and also on a third set based on a repartition of the continuous rating scale data into two categories. This was useful for comparing results based on the 64-point rating scale data and on binary-decision data. For convenience, the results are discussed as if there were three separate data sets.

At the end of each experimental trial involving the continuous rating scale, an electronic measurement of slider position was converted into one of 64 rating categories. To generate the derived-binary data, the 64-point scale was partitioned at its midpoint, which was equivalent to bisecting the original continuous scale. Ratings of 32 or less were equated with a *"no"* decision while ratings of 33 or more were equated with a *"yes"* decision. The resulting data is called *derived binary-decision* data. Each of the continuous rating scale and *push-button binary-decision* experiments involved a different set of replications, where each separate replication is associated with new unique noise samples. The reason for the third data set is that discrepancies in results between the 64-point rating scale data and the push-button binary-decision data could be due simply to unique noise sampling variablity, rather than to an underlying pattern of difference due to the decision methodology. A more direct comparison that minimises such sampling variability is between the 64-point rating scale data and the derived binary-decision data, both of which share the same unique noise samples. Any discrepancies in results could be due only to collapsing the rating scale from 64 points down to 2, but not to independent and unknown samples of unique noise.

**ROC and GOC Results.** The single-replication ROC curves from each of the three data sets, across all observers, are shown in the left-hand panels of Figure 8.2, along with the theoretical 2IFC ROC curve. There was obvious variability both across observers and within observers. It is also clear that none of the ROC curves followed the theoretical ROC curve and that ROC performance was well down from theoretical performance. Figures 8.2(a) and 8.2(c) graphically show that not only did unique noise depress the level of binary-decision performance, but that it also affected an observer's apparent bias in the task.

The 64-replication mean ROC and GOC curves for Figures 8.2(a), 8.2(c) and 8.2(e) are given in Figures 8.2(b), 8.2(d) and 8.2(f) respectively. The mean ROC curves showed relatively poor performance compared to theory. In contrast, the GOC curves were much closer to the theoretical ROC curve and showed much better performance in the 2IFC task compared to the mean ROC curves.

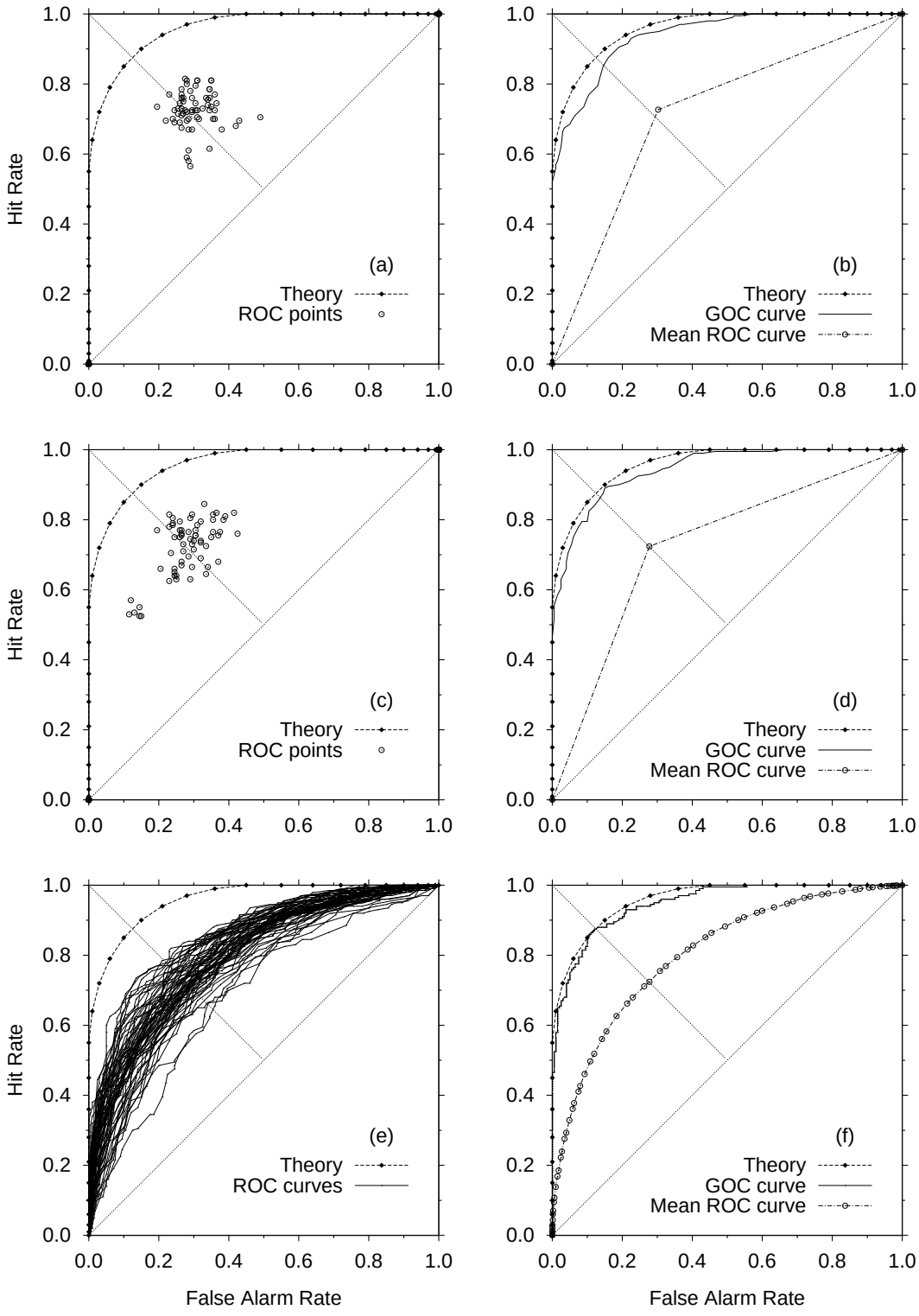


FIGURE 8.2: Single-replication ROC curves (left-hand panels), and mean ROC and GOC curves (right-hand panels) for all 64 replications for all observers in the discrete 2IFC experiments. Panels (a) and (b) are for the push-button binary-decision data, (c) and (d) are for the derived binary-decision data and (e) and (f) are for the 64-point rating scale data (after Lapsley Miller et al., 1998, Figures 5, 7 and 8).

Unbiased 2IFC binary-decisions should lie on the negative diagonal, but the single-replication ROC curves (points) in Figures 8.2(a) and 8.2(c) generally did not. The mean ROC curves in Figures 8.2(b) and 8.2(d), did lie on or very close to the negative diagonal. This suggests that observers were unbiased on average, but that unique noise sampling variability across replications resulted in apparent bias on individual replications.

The mean ROC points for the push-button binary-decision data and derived binary-decision data in Figures 8.2(b) and (d) were very similar to each other. This suggests that the overall level of unique noise was the same for the continuous rating scale and the push-button experiment. The 64-point rating scale GOC curve was the closest to the theoretical ROC curve, and the push-button binary-decision GOC curve was the furthest away. A comparison between the GOC curves in Figures 8.2(d) and (f) shows that after 64-replications were combined, the GOC curve based on derived binary-decisions was very slightly lower than the GOC curve based on 64-point rating scale decisions. Since GOC curves tend towards their final form asymptotically (as replications are added), small differences between GOC curves that are close to asymptotic performance reflects a relatively large difference in the number of replications needed to reach a given level. This can be shown using FORAs.

### 8.1.3 FORA results

**FORAs based on $\mathcal{A}$.** The number of replications of each experiment was too large to compute ACA for all combination-sizes. Instead, partial-ACA was computed for combination-sizes 1 to 5, and for the complementary sizes from 59 to 64. This resulted in FORAs that were defined only at their outer points.[6] FORAs based on $\mathcal{A}$ are presented in Figure 8.3 along with their respective log-log plots. The FORAs for the two binary-decision data sets were similar, although the derived binary-decision FORA was consistently better than the push-button FORA, by roughly the same amount at each combination-size. The 64-point rating scale FORA was much better than either of the two binary-decision FORAs. Despite the differences in FORA location across the three data sets, the three estimated asymptotes were remarkably similar to each other (horizontal lines in Figure 8.3(a)), and each compared well to theoretical performance.

Data points on the log-log plot for the two binary-decision FORAs (Figure 8.3(b)) were virtually identical to each other. The log-log plot for the 64-point rating scale FORA was consistently lower than the two binary-decision plots, which reflects smaller increments in the 64-point rating scale FORA. The outer points on each of the three log-log plots indicated that a full 63-point data series would curve downwards, implying that each asymptote underestimated theoretical performance.

[6] This was described in Section 7.3 in the analysis of Whitmore et al.'s (1993) experiment.

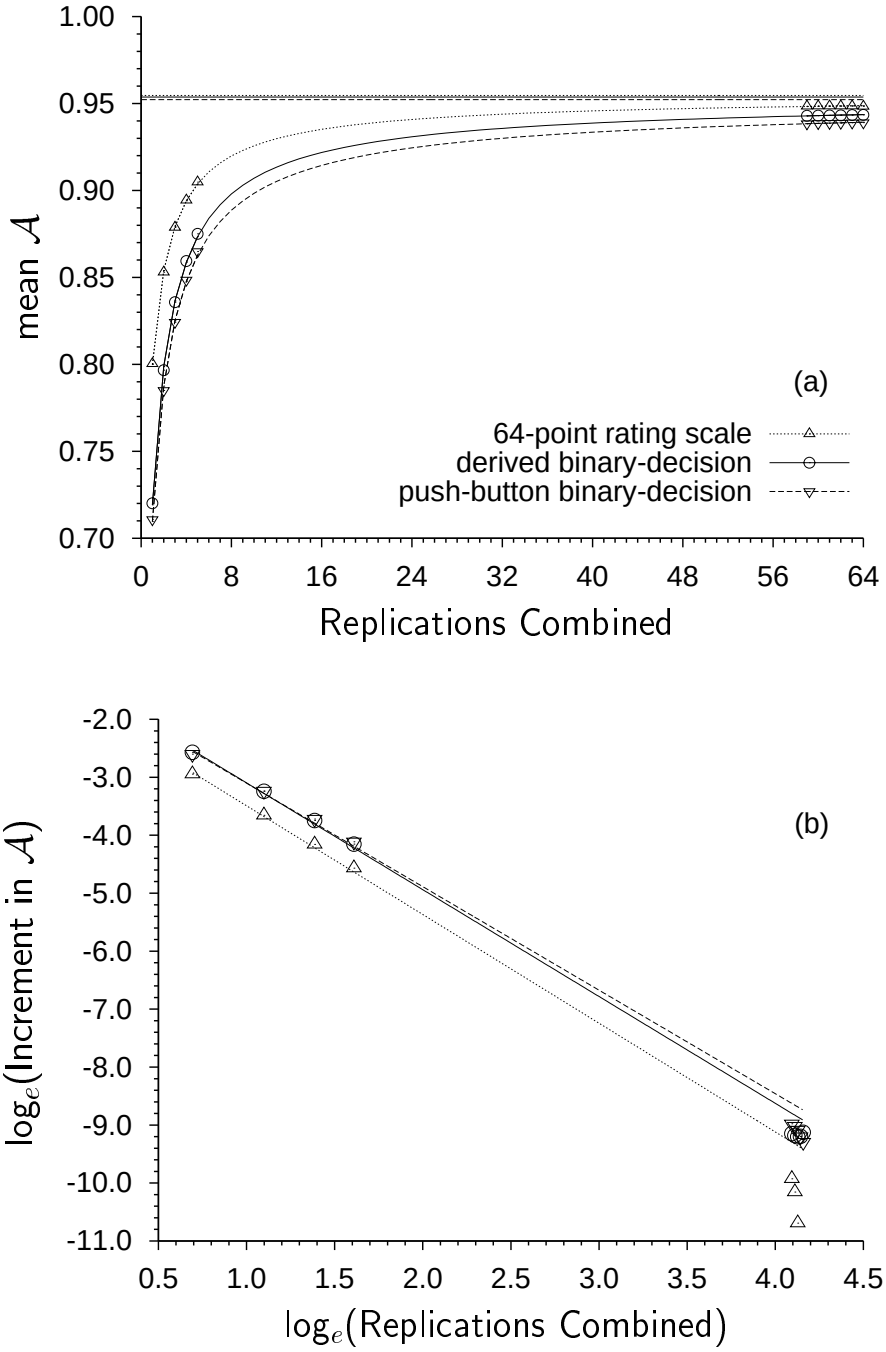


FIGURE 8.3: (a) Outer points on the 64-replication FORAs based on $\mathcal{A}$ for Lapsley Miller et al.'s (1998) discrete case 2IFC experiments. Open symbols denote data points, horizontal lines denote asymptotes, and curved lines denote regression functions. These are keyed to the data sets as follows: $\bigtriangledown$ and - - - - for the push-button binary-decision, $\bigcirc$ and —— for the derived binary-decision and $\bigtriangleup$ and ····· for the 64-point rating scale. (b) The accompanying log-log plot for each of the three data sets. Straight lines indicate log-log relationships based on parameters from the FORA regression function for each data set. Symbols and line-types denote the same data sets as for panel (a).

Both of the binary-decision FORAs based on $\mathcal{A}$ started near 0.72 and increased to around 0.94 after 64 replications. In contrast, the 64-point rating scale FORA started at $\mathcal{A} = 0.80$ and increased to almost 0.95. The final values on each FORA (after 64-replications) were very similar, because the GOC curves shown in Figure 8.2 were all very similar. The asymptotic values of $\mathcal{A}$ for the three data sets were also similar, at 0.9522, 0.9537 and 0.9546 for the push-button binary-decision, derived binary-decision, and 64-point rating scale FORAs respectively. All three values compared very well with the theoretical $\mathcal{A}$ of 0.9588, but all three values underestimated theoretical performance slightly. This was consistent with the log-log plots, each of which implied that they curved downwards slightly.

**FORAs based on $d'$, $\mathcal{D}_2$ and $P(C)$.** FORAs for $d'$, $\mathcal{D}_2$ and $P(C)$ were also calculated, but are not plotted here. The FORAs and log-log plots for $d'$ and $\mathcal{D}_2$ were smooth and very similar in form to those in Figure 8.3 whereas those for $P(C)$ were much more variable. FORA results for each of the three data sets, and for all four measures of sensitivity, are listed in Table G.3 in Appendix G.

The relationship between $\mathcal{A}$ and $d'$, and between $\mathcal{A}$ and $\mathcal{D}_2$, accentuated differences among the data sets. For $d'$, the estimated asymptotes were 2.4817, 2.4934 and 2.4619 for the push-button binary-decision, derived binary-decision, and 64-point rating scale FORAs, respectively, compared to the theoretical value of $d'$, which was 2.4556. For $\mathcal{D}_2$, the estimated asymptotes were 0.7917, 0.7824 and 0.7527 bits, respectively, compared to the theoretical value of $\mathcal{D}_2$, which was 0.7520 bits. Curiously, the 64-point rating scale data provided the largest asymptotic $\mathcal{A}$ but the smallest asymptotic $d'$ and $\mathcal{D}_2$. FORAs from all three data sets *underestimated* the theoretical value of $\mathcal{A}$, and *overestimated* the theoretical values of $d'$ and $\mathcal{D}_2$. Viewed separately, all three data sets provided good approximations to the theoretical performance value for all three measures, $\mathcal{A}$, $d'$ and $\mathcal{D}_2$. The estimated asymptote from the 64-point rating scale data, however, was consistently the best (i.e. closest to theory) for all three measures.

The FORAs for $P(C)$ for all three data sets were not smooth functions, and were much more variable than those based on $\mathcal{A}$, $d'$ and $\mathcal{D}_2$. The FORA based on $P(C)$ for the derived binary-decision data was so variable that the regression procedure could not converge. The value of $P(C)$ based on the 64-replication GOC curve for each data set was much better than the average single-replication value of $P(C)$, because the GOC curves were much better than the mean ROCs (right-hand panels in Figure 8.2). Unlike the other three measures, the initial values on the $P(C)$ FORAs were very similar for all three data sets. This is because $P(C)$ was calculated from where an ROC or GOC curve intersected the negative diagonal. The value of $P(C)$ for the 64-replication 64-point rating scale GOC curve was exactly equal to the theoretical $P(C)$ value of 0.8750, because the 64-replication GOC curve happened to intersect the negative diagonal at exactly the same point that the theoretical ROC curve did (Figure 8.2(f)).

**Effect of using the trapezoidal rule to calculate $\mathcal{A}$.** Much of the discrepancy between the continuous rating scale FORA and the binary-decision FORAs related to the number of points on a GOC curve and the way in which $\mathcal{A}$ was calculated. Empirical values of $\mathcal{A}$ were calculated using the trapezoidal rule (McNicol, 1972; Bamber, 1975). This rule can grossly underestimate theoretical values of $\mathcal{A}$ whenever a rating scale ROC curve provides a poorly defined approximation to a theoretical ROC curve (Bamber, 1975, Figure 6). This can occur when the number of ROC (or GOC) points is small. In general, there are $\xi$ points on a $\xi$-replication GOC curve based on binary-decision data (excluding the points (0,0) and (1,1)). Underestimation of $\mathcal{A}$ in the binary FORAs in Figure 8.3(a) was large when $\xi$ was small, but not when $\xi$ was moderate to large, ($\xi > 30$, say).

The regression-FORAs in Figure 8.3(a) suggested that the average performance based on binary-decision data was always worse than that based on continuous rating scale data within the range of replications ($1 \leqslant \xi \leqslant 64$). A correction for underestimation due to the trapezoidal rule could be applied, similar to the use of $A'$ for single-point ROC curves (Pollack & Norman, 1964; Smith, 1995).[7] Another possible correction could be to fit an ROC curve to each GOC curve calculated during ACA (Taylor, 1984), and use the area under the fitted curve as the sensitivity value for each GOC curve. This process, described in Section 6.2.4, was not implemented here because of its extra complexity and uncertain effects. Rather than correcting for underestimation due to the trapezoidal rule, it is simpler to use a multiple-point rating scale instead of a two-point rating scale.

### Comparison between rating scale and binary-decision data

The GOC curves, FORAs and subsequent asymptotes for all measures support the use of a multiple-point rating scale instead of a two-point rating scale, whether the two-point scale is derived from push-button methodology or from a bisection of a continuous rating scale. The GOC curve and FORA based on the derived binary-decision data are worse than their equivalents based on the 64-point rating scale data, and yet this is essentially the same data set. Careful examination of Figure 8.3(a) shows that that the binary-decision experiments required at least twice as many replications than the rating scale experiment in order to achieve a given performance level. This was true, even after a point when the GOC curves based on binary-decision data were well defined in the ROC space (after 32 replications, say). When only a small number of binary-decision replications were combined, average performance was relatively poor was due to the resolution of GOC curves in the ROC space. Once many replications were combined (say 40 or more), GOC curves would have been adequately defined in the ROC space, but average performance was still poorer. This may have been due to loss of information due to categorisation (Watson et al., 1964). Both problems suggest there are reasons to prefer the use of multiple-point rating scales in GOC and ROC analysis.

[7] The measure $A'$ itself would not work for this purpose, because GOC curves based on two or more replications are comprised of more than a single point.

**Watson et al.'s (1964) study.** In the context of single-replication ROC curves, Watson et al. (1964) showed that the amount of information available in rating data that is generated by partitioning a continuous rating scale increases with the number of rating categories. They showed that for their data, there was a noticable increase in information, even going from two to three categories, and that improvement extended up to 20 categories.

**Taylor's (1984) simulations.** Further results supporting the use of multiple-point rating scales were reported by Taylor (1984), who ran computer simulations of unique-noise-affected observers in GOC experiments. Taylor's simulations were like implementations of a dice game, extended to model unique noise (Watson, 1963). Common noise was distributed according to Gaussian distributions of different means and unequal variances, and unique noise was distributed as an additive Gaussian variable. The simulation data was analysed in terms of two-point and 20-point rating scales (essentially partitions of the decision axis) using ACA to calculate FORAs based on $d_z$. Taylor (1984) found for both data sets that $d_z$ improved as a function of replications added, and approached the known theoretical value of $d_z$ in relatively shallow FORAs. The FORA based on the 20-point rating scale was consistently better (by 0.1-0.2 in the value of $d_z$) than the FORA based on binary-decision data. Taylor used the simulated data to estimate the unique-to-common noise variance ratio, $k$, (which was set to $\simeq 1.9$), and found that much better estimates were made using the 20-point data, compared to the binary-decision data. The error in estimates were about 0-10% for the 20-point data, depending on the number of replications combined, but anywhere from 10-50% for the binary-decision data.

## Conclusion

These experiments and simulations showed that it is more efficient to average out unique noise using a rating scale with moderate to high resolution than it is using a binary-decision scale. GOC analysis works best if the rating scale is well-defined. As more replications are combined in GOC analysis, the mean rating per stimulus tends to some expected value for each stimulus. In order for a set of mean ratings to approximate their expected values (over an entire stimulus set), fewer replications are needed if a continuous rating scale is finely partitioned than if it is coarsely partitioned. Similar arguments apply to partitions of decision axes, which theoretically underlie multiple-point rating scales (including binary-decision scales). If the goal in a discrimination task is to achieve the best performance, then highly partitioned continuous rating scales should be used when possible, unless there are very good reasons not to.

## 8.2 Amplitude discrimination FORAs for four observers

A multiple-replication, 2IFC aural amplitude discrimination experiment was run as part of Lapsley Miller et al.'s (1998) research project. The task was to decide in which of two observation intervals a noise signal had been added to a noise masker. Decisions were made using a continuous rating scale. The theoretical ROC curve for this experiment was unknown, but the decision axis was assumed to be continuous because of the nature of the stimuli.

### 8.2.1 Method

**Observers.** The same four observers from the discrete case, 2IFC frequency discrimination experiments in Section 8.1 also took part in this experiment.[8] Like the previous experiments, each of the observers ran multiple practice sessions until they demonstrated proficiency at the task.

**Stimuli.** The signals for the experiment were short-duration, band-pass filtered Gaussian noise transients with an equivalent rectangular bandwidth of 92 Hz, centred at 250 Hz. The maskers were short-duration, low-pass filtered Gaussian noise transients with an equivalent rectangular bandwidth of 1500 Hz. The signals and maskers were of the same duration, and were gated together using a Kaiser window with a shape parameter of 9. The absolute duration of the window was 20 ms, which gave an equivalent rectangular duration of 8.2 ms. The signal-to-noise ratio was 7.5 dB, and the gated masker had a spectrum level of 69 dB SPL. During experimental sessions, an 4 kHz low-pass analog Gaussian noise masker ran continuously at a spectrum level 33 dB SPL. All stimuli were presented diotically.

The same experimental system and equipment were used as in discrete case experiment described the preceding section, including the controlling computers, DAC, sound chamber, slider panel, headphones and headsets. The main differences in stimuli for this experiment were the bandwidth of the continuous masker (4 kHz), the clocking rate of the DAC (13800 Hz), the post-DAC filter (3 kHz low-pass), and the method of digital signal generation.

**Signal generation.**[9] The reproducible digital transients used in this experiment were based on computer-generated inverse fast Fourier transforms (IFFTs), implemented on an HP 9826 computer. A radix-2 IFFT was used to generate 32768-point digital time series. At the clocking rate of the DAC, each IFFT produced a 2374 ms time series, from which four non-overlapping sections of 20 ms duration (276 points) were selected at random, one for each of the four observers.

[8] Observers are numbered the same as in the preceding section.

[9] A complete description of the signal generation method is available in Lapsley Miller (1999).

Two separate sets of IFFTs were run, a signal-alone set and a noise-alone set. The spectrum input to each IFFT for either set was a random sample from a band-limited rectangular spectrum with zero power outside of the band. The amplitude within each spectral bin was uniformly distributed across IFFTs, and the input values were independent across spectral bins. The input to each signal-alone IFFT was a band-pass spectrum with an absolute bandwidth of 35 Hz, which contained 83 spectral components and was centred at 250 Hz. The input to each noise-alone IFFT was a low-pass spectrum with an upper cutoff of 1500 Hz, which contained 3561 spectral components.

Signal-alone and noise-alone transients were generated separately and stored on disk as a series of floating-point values. Prior to windowing, the distribution of instantaneous values for the signal-alone waveforms was Gaussian out to 2.5 standard deviations from the mean, and that for the noise-alone waveforms was Gaussian out to 3 standard deviations.

The signal-alone and noise-alone transients were randomly paired, additively mixed, windowed and converted into signal-plus-noise digital code sequences. Similarly, noise-alone transients by themselves were windowed and converted into noise-alone digital code sequences. After windowing, the equivalent rectangular bandwidth of the signal-alone transients was 92 Hz, and the equivalent rectangular bandwidth of the noise-alone transients remained the same at 1500 Hz. All code sequences were stored on disk so that the waveforms generated by the DAC were reproducible. Each of the floating-point signal-alone and noise-alone transients stored on disk contributed to one and only one stimulus transient. Each of the four observers had a unique stimulus set. Each IFFT that was computed contributed to one and only one transient per observer, in order to minimise the waveform correlations in a stimulus set.

**Experimental design.** Each of the four observers ran 16 replications, consisting of 1500 trials per replication, 750 trials per 2IFC event. These were run in sessions of 500 trials that took about 20 minutes to complete. A different haphazard trial sequence was used on each replication, and the sequence of 2IFC events was run-limited so that the same event could not occur more than 4, 5 or 6 trials in a row (the value was randomly chosen and changed from trial to trial).

Each trial consisted of a 100 ms warning interval, two 20 ms observation intervals separated by a 500 ms inter-stimulus interval, a 1500 ms decision interval, and a 900 ms reset interval. The reset interval was a minimum duration. The next trial could not begin until the slider had been reset to the extreme left, so the slider always started in the same position on each trial. A set of LED lights were switched on and off to mark the trial intervals. No trial-by-trial knowledge of results was given, but observers could later view their single-replication ROC curves at the conclusion of each replication.

The continuous rating scale was used to indicate confidence that the $\langle \mathrm{SN}, \mathrm{N} \rangle$ event order had occurred—the extreme left indicated zero confidence while the extreme right indicated 100% confidence, with increasing confidence indicated by an increasing slider

position going from left to right. The 12 cm long continuous rating scale was measured electronically, partitioned evenly into 64 categories, and ratings were stored as integers from 1 to 64.

### 8.2.2 Results

**ROC and GOC results.** Single-replication ROC curves for each observer are presented in Figure 8.4. There was appreciable variability across replications, both within and across observers. Observer 4 showed the most variability, having a standard deviation of values of $\mathcal{A}$ of 0.03, compared to about 0.02 for each of the other observers. $P(\mathrm{C})_{\mathrm{2IFC}}$, calculated from where an ROC curve crossed the negative diagonal, varied by as much as 0.09 (for Observer 4). The 16-replication mean ROC and GOC curves for each observer are shown in Figure 8.5. The mean ROC curve indicated expected single-replication performance and the GOC curve indicated the improvement in performance that was possible once unique noise was reduced. Observer 4 showed the best mean ROC and GOC performance of all four observers. Although the mean ROC performance differed among the other three observers, their GOC performance was fairly similar. This can also be seen in their FORAs.

**FORA results.** Each observer's 16-replication FORA and log-log plot based on $\mathcal{A}$ is shown in Figure 8.6. FORA values and parameters and estimated asymptotes for each observer (for the $\mathcal{A}$ measure only) are given in Table G.4 in Appendix G. The FORAs highlight individual differences that are harder to recognise in the ROC space alone. The FORAs in Figure 8.6 showed that of the four observers, Observer 3 in (circles and solid lines) had the worst average single-replication value of $\mathcal{A}$ of the four observers, and yet had the second best asymptote.

Observer 4 clearly had the best performance (triangles in Figure 8.6) of any of the four observers. In fact, the average single-replication value of $\mathcal{A}$ for Observer 4 (0.8077) was slightly better than the asymptote for Observer 1 (0.8058). This implied that Observer 1 could not have expected to perform better, on average, than Observer 4, no matter how many replications were run.

The FORAs for Observers 1 and 3 started at almost the same level, where the average performance for Observer 3 was slightly worse than performance for Observer 1. However, their FORAs crossed between 1 and 2 replications, and GOC performance for Observer 3 was better, on average, than performance for Observer 1, once data from more than one replication was combined.

The initial value of $\mathcal{A}$ for Observer 2 (0.7652) was more than 0.10 greater than the initial value for Observer 3 (0.6640). Nevertheless, the asymptote for Observer 3 (0.8417) was almost 0.02 greater than the asymptote for Observer 2 (0.8226). The large initial difference in average ROC performance between Observers 2 and 3 was solely due to unique

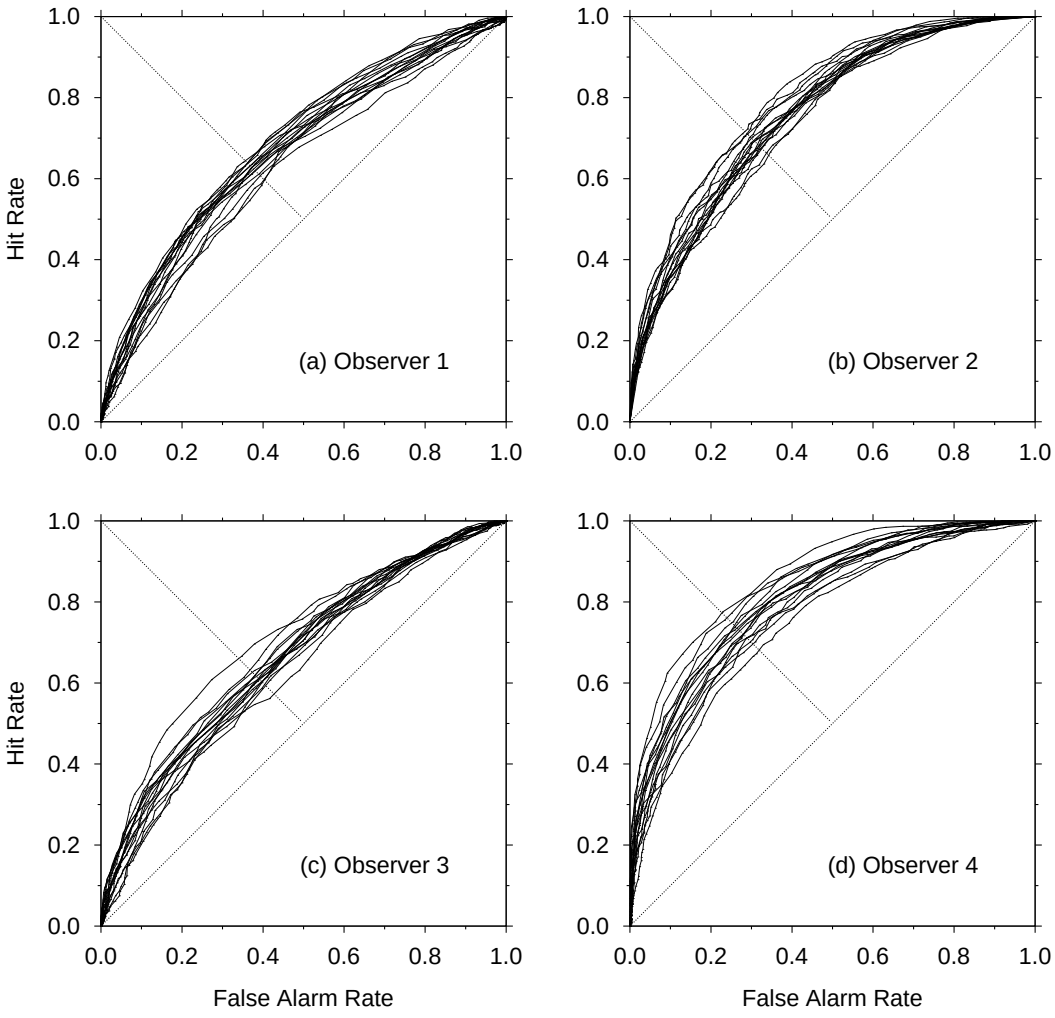


FIGURE 8.4: All 16 single-replication ROC curves for each observer in Lapsley Miller et al.'s (1998) 2IFC amplitude discrimination experiment.

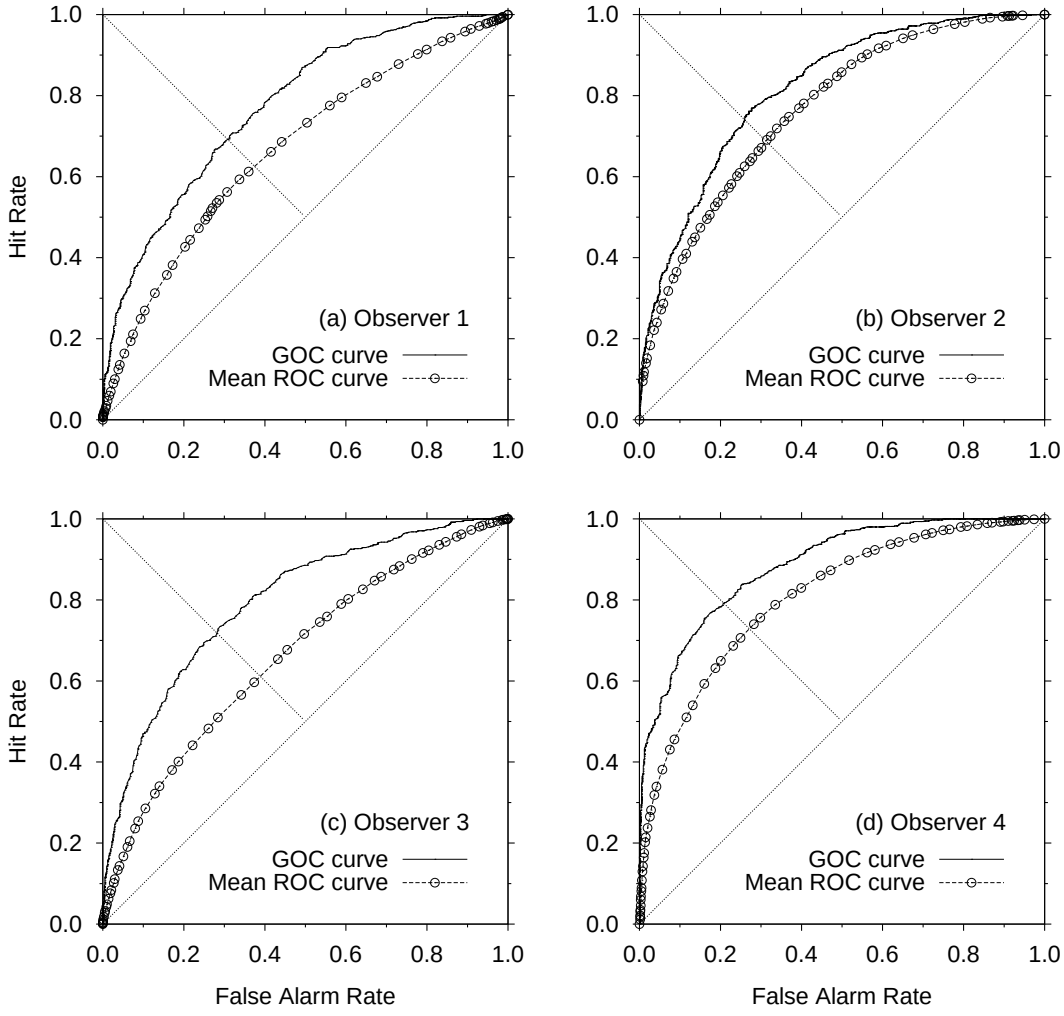


FIGURE 8.5: The 16-replication mean ROC curves and GOC curves for each observer in the 2IFC amplitude discrimination experiment (after Lapsley Miller et al., 1998, Figure 11).

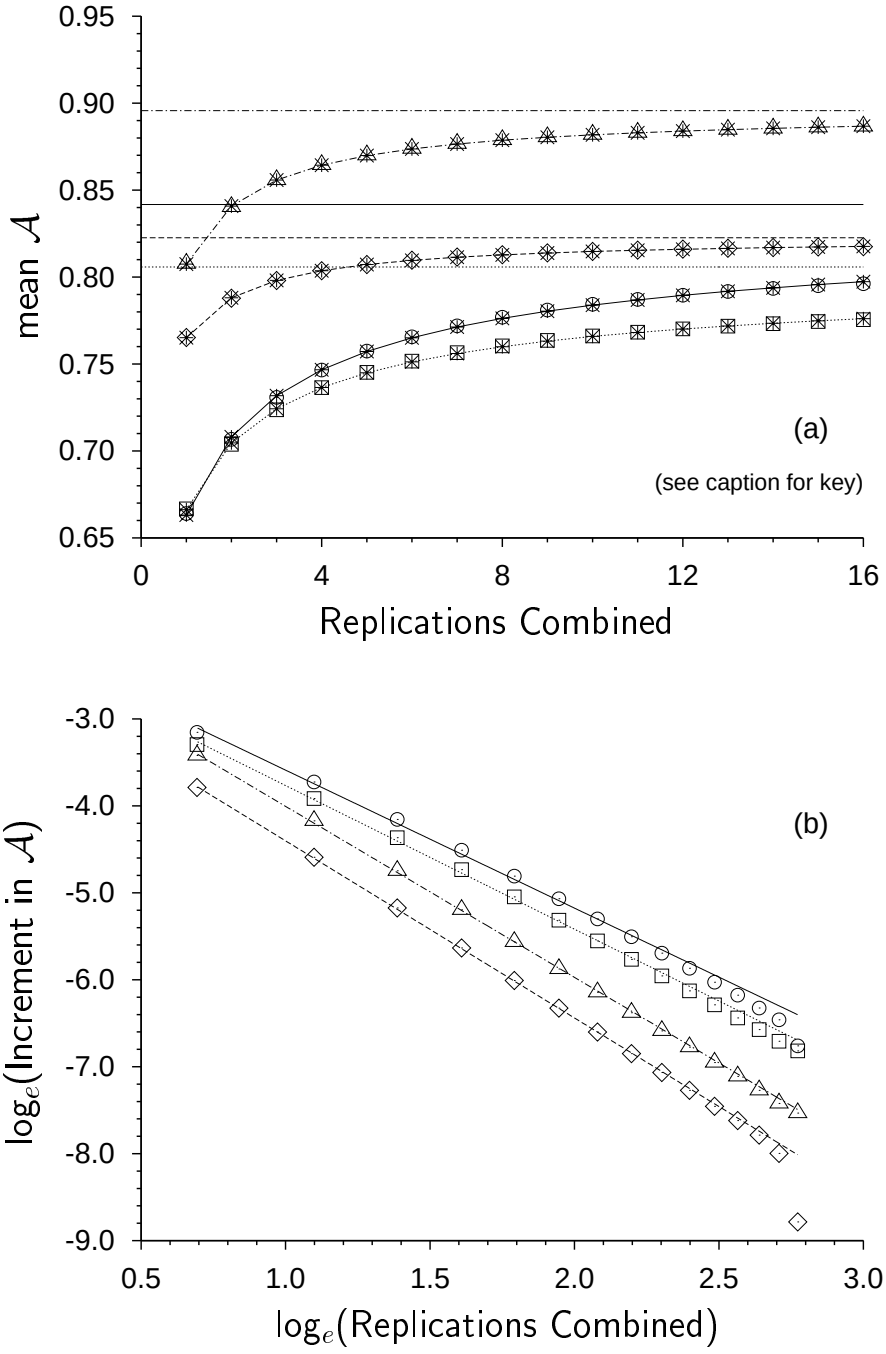


FIGURE 8.6: The 16-replication FORA and log-log plot for each observer in Lapsley Miller et al.'s (1998) continuous case 2IFC experiment. Key: Observer 1: □ and ·····; Observer 2: ◇ and - - - -; Observer 3: ○ and ——; Observer 4: △ and -·-·-·. (a) Mean value of $\mathcal{A}$ as a function of the number of replications combined. Horizontal lines denote asymptotes; curved lines with ∗ denote regression functions. (b) The accompanying log-increment in area versus log of replications combined for each observer, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each observer.

noise—inconsistent decision making—rather than any inherent, unique-noise-free ability to discriminate between events. The reversal of relative performance for the same two individuals also occurred in Whitmore et al. (1993) experiment (described in Chapter 7). In a much larger series of experiments, which are the topic of Section 8.4.2, Lapsley Miller (1999) found that inter-observer FORA performance reversal occurred 22% of the time, which suggests that this type of result is common, and that comparisons of observers based on single-replication performance should be done with caution.

Equation 6.5 provided excellent regression fits to the empirical FORAs across a wide range of performance levels. The log-log plots were fairly linear for all observers, with a slight tendency to curve downwards. This implied that the asymptotes overestimated true asymptotic performance, at least to some small degree. The log-log plot for Observer 3 (circles in Figure 8.6(b)) was the most curved. This relates to the fact that the regression-FORA for Observer 3 (Figure 8.6(a)) was slightly off-centre compared to the data series upon which it was based. The extent to which the asymptote for Observer 3 is overestimated is not known. However, the log-log plot and amount of improvement in the FORA in Figure 8.6 are both comparable to what was shown in Figures 6.3 and 6.4 for Taylor et al.'s (1991) experiment (Chapter 6). There it was known that the asymptote overestimated the theoretical value of $\mathcal{A}$ by 0.0035. Based on this analogy, the asymptote for Observer 3 in Figure 8.6(a) could be overestimated by perhaps 0.010 or less. This still places Observer 3 asymptotically ahead of Observers 1 and 2.

**Conclusion**

Substantial unique noise effects were successfully reduced in Lapsley Miller et al.'s (1998) 2IFC amplitude discrimination experiment by GOC analysis and by ACA. The FORAs calculated from this data set provided further evidence that the data pattern described by Equation 6.5 holds across different experiments. The FORAs showed that among a group of observers, relative performance based on single-replication ROC analysis, and even GOC analysis based on 16 replications, *does not necessarily reflect relative asymptotic unique-noise-free performance.*

## 8.3 FORAs over a wide range of performance levels

Results from a previously unpublished, multiple-replication, 2IFC aural amplitude discrimination experiment are presented here. The task was to decide in which of two observation intervals a noise signal had been added to a noise masker. Five different signal-to-noise ratios were used, and the performance of observers ranged from near-chance to near-perfect. The results demonstrate ACA and FORA regression over a full range of performance levels, particularly at very high levels. Unique-noise-free psychometric functions are also estimated.

The 2IFC experiment was one of three previously unpublished amplitude discrimination experiments. Results for the other two experiments are not presented here, but they are mentioned because of an interwoven experimental design. The 2IFC experiment is reported because it is the best for demonstrating FORA regression at high performance levels ($d' > 3$). This 2IFC experiment included the highest performance levels in any of the data sets presented in this thesis.

### 8.3.1 Method

**Observers.** Two experienced observers took part in the experiment. Both had participated in previous SIFC and 2IFC amplitude discrimination experiments, and both were familiar with continuous rating scale sliders.[10]

#### Stimuli and equipment

The signals for the experiment were short-duration, band-pass filtered Gaussian noise transients with an equivalent rectangular bandwidth of 50 Hz, centred at 500 Hz. The maskers were short-duration, low-pass filtered Gaussian noise transients with an equivalent rectangular bandwidth of 4000 Hz. The signals and maskers were of the same duration, and were gated together using a Kaiser window with a shape parameter of 9. The absolute duration of the window was 48.6 ms, which gave an equivalent rectangular duration of 20 ms. Signals were presented at five different signal-to-noise ratios (SNRs), namely $-5$, 0, 4, 8 and 12 dB. The gated masker had a spectrum level of 76 dB SPL. During experimental sessions, an 8 kHz low-pass analog Gaussian noise masker ran continuously at a spectrum level 20 dB SPL. All stimuli were presented diotically.

The experiment was run using an IBM-compatible personal computer (PC) and a Hewlett Packard 82324A High-Performance Measurement Coprocessor, housed inside the PC.[11] The coprocessor controlled experimental trial sequences and data collection, whereas the PC controlled stimulus production. The computer also housed a Turtle Beach Monterey soundcard which was used for waveform production under the control of the PC. The Monterey soundcard was used for its CS4329 16-bit digital-to-analog converter (DAC), which was clocked at 44100 Hz, and followed by an on-board 22 kHz low-pass analog smoothing filter.[12] The output of the soundcard was attenuated, mixed with the continuous analog masker, and passed to a headset amplifier. Stimuli were presented to observers

[10] These were the same two observers who took part in the amplitude discrimination experiment in Chapter 7. Observer numbering has been kept the same.

[11] The PC had a 33 MHz Intel 486DX CPU, 20 MB of memory, and ran under MS-DOS 6.22. The coprocessor had a 16 MHz Motorola MC 68030 CPU, 2 MB of memory, and ran HP Basic 5.0.

[12] The soundcard was chosen for its specifications. The combined DAC and filter system had a noise floor (specified from the documentation) of -120 dB over the audible range. The soundcard was originally configured with a Turtle Beach Rio MIDI synthesizer card attached to it. This synthesizer was removed as it was not required, and the noise floor decreased further without it. The filter had a flat response below 20 kHz, a 3dB bandwidth of 22090 Hz and a rolloff of 331 dB per octave.

in a sound-attenuated chamber, through TDH-39 headphones mounted in Rudmose Tracor RA-125 headsets with MX-41/AR cushions. Apart from the digital waveform generation and production, much of the rest of the experimental system concerned with stimulus production and shaping was the same as for the other experiments described in this and previous chapters. The equipment that was common across experiments included the passive attenuators, analog mixer, headset amplifier, make and model of headset, rating sliders, and soundchamber.

**Signal generation.** The digital transients used in this experiment were computer-generated using inverse fast Fourier transforms (IFFTs). Transients were generated and stored as digital code sequences on disk, so that they could be reproduced across replications. Code sequences were generated using an IBM-compatible PC.[13] A radix-2 IFFT algorithm was used to generate $2^{21}$-point digital time series. At the clocking rate of the DAC, each IFFT produced a 47554 ms time series, from which non-overlapping sections of 48.6 ms duration (2143 points) were selected. Two separate sets of IFFTs were run, a signal-alone set and a noise-alone set. The spectrum input to each IFFT for either set was a random sample from a band-limited rectangular spectrum with zero power outside of the band. The amplitude within each spectral bin was uniformly distributed across IFFTs, and the input values were independent across spectral bins. The input to each signal-alone IFFT was a band-pass spectrum with an absolute bandwidth of 44 Hz, which contained 2093 spectral components and was centred at 500 Hz. The input to each noise-alone IFFT was a low-pass spectrum with an upper cutoff of 4000 Hz, which contained more than 190 000 spectral components. After windowing, the equivalent rectangular bandwidth of the signal-alone transients was 50 Hz, and the equivalent rectangular bandwidth of the noise-alone transients remained the same at 4000 Hz.

Multiple, non-overlapping, 2143-point sections were selected from the 2.1 million-point, long-duration time series produced by each IFFT. Each section was used to generate a stimulus transient. Ideally, a different IFFT would be used to calculate each section, to ensure uncorrelated stimuli, but the computation time per IFFT made this prohibitive. Instead, multiple sections were selected per IFFT, in a way that took the autocorrelation function of the IFFT into account. One hundred sections were randomly selected from each signal-alone IFFT, and two hundred sections from each noise-alone IFFT. The autocorrelation function depended on the bandwidth ($W$) of the IFFT input spectrum, and essentially dropped to zero (on average) for time lags greater than $4/W$ seconds. The minimum lag that was used was about $14.5/W$ for the signal-alone sections, and $8/W$ for the noise-alone sections, each of which was partly related to the number of sections chosen per IFFT. As well as setting a minimum lag, the starting point of each of the sections was randomly chosen from a range of 2143 values, to further minimise correlation.

[13] The PC used for signal-generation, which was not the experimental PC, had a 99 MHz Intel 486DX CPU, 64 MB of memory, and ran under MS-DOS 6.22.

Signal-alone and noise-alone transients were generated separately and stored on disk as a series of floating-point values. Prior to windowing, the distribution of instantaneous values was Gaussian out to 3 standard deviations for the signal-alone waveforms, and was Gaussian out to 4 standard deviations for the noise-alone waveforms.

The signal-alone and noise-alone transients were randomly paired, additively mixed, windowed and converted into signal-plus-noise digital code sequences. Similarly, noise-alone transients by themselves were windowed and converted into noise-alone digital code sequences. All code sequences were stored on disk so that the waveforms generated by the DAC were reproducible. Each of the floating-point signal-alone and noise-alone transients stored on disk contributed to one and only one stimulus transient.

### Experimental Design

Each observer replicated the experiment eight times using the set of reproducible stimulus transients. Although the reproducible transients were the same across observers, each observer had a unique set of pairings of $SN$ and $N$ transients, one pairing per 2IFC trial. The unique pairings meant that each observer, in effect, had a separate set of 2IFC stimuli. As a result, GOC analysis could be done only within observers and not across observers.

The task of the observer was to decide in which of two observation intervals the signal had occurred. A continuous rating slider was used to indicate confidence that the $\langle \mathrm{SN}, \mathrm{N} \rangle$ event order had occurred. The extreme left indicated zero confidence that $\langle \mathrm{SN}, \mathrm{N} \rangle$ had occurred, while the extreme right indicated 100% confidence, with increasing confidence indicated by an increasing slider position. The 12 cm long continuous rating scale was partitioned evenly into rating categories, and ratings were stored as integers from 1 to 1980. For the purposes of data analysis, the 1980-point scale was uniformly partitioned into a 600-point scale (due to memory limitations in the data analysis program).

Each trial consisted of a 300 ms warning interval, two 49 ms observation intervals separated by a 500 ms inter-stimulus interval, a 1250 ms decision interval, and a 1250 ms reset interval. The reset interval allowed the observer to reset the slider. The next trial could not begin until the slider had been reset to the extreme left. A set of LED lights on the slider panel were switched on and off to mark the trial intervals. No trial-by-trial knowledge of results was given, but observers could view their single-session ROC curves (for 5 SNRs) at the end of each session, and their single-replication ROC curves after each completed replication.

The 2IFC experiment was one of three amplitude discrimination experiments that were run concurrently. The other two experiments involved SIFC tasks. One SIFC experiment used the same types of noise signals as the 2IFC experiment, and the other used tonal transients of the same duration and center frequency as the other two conditions. The stimulus sets were different in each experiment. Trial blocks from all three experiments were interwoven in a haphazard sequence, constrained so that sessions from the same

experiment were not run more than twice in a row. To reacquaint the observer with the type of stimuli for each 2IFC session, six preview trials were run at the beginning of each session using stimuli from the higher SNRs (12 and 8 dB), alternating between 2IFC events on successive trials. The stimuli used in these preview trials were not part of the main stimulus set and no data was collected for them.

Each replication consisted of 1000 trials per SNR, or 5000 trials in total. Trials were run in 20 sessions of 250 trials per session per replication, and SNRs were intermixed within sessions (Tucker, Evans, & Jeffress, 1967; Emmerich, 1968b). Each observer completed 160 2IFC sessions in total over the eight replications. Each session took about 12 minutes to run. A different haphazard sequence was generated for each observer and for each replication, so that any time-order effects based on the trial sequence would contribute to unique noise rather than common noise (and could later be removed by GOC analysis). Each haphazard 5000-trial sequence was partitioned into sections of 250 trials to determine the trial sequence for each session. Within a session, no constraint was made on the number of trials per SNR. The trial sequence was run-limited so that the same 2IFC event *and* SNR could not occur more than 3 or 4 times in a row (with either 3 or 4 randomly chosen with equal probability on each trial).

Practice sessions were run over the course of several months to familiarise the observers with the tasks, and to finalise interval timings and stimulus parameters. Most of the later practice sessions were run with SNRs set from $-5$ dB to 15 dB. The levels were later dropped to $-5$, 0, 4, 8 and 12 dB, because the task was found to be too easy at 15 dB.

Experimental sessions were run over a 6 month time period. There were no restrictions as to how many sessions either observer could or should run in any given day, nor the time of day to run. On average, each observer completed one 2IFC session per day. Part way through the first replication, Observer 2 halted his data collection due to circumstances unrelated to the experiment. Sessions began again after a break of seven weeks. The longest break Observer 1 took from data collection was 11 days.

### 8.3.2 ROC and GOC results

ROC, mean ROC and GOC curves were calculated separately for each observer at each signal-to-noise ratio. These are presented in Figure 8.7. Panels(a) and (b) show each observer's 8 ROC curves at each SNR. Performance for each observer was affected by unique noise at all SNRs. The most consistent set of ROC curves was for Observer 2 at the highest SNR (12 dB). Apart from that subset of data, ROC curves for a given SNR varied appreciably across replications, both in the shape of the curve and in the implied level of performance. The standard deviation of $\mathcal{A}$, across replications at each SNR, ranged from 0.007 to 0.018 for Observer 1, and from 0.002 to 0.025 for Observer 2. These values are small, because they are based on the area under an entire ROC curve. For a given false alarm rate, the hit rate at each SNR could vary by as much as 0.2,

depending on the SNR. The ROC variability shown in Figures 8.7(a) and 8.7(b) is typical of such experiments. Many similar examples are given in Lapsley Miller (1999), whose experiments are described in Section 8.4.

Mean ROC curves for each SNR and for each observer are shown in Figures 8.7(c) and 8.7(d). Except at the lowest SNR (−5 dB), Observer 2 performed better on average than Observer 1, particularly at the highest SNR. A similar pattern also held for the 8-replication GOC curves, which are shown in Figures 8.7(e) and 8.7(f), with Observer 2 having better GOC performance at all but the lowest SNR.

There was a general asymmetry in both the mean ROC curves and the GOC curves. Mean ROC and GOC curves indicated slightly higher performance to the right of (or above) the negative diagonal when compared to the left of (or below) the negative diagonal. The GOC curve for Observer 2 at 8 dB was an exception to this, with the asymmetry going the other way. The asymmetry is curious in light of the theoretical result that 2IFC ROC curves should be symmetrical in the negative diagonal (Green & Swets, 1974). Possible reasons include time-asymmetric masking effects or memory effects. An inspection of histograms of ratings for each observer at each SNR did not reveal any obvious trend, but the data was so variable that *any* small effect would be swamped. Since only eight replications were run, and the observers were inconsistent, unique noise sampling variability could fully account for the results. Similarly, common noise sampling variability, due to stimulus sampling and pairing, could also account for the results (even though there were 1000 trials per event). There is a paucity of 2IFC ROC curves in the literature, and the symmetry of *empirical* 2IFC ROC curves has yet to be assessed. While 2IFC ROC symmetry may be predicted in theory (Green & Swets, 1974; Egan, 1975), Figure 8.7 indicates that is not necessarily the case in practice.

### 8.3.3 FORA results

FORAs for $\mathcal{A}$ and $d'$ are described here. FORA values, parameters and estimated asymptotes for this experiment are given in Table G.5 in Appendix G.

**FORAs based on $\mathcal{A}$.** FORAs based on $\mathcal{A}$ were calculated for each observer at each SNR. These are presented with their associated log-log plots in Figure 8.8 for Observer 1, and in Figure 8.9 for Observer 2. Both ROC and GOC performance (the first and last FORA points, respectively) improved with increasing SNR, as expected. The set of asymptotes for each observer also followed the order of SNRs. This was expected, but was not guaranteed. In multiple-SNR experiments, even for the same observer, the order of asymptotes across SNRs does not *have* to reflect the order of mean ROC curves, or GOC curves, across SNRs.

The regression-FORAs in Figures 8.8 and 8.9 fitted the empirical FORAs so well that regression points are indistinguishable from data points on the scale that is shown (with

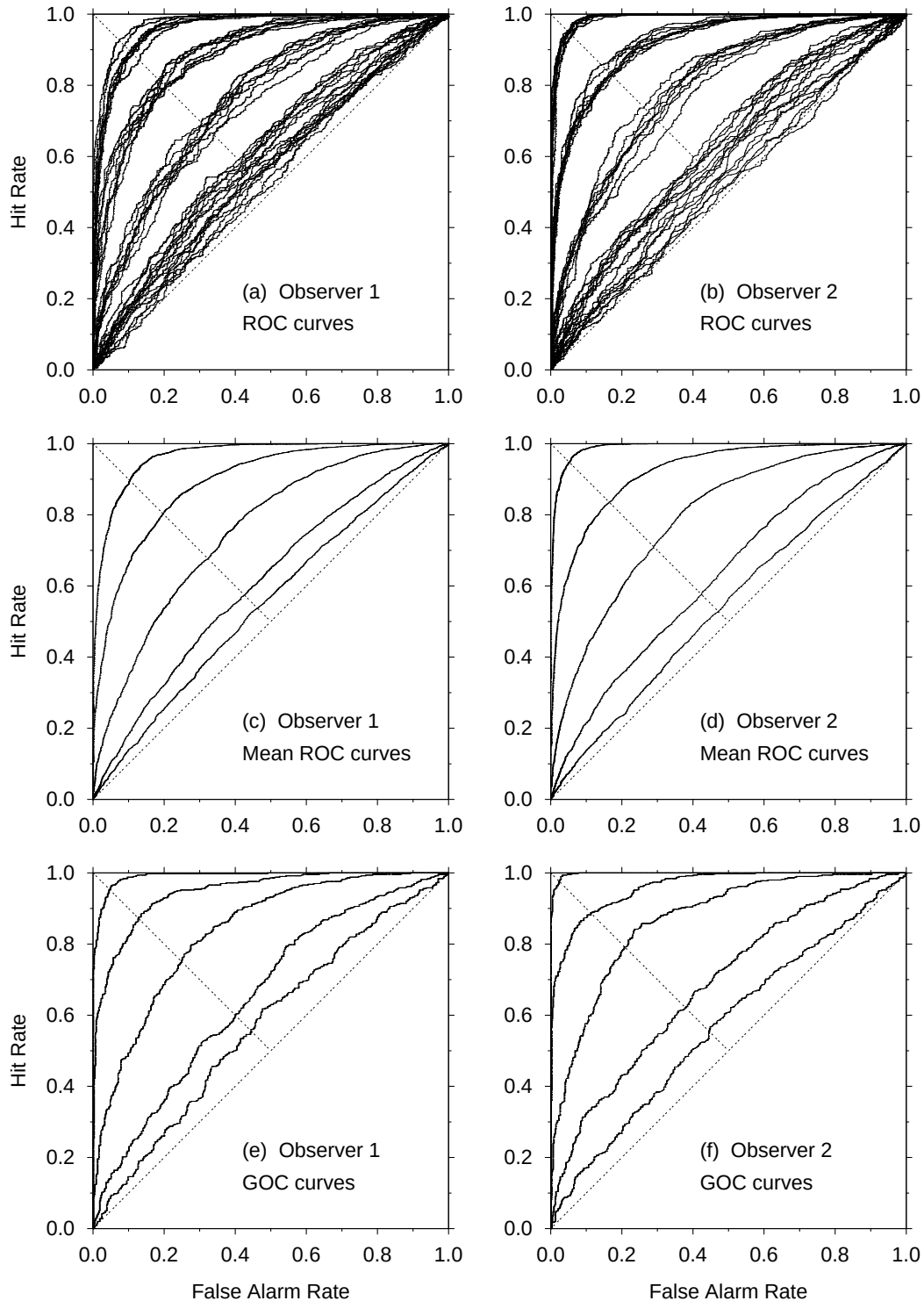


FIGURE 8.7: ROC, mean ROC and GOC curves for each observer at each signal-to-noise ratio (SNR) in the 2IFC amplitude discrimination experiment. The SNRs were $-5$, 0, 4, 8 and 12 dB. Within each panel, curves improve towards the top-left with increasing SNR. Left-hand panels show curves for Observer 1 and right-hand panels show curves for Observer 2. The ROC curves for $-5$ dB and 0 dB in panels (a) and (b) overlap slightly, but are generally distinct, with the curves for 0 dB being higher than those for $-5$ dB.

$\mathcal{A}$ between 0.5 and 1.0). The log-log plots were all highly linear for the first four SNRs, but were less so for the highest SNR. Observer 2 performed better than Observer 1 at all but the lowest SNR (−5 dB). Potential improvements in $\mathcal{A}$ were smallest at the lowest and highest SNRs, and largest at the middle SNR (4 dB). The potential improvement was as much as 0.1 in value at 4 dB for both observers.

There were two unusual patterns in the FORAs based on $\mathcal{A}$. First, although the empirical FORA for Observer 1 at −5 dB was higher than that for Observer 2, the asymptote for Observer 2 was higher than that for Observer 1. Second, the linear form of the log-log plot for Observer 2 broke down at the highest SNR (12 dB). This particular FORA was associated with a very high performance level ($\mathcal{A} > 0.99$). Each of these patterns are examined in detail later. Performance across observers at −5 dB is discussed in Section 8.3.4, and performance for Observer 2 at 12 dB is discussed in Section 8.3.5.

**FORAs based on $d'$.** FORAs based on $d'$ were also calculated for each observer at each SNR, shown in Figure 8.8 for Observer 1, and in Figure 8.9 for Observer 2. The pattern of results was different for $d'$ compared to the results based on $\mathcal{A}$, which reflected the fact that (as a measure) $d'$ is unbounded above, whereas $\mathcal{A}$ is bounded above. It is clear in Figures 8.8 and 8.9 that not only only did $d'$ increase with SNR, but that the amount of improvement in $d'$ also increased with SNR.[14] Potential improvement in asymptotic performance, compared to single-replication performance, was generally large. Values of $d'$ improved by 30-70% for each observer at the upper four SNRs, and by up to 100% at the lowest SNR. At the highest SNR, $d'$ improved in value by more than 1.0 at the highest SNR.

**Log-log plots based on $\mathcal{A}$ and $d'$.** In general, the total amount of improvement in a FORA affects the location of the associated log-log plot, regardless of the measure of sensitivity. If the total FORA improvement is large, then FORA increments are large, and consequently, the associated log-log plot is higher. If the total FORA improvement is small, then FORA increments are small, and consequently, the log-log plot is lower down. The log-log plots based on $d'$ in (Figures 8.10 and 8.11) increased with increasing SNR, whereas the log-log plots based on $\mathcal{A}$ (Figures 8.8 and 8.9) overlapped for different SNRs. This reflected the fact that $d'$ was unbounded above, and the total improvement in $d'$ increased with increasing SNR, whereas $\mathcal{A}$ was bounded above at unity, so the total improvement was constrained at the higher SNRs.

A concept introduced previously in Section 6.2 is the *relative curvature* of a FORA, indicated by the $\mu$ parameter, which is the slope of a linear log-log plot. The relative curvature is an indication of the rate at which a FORA approaches its asymptote, regard-

[14] Except for Observer 2 for 4 dB versus 8 dB. The improvement in $d'$ is 0.6020 for 4 dB, compared to 0.5891 for 8 dB.

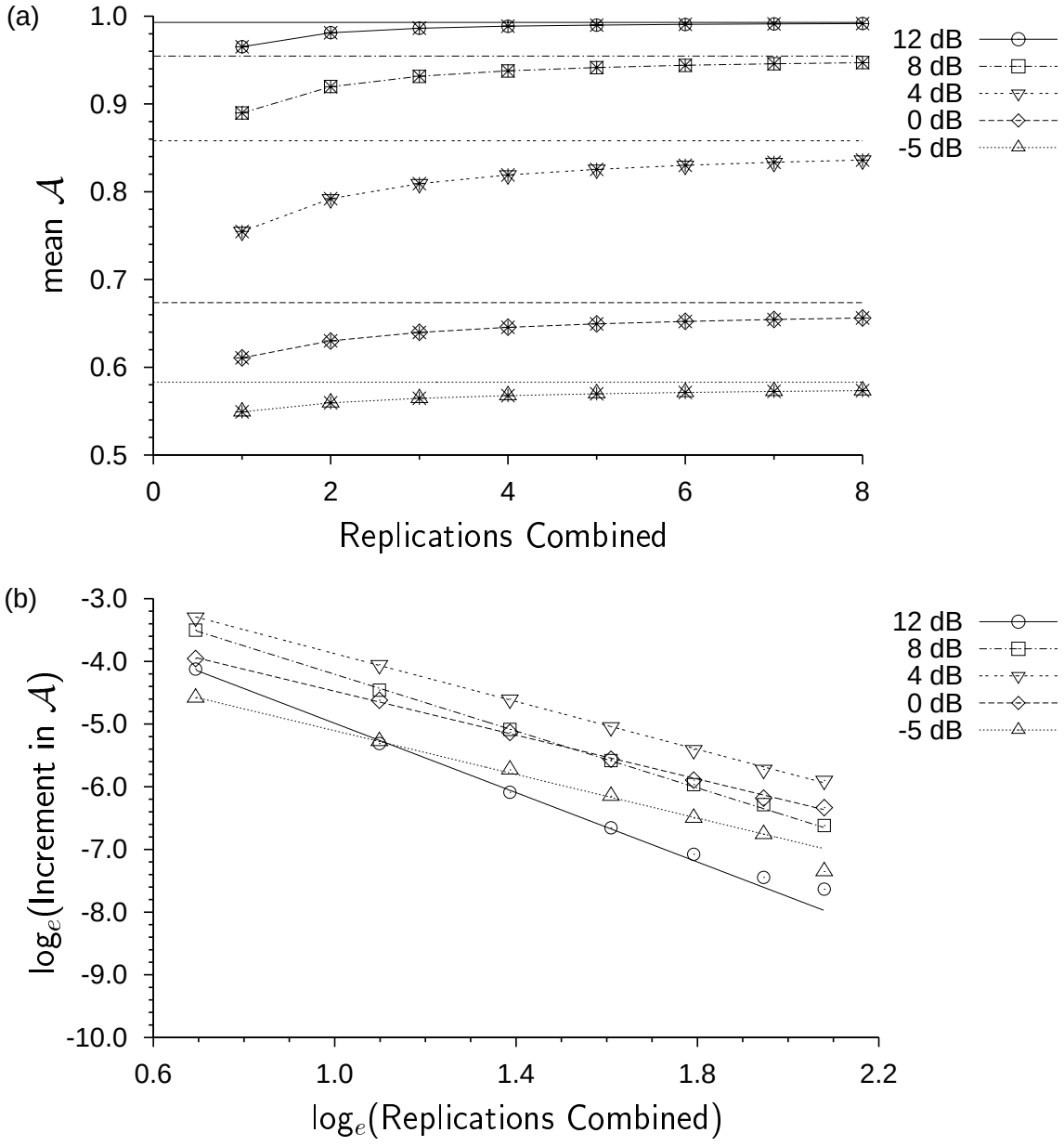


FIGURE 8.8: The 8-replication FORA and log-log plot based on $\mathcal{A}$ for Observer 1 at each signal-to-noise ratio (SNR). (a) Mean value of $\mathcal{A}$ as a function of the number of replications combined. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (b) The accompanying log-increment in area versus log of replications combined for each SNR, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each SNR.

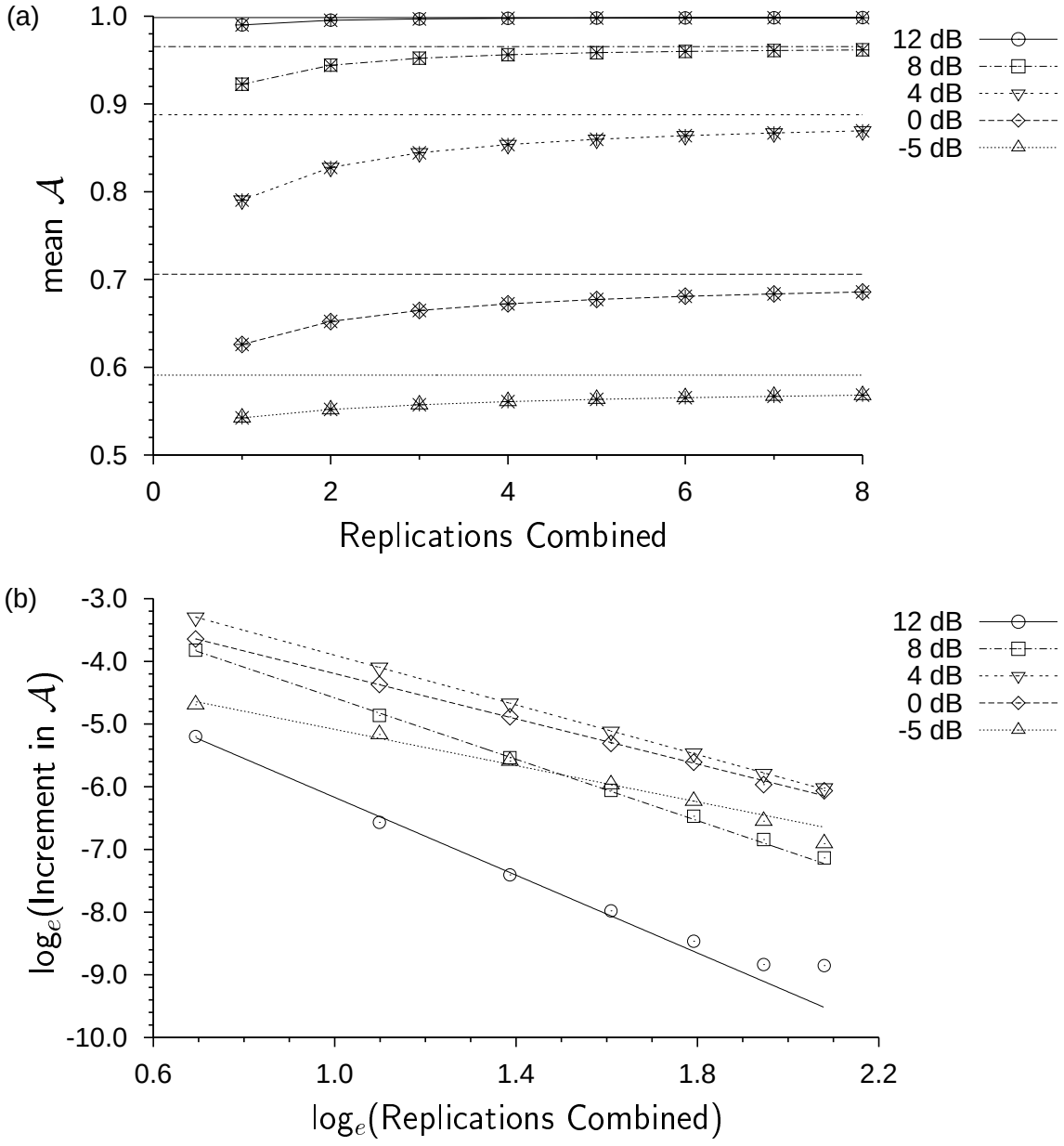


FIGURE 8.9: The 8-replication FORA and log-log plot based on $\mathcal{A}$ for Observer 2 at each signal-to-noise ratio (SNR). (a) Mean value of $\mathcal{A}$ as a function of the number of replications combined. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (b) The accompanying log-increment in area versus log of replications combined for each SNR, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each SNR.

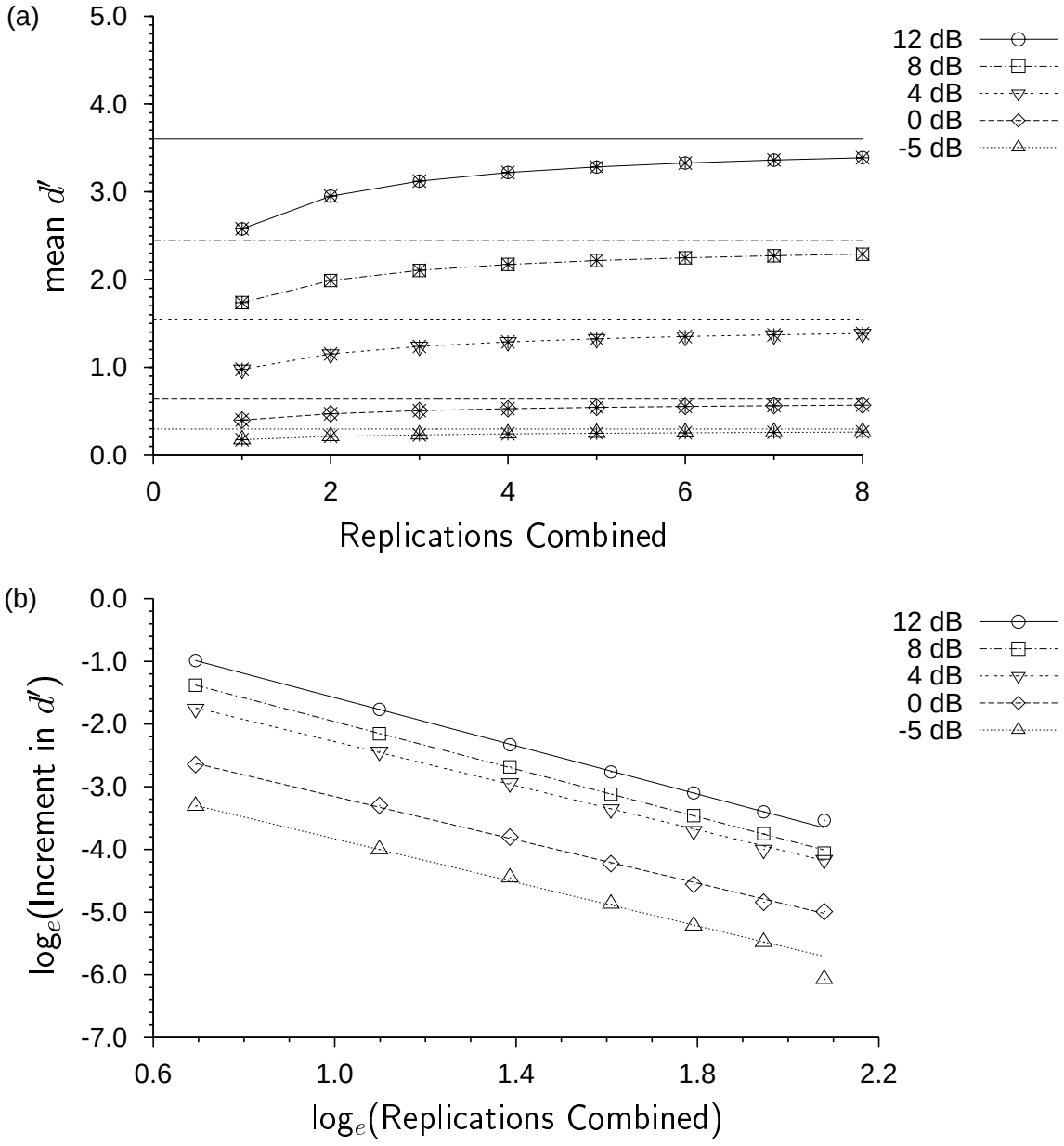


FIGURE 8.10: The 8-replication FORA and log-log plot based on $d'$ for Observer 1 at each signal-to-noise ratio (SNR). (a) Mean $d'$ value as a function of the number of replications combined. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (b) The accompanying log-increment in $d'$ versus log of replications combined for each SNR, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each SNR.

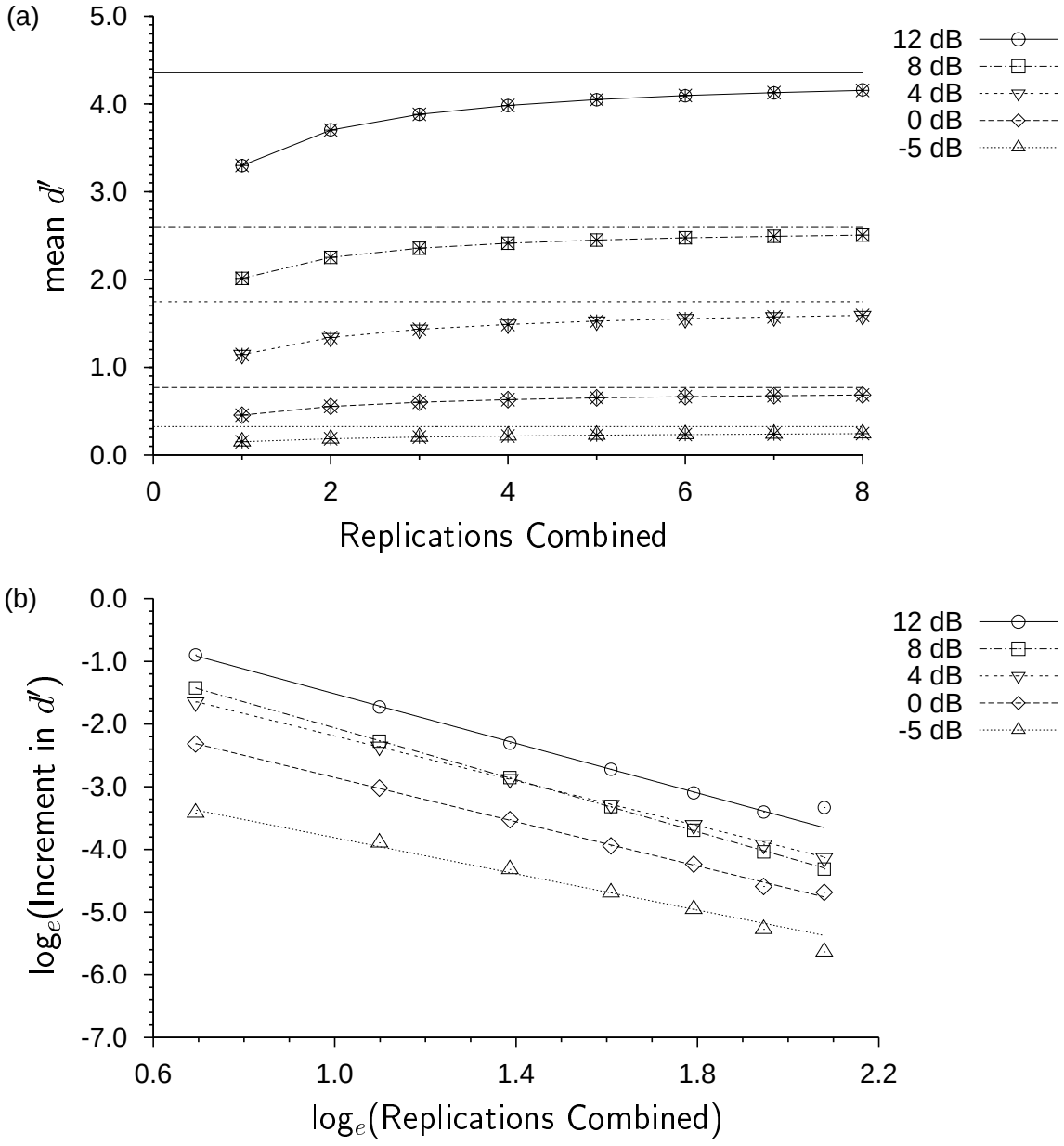


FIGURE 8.11: The 8-replication FORA and log-log plot based on $d'$ for Observer 2 at each signal-to-noise ratio (SNR). (a) Mean $d'$ value as a function of the number of replications combined. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (b) The accompanying log-increment in $d'$ versus log of replications combined for each SNR, where straight lines indicate log-log relationships based on parameters from the FORA regression function for each SNR.

less of the total improvement of the FORA. The steeper the slope of the log-log plot, the greater the relative curvature, which implies that fewer replications are needed to attain a given proportion of the possible improvement in performance.

For each observer, the slopes of the log-log plots based on $d'$ were similar across SNRs. In contrast, the slopes of the log-log plots based on $\mathcal{A}$ increased systematically with increasing SNR and absolute performance levels, regardless of the location of the log-log plot. This showed that the relative curvatures of FORAs based on $d'$ were similar across performance levels, whereas the relative curvatures of FORAs based on $\mathcal{A}$ increased with performance level. Lapsley Miller (1999) found similar data patterns for FORAs based on $\mathcal{A}$ and on $d'$ over a wide range of stimulus parameters.

### 8.3.4 FORAs at low performance levels

FORAs for both observers at the lowest SNR ($-5$ dB) appeared flat in Figures 8.8 to to 8.11, because of the plotting scales that were used. These FORAs are re-presented here in more detail. Figure 8.12(a) shows the FORAs based on $\mathcal{A}$, and Figure 8.12(b) shows the FORAs based on $d'$. Figure 8.12 shows how FORAs that have only a small total improvement, such as those at the lowest SNR, also follow the same form as FORAs that have a much larger total improvement, such as those at moderate to high SNRs. This much can also be inferred from the log-log plots presented in Figures 8.8 to 8.11, which are all fairly linear at $-5$ dB.

The results at $-5$ dB imply an order reversal of observers according to performance value. GOC results based on one to eight replications suggest that Observer 1 was better than Observer 2. The asymptotes, however, suggest that Observer 2 would be better than Observer 1 if more replications were run. The regression-FORAs may be extrapolated to a point where they cross each other. The functions based on $\mathcal{A}$ would cross after 46 replications at $\mathcal{A} = 0.5802$. The functions based on $d'$ would cross after 48 replications at $d' = 0.2865$. Similar types of crossover were also found in experiments reported in other chapters.

### 8.3.5 Ceiling effects at high performance levels

Out of the data sets presented so far in this chapter, the highest performance of all occurred for Observer 2 at 12 dB in the current 2IFC experiment. The FORA based on $\mathcal{A}$ had an initial value of 0.9900, which reached 0.9984 after 8 replications, and implied an asymptote at 0.9985. This particular data gives the best (and perhaps only) example of a ceiling effect in FORA regression. By comparison, the FORA for Observer 1 at 12 dB implied an asymptote at $\mathcal{A} = 0.9929$, but even that level is not high enough to show a

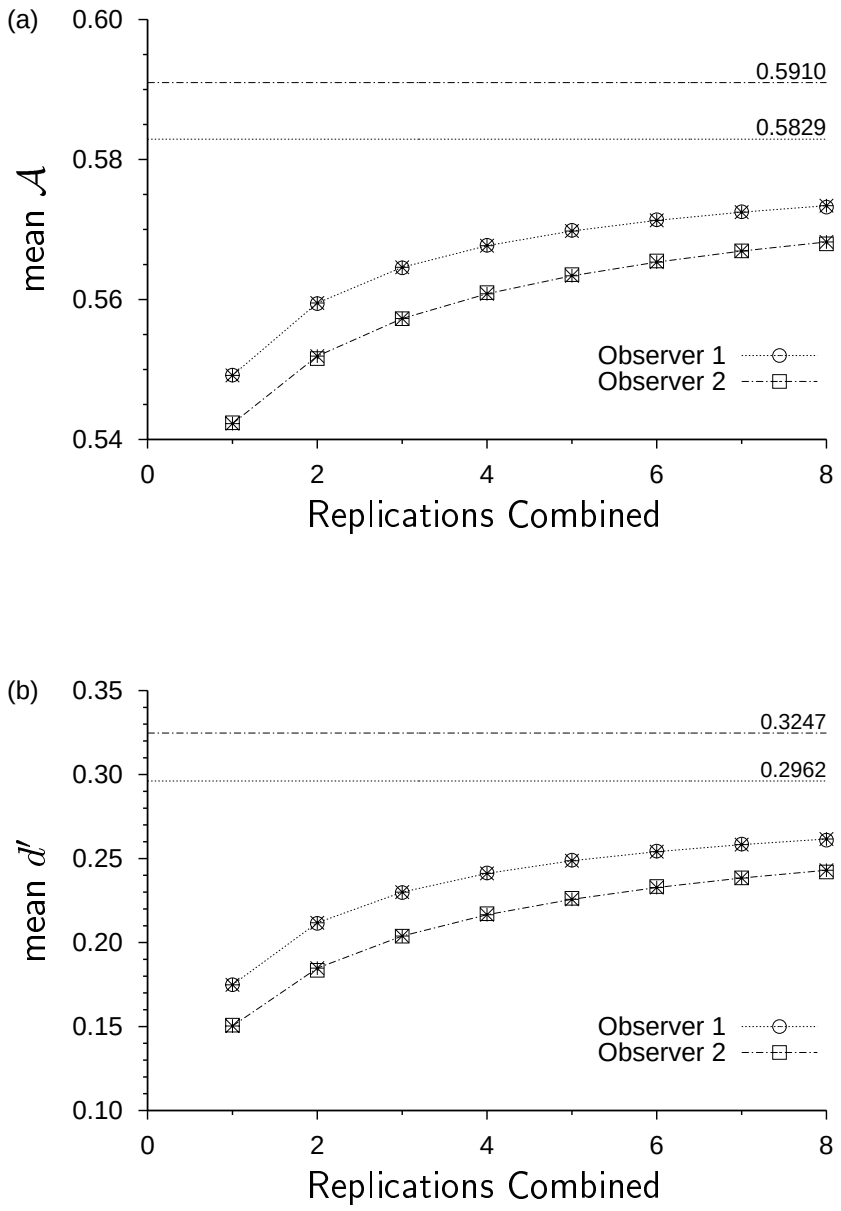


FIGURE 8.12: Detail of the lowest FORAs (at 5 dB) for both observers, based on $\mathcal{A}$ and $d'$. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. Observer 1 showed better GOC performance (for the finite number of replications run), but Observer 2 showed better asymptotic performance. (a) Mean $\mathcal{A}$ as a function of replications added (taken from Figures 8.8 and 8.9). (b) Mean $d'$ as a function of replications added (taken from Figures 8.10 and 8.11).

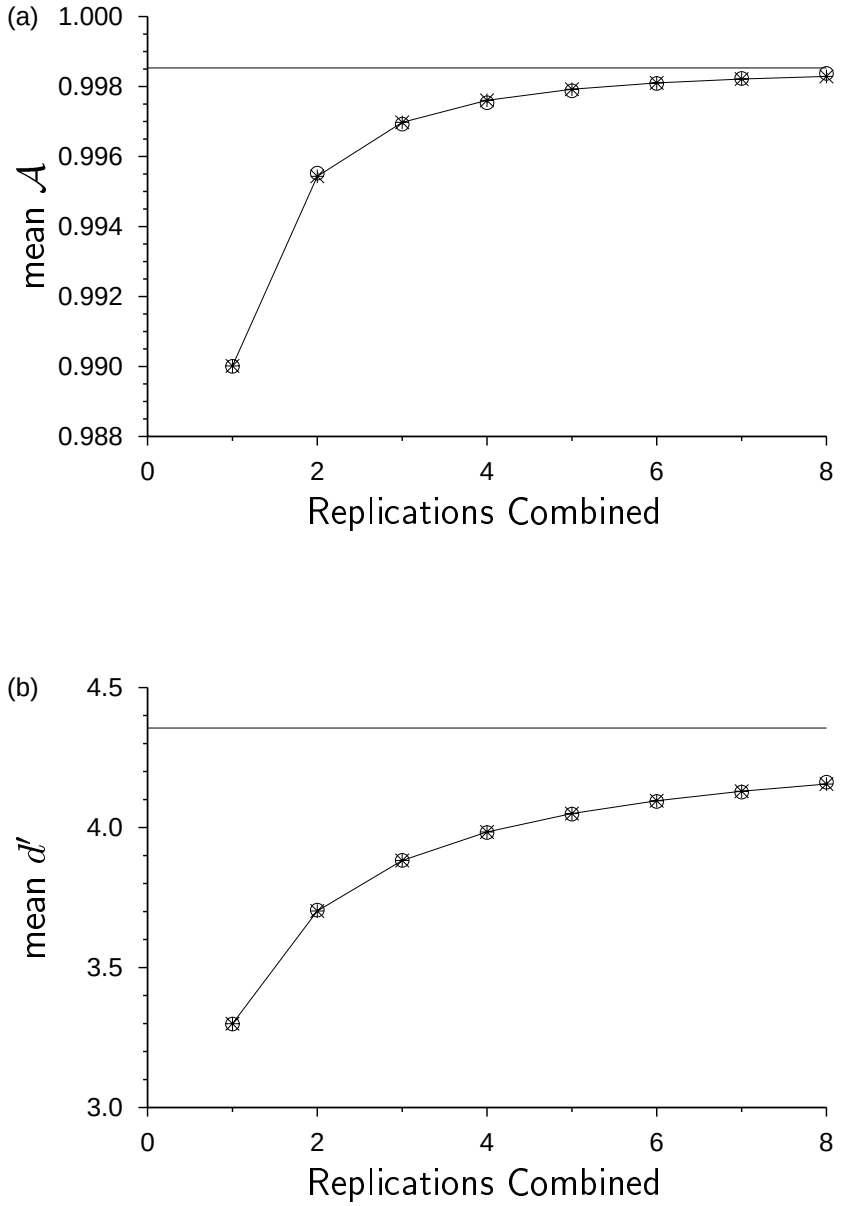


FIGURE 8.13: Detail of FORAs at the highest performance level (at 12 dB) for Observer 2, based on $\mathcal{A}$ and $d'$. Hollow symbols denote data points, horizontal lines denote asymptotes, and "$*$" points joined by line segments denote regression functions. (a) Mean $\mathcal{A}$ as a function of replications added (taken from Figure 8.9). (b) Mean $d'$ as a function of replications added (taken from Figure 8.11).

definite ceiling effect.[15]

The steepest slope for Observer 2 in Figure 8.9(b) was for 12 dB, showing that the highest FORA in Figure 8.9(a) had the greatest relative curvature. The FORA only seemed very flat in Figure 8.9(a) because of the scale on which it was plotted. This FORA, based on $\mathcal{A}$, and the associated FORA based on $d'$, are re-presented in detail in Figure 8.13. Data points on the FORA based on $\mathcal{A}$ were approximately fitted by the regression equation, but the fit was worse than most of the other FORAs based on $\mathcal{A}$ that have been shown up to this point. In contrast, the FORA based on $d'$ (Figure 8.13(b)) showed a much better regression fit.

Together, Figures 8.13(a) and 8.13(b) demonstrate a ceiling effect, which is related to the fact that $\mathcal{A}$, as a measure, is bounded above at unity. The ceiling effect is small, but demonstrable. The estimated asymptotic value of $\mathcal{A}$ was 0.9985, and the asymptotic value of $d'$ was 4.3555. The $d'$-equivalent of 0.9985 is 4.2063, however, which is slightly less than 4.3555. Although the FORA based on $\mathcal{A}$ was somewhat distorted, the FORA based on $d'$ showed that there was a relatively smooth pattern underlying the data.[16] The empirical log-log plot based on $d'$ (open circles in Figure 8.11(b)) fell almost straight onto the FORA's equivalent linear function, apart from the last point, and was much closer to the data than the related log-log plot based on $\mathcal{A}$ (open circles in Figure 8.9(b)).

These results show that FORA regression can be sensibly applied to data associated with very high performance levels. The regular form of the regression function may not describe FORAs based on $\mathcal{A}$, because the data could be distorted by ceiling effects. Use of an unbounded measure such as $d'$ can correct for ceiling effects, and sensible prediction of asymptotic performance levels for $d' > 4$ is in fact possible.

### 8.3.6 Psychometric functions

A psychometric function shows how performance changes as a function of a stimulus parameter, such as SNR. Psychometric functions are often used to compare performance across individuals, groups or experimental conditions. Such functions are usually calculated based on single-replication experiments only, and the confounding effects of unique noise on performance are incorporated within the results. ACA and FORA regression, however, allow comparisons to be made based on estimated unique-noise-free performance.

Psychometric functions for both observers in the current experiment are presented in Figures 8.14 and 8.15. Figure 8.14 shows $\mathcal{A}$ as a function of SNR, where the SNR

[15] There is a hint of a small ceiling effect for Observer 1 at 12 dB because of the similarity of his log-log plot (open circles in Figure 8.10(b)) with that of Observer 2 for the same SNR (open circles in Figure 8.11(b)). In contrast to ceiling effects, it is hard to show any floor effects with FORAs because FORAs tend to increase away from chance performance rather than tend towards it.

[16] There is no *necessary* reason why the estimated asymptotic value of $\mathcal{A}$ should be less than or equal to 1, even though the data must be. The fact that the FORA for $\mathcal{A}$ does converge to a sensible value is encouraging. Examples of asymptotes converging *above 1* are given in Section 8.4.2.

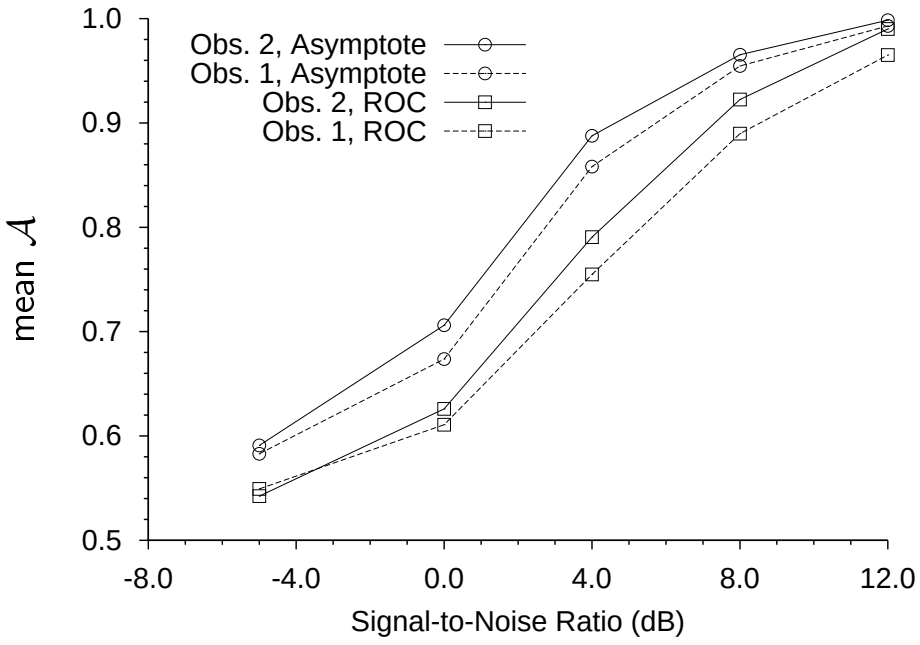


FIGURE 8.14: Single-replication and asymptotic psychometric functions for each observer, showing $\mathcal{A}$ as a function of signal-to-noise ratio in dB. Dashed lines are for Observer 1 and solid lines are for Observer 2. For each observer, the lower function shows average ROC performance, while the upper function shows asymptotic performance.

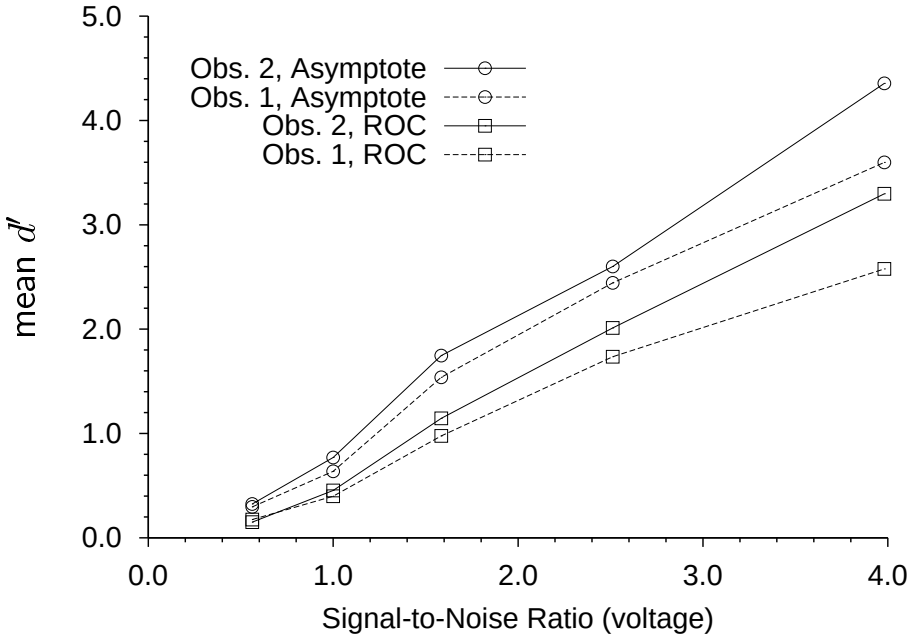


FIGURE 8.15: Single-replication and asymptotic psychometric functions for each observer, showing $d'$ as a function of signal-to-noise ratio (expressed as a voltage ratio). Dashed lines are for Observer 1 and solid lines are for Observer 2. For each observer, the lower function shows average ROC performance, while the upper function shows asymptotic performance.

is expressed in dB, and Figure 8.15 shows $d'$ as a function of SNR, where the SNR is expressed as a voltage ratio. Large potential improvements in performance at the highest SNR can be seen using $d'$, which would otherwise remain hidden if using $\mathcal{A}$. In each graph, dashed lines are for Observer 1 and solid lines are for Observer 2. The lower of the two functions for each observer shows average ROC performance at each SNR, based on the first point of each FORA, while the upper function shows each individual's asymptotic performance at each SNR.

For each observer and for either measure, the asymptotic psychometric function was higher than the ROC psychometric function. This is expected since all FORAs were increasing functions. The psychometric functions show that Observer 2 was better than Observer 1 in the task, both asymptotically, and in terms of ROC performance (except for ROC results at $-5$ dB). Figure 8.14 suggests that the ROC psychometric functions were approximately the same shape as the asymptotic psychometric functions, but were shifted horizontally by approximately 2-3 dB. The psychometric functions in Figure 8.15) showed that $d'$ was approximately linear with SNR, expressed as a voltage ratio, and that the effect of unique noise was to decrease the slope of the psychometric functions for each observer. This is roughly consistent with a constant attenuation, in dB, which would equate to a multiplicative change in voltage ratio. When converted to power ratios, attenuations of 2-3 dB imply that the effect of unique noise was similar to having stimulus maskers with 60%-100% more power than the maskers that were used. This indicates that the amount of unique noise was approximately the same as the amount of common noise, which is consistent with what has been found elsewhere (Swets et al., 1959; Watson, 1963; Green, 1964; Ahumada et al., 1975; Spiegel & Green, 1981; Siegel & Colburn, 1989).

Lapsley Miller (1999) made use of asymptotic psychometric functions to evaluate human amplitude discrimination (also using Gaussian noise signals). In her experiments (described in Section 8.4.2), Lapsley Miller found that single replication performance was typically attenuated from asymptotic performance by 2.5-4.0 dB, and attenuation was generally unaffected by stimulus duration. As power ratios, these attenuations imply that the effect of unique noise was similar to having stimulus maskers with 80%-150% more power than the maskers that were used, which is also consistent with previous measurements of unique noise.

**Summary.** Psychometric functions can be used to show how performance changes as a function of SNR, both with and without the effects of unique noise. Experimental data indicates that the effect of unique noise on aural amplitude discrimination is to attenuate performance by the equivalent of 2-4 dB in SNR. Psychometric functions derived from FORAs may also provide an extra tool for evaluating theories and models of observers, because any theory or model should account for both asymptotic performance and single-replication performance.

## 8.4 FORAs in Lapsley Miller's (1999) experimentation

Apart from Taylor's (1984) experiments with pigeons, and what is presented in this thesis, the only other study that has used ACA to calculate FORAs is by Lapsley Miller (1999). Her Ph.D. thesis investigated the joint role played by bandwidth, $\mathcal{W}$, and duration, $\mathcal{T}$, in amplitude discrimination in human hearing. She made extensive use of ACA and the FORA regression procedure to estimate unique-noise-free performance from her data sets. The aim of her project was to investigate hearing, and FORAs provided one method for doing this. A verbal summary of her results is given here, with an emphasis placed on data patterns rather than on their implications for amplitude discrimination. No graphs are presented.

### 8.4.1 Method

**Observers.** Three observers took part in the experiments. Only results for the first two observers are reported here, because more extensive data sets were available for them. Observer 1 was an experienced observer, and had taken part in Lapsley Miller et al.'s (1998) experiments described in Sections 8.1 and 8.2. Observer 2 was a novice observer. Observers ran more than 30 000 practice trials each to familiarise themselves with the variety of stimulus conditions. The stimulus sets used for practice were different from those used for the main data collection.

**Stimuli.** An extensive series of SIFC experiments was run, involving 18 different experimental conditions. Each condition had signals based on a different pair of $\mathcal{W}$ and $\mathcal{T}$ values,[17] and each condition was an experiment in its own right. $\mathcal{W}$ ranged from 2.5 Hz to 160 Hz in octave steps, and $\mathcal{T}$ ranged from 400 ms to 6.25 ms, also in octave steps. $\mathcal{W}$ and $\mathcal{T}$ were chosen so that the product $\mathcal{WT}$ took on values of 1, 2 and 4. There were seven conditions for $\mathcal{WT} = 1$, six conditions for $\mathcal{WT} = 2$, and five conditions for $\mathcal{WT} = 4$, making up 18 conditions in all. Stimuli in each condition were presented at five different SNRs, which ranged from 0 dB to 16 dB for $\mathcal{WT} = 1$, from $-4$ dB to 12 dB for $\mathcal{WT} = 2$ and from $-8$ dB to 8 dB for $\mathcal{WT} = 4$, all in 4 dB steps for each $\mathcal{WT}$. Detectability increases as $\mathcal{WT}$ increases, so different SNRs were used at each $\mathcal{WT}$ to ensure that the performance range was similar for each $\mathcal{WT}$.

The experimental system was the same described for the experiment in Section 8.3. The task of an observer was to detect a Gaussian noise signal in the presence of a Gaussian noise

[17] Lapsley Miller used the essential-92.4% bandwidth and essential-92.4% duration measures to specify $\mathcal{W}$ and $\mathcal{T}$. These measures describe the interval required to constrain 92.4% of a transient's energy in the frequency domain and time domain, respectively (Landau and Pollack, 1961; cited in Lapsley Miller, 1999). Algorithms for calculating these measures in discrete time are given in her Appendix B. The Kaiser window that was used in the experiments had a shape parameter of 9, for which the equivalent rectangular duration was equal to $\mathcal{T}$, and the absolute duration was equal to $2.43 \times \mathcal{T}$.

masker. Signals and maskers were gated together and presented diotically. The spectrum level of the Noise-alone transients was 60 dB SPL. During experimental sessions, an 8 kHz low-pass analog Gaussian noise masker ran continuously at a spectrum level 20 dB SPL.

**Experimental design.** Each observer ran six replications at each of 18 experimental conditions. Each replication of each condition consisted of the presentation of 3000 trials, 500 *SN* trials for each of the five SNRs, plus 500 *N* trials. Over six replications, each observer completed 18 000 trials per condition. Stimuli were presented in a different haphazard order for each replication and for each observer. For each replication, and across SNRs, trials were run in eight sessions of 375 trials for $\mathcal{WT} = 1$ and 10 sessions of 300 trials for $\mathcal{WT} = 2$ and for $\mathcal{WT} = 4$. All of the data for $\mathcal{WT} = 1$ was collected first, then the data for $\mathcal{WT} = 2$, and finally the data for $\mathcal{WT} = 4$. Only one condition was run per session, and a haphazard order of conditions was used across sessions. SNRs were intermixed within sessions (Tucker et al., 1967; Emmerich, 1968b). Within a session, no constraint was made on the number of trials per SNR, nor on the number of trials per event. The trial sequence was run-limited so that the same event *and* SNR could not occur more than 3, 4 or 5 times in a row (where 3, 4 or 5 was randomly chosen with equal probability on each trial). A series of short preview trials were run at the start of each session to remind observers of the type of stimuli for the current session. The stimuli used in the preview trials were not used for the main data collection.

Lapsley Miller used the same continuous rating scale methodology as described for other experiments in this thesis. The position of the rating slider was measured electronically and converted into a 2000-point rating scale. Each SIFC trial consisted of a brief warning interval ($\simeq$ 100 ms), an observation interval whose length was determined by the absolute duration of the stimuli, a decision interval of 1250 ms and a minimum reset interval of 300 ms. A set of LED lights on the decision panel were switched on and off to mark the trial intervals. No trial-by-trial knowledge of results was given, but observers viewed their single-session ROC curves (for five SNRs) at the end of each session, and viewed single-replication ROC curves after each completed replication.

### 8.4.2 Results

As well as ROC and GOC curves, Lapsley Miller presented FORAs, log-log plots and asymptotes for all of the conditions, for each observer and for different measures of sensitivity. GOC analysis was only performed within observers, and not across observers. FORA results were presented for each observer, condition, SNR, and over four measures of sensitivity, $\mathcal{A}$, $d'$, $\mathcal{D}_2$ and $\mathcal{D}_6$. The focus here is mainly on the results for $\mathcal{A}$ and $\mathcal{D}_6$, although some results based on $d'$ and $\mathcal{D}_2$ are also described.

### FORAs based on two-event measures

The results fell over a full range of possible performance, from near-chance to near-perfect, depending on the particular condition and SNR. For each of the 18 conditions, and for each observer, asymptotic values of $\mathcal{A}$ lay between 0.55 and 0.65 at the lowest SNR (with FORA data lying below the asymptote), and between 0.95 and 1.00 at the highest SNR. The sets of FORAs for each condition were similar in form to the FORAs presented in Figures 8.8 to 8.11 in the previous section, except that they were based on six replications.

With five SNRs per condition, and 18 conditions, there were 90 FORAs per observer for a given measure of sensitivity. The number of performance values that were averaged to calculate each point on a FORA was small compared to numbers used in other experiments. Because only six replications were run, ${}^6C_\xi$ was either 1, 6, 15 or 20 for $\xi = 1 \ldots 6$. Nevertheless, mean FORAs were generally stable, and regression-FORAs and log-log plots could be estimated for most of the data sets.

Within each condition, the log-log plots became steeper, and $\mu$ decreased with increasing SNR, for FORAs based on $\mathcal{A}$ and $\mathcal{D}_2$. The slope was more homogeneous for FORAs based on $d'$, and did not vary systematically across SNRs. A similar pattern occurred in the previous section (Figures 8.8 to 8.11), and occurred because $\mathcal{A}$ and $\mathcal{D}_2$ are bounded in nature, whereas $d'$ is unbounded.

### Relative reversal of observer order based on ROC and asymptotic performance

Both observers performed comparably across most conditions. The observer with the higher single-replication value of $\mathcal{A}$ usually had the higher asymptotic value as well, but this was *often not the case*. In 20 out of the 90 condition-and-SNR pairings, the observer with the smaller initial value of $\mathcal{A}$ had the higher asymptotic value. Relative-performance reversal was scattered among the many conditions and SNRs without an obvious pattern, except for the $\mathcal{WT} = 2$, 400 ms, 5 Hz condition. In this condition, the initial $\mathcal{A}$ values for Observer 2 were better than those for Observer 1 at all SNRs, whereas the opposite was true for their asymptotes.

### Summary of correlations

Tables E.1 to E.6 in Lapsley Miller (1999, Appendix E) list each observer's six-replication FORA results based on $\mathcal{A}$, $d'$ and $\mathcal{D}_2$, including correlations ($r$) for all of the associated 5-point log-log plots. Values of $r$ for log-log plots based on $\mathcal{A}$ have been converted to $r^2$, and are summarised here.

Out of the 90 FORAs for Observer 1, there were 72 FORAs with $r^2$ greater than or equal to 0.9980, 14 with $r^2$ between 0.9950 and 0.9980, two with $r^2$ between 0.9800 and 0.9950, and two relatively stray FORAs, one with $r^2 = 0.9185$ and one with $r^2 = 0.1950$. For Observer 2, there were 75 FORAs with $r^2$ greater than or equal to 0.9980, 14 with

$r^2$ between 0.9950 and 0.9980, and one with $r^2 = 0.9850$. Since $r^2$ on a log-log plot indicates how well FORA data is fitted by Equation 6.5, these results show that the FORA regression was extremely robust across observers, parameter values ($\mathcal{W}$, $\mathcal{T}$, and SNR) and performance levels (without regard to stimulus parameters).

**Non-converging and over-converging FORAs**

The vast majority of FORAs shown in this and previous chapters, and in Lapsley Miller (1999), can be successfully fitted by Equation 6.5 and extrapolated to estimate unique-noise-free performance. There was a small set of Lapsley Miller's (1999) data, however, for which FORA extrapolation broke down.

There is no necessary reason why an asymptotic value of $\mathcal{A}$ (or of $\mathcal{D}_2$) should be less than or equal to 1, even though the empirical FORA must be. Some of the FORAs in Lapsley Miller's data either did not converge, because the $\mu$ parameter was greater than $-1$, or over-converged to impossible values such as $\mathcal{A}_\infty > 1$ or $(\mathcal{D}_2)_\infty > 1$. FORAs based on $d'$ cannot over-converge because $d'$ is unbounded above, so they either converge to a finite value or they do not converge at all.

Sensible convergence in Lapsley Miller's data was by far the rule, but there were exceptions. Of the 180 FORAs based on $\mathcal{A}$ from either observer, there was only one that did not converge at all, and one that over-converged beyond $\mathcal{A} = 1$. The latter was for Observer 1 at $\mathcal{WT} = 4$, $\mathcal{T} = 400$ ms, $\mathcal{W} = 10$ Hz, and 8 dB (the highest SNR). The regression-FORA based on $\mathcal{A}$ started at 0.8129 and over-converged to an asymptote at $\mathcal{A} = 1.0083$. In spite of this, the regression was very reasonable, because $r^2$ for the log-log plot was 0.9994. For the same data analysed using $d'$, the regression-FORA started at $d' = 1.2606$ and converged to an asymptote at $d' = 5.6978$ ($r^2$ was also 0.9994 based on $d'$). The results for this condition may be contrasted with results for a similar condition, with a similar pattern of results, but which did converge sensibly. For Observer 1 at $\mathcal{WT} = 2$, $\mathcal{T} = 400$ ms, $\mathcal{W} = 5$ Hz, and 8 dB (which only differed to the over-converging condition in the value of $\mathcal{W}$), the regression-FORA based on $\mathcal{A}$ started at 0.8351 and converged sensibly to an asymptote at $\mathcal{A} = 0.9999$, with $r^2 = 0.9988$. The regression-FORA based on $d'$ for the same data started at $d' = 1.3809$ and converged to $d' = 4.8519$, with $r^2 = 0.9982$. If the underlying level of performance is very high (say $d' \simeq 5$), there may be only a very fine line between over-convergence and sensible convergence for FORAs based on $\mathcal{A}$.

Out of all 180 regression-FORAs based on $\mathcal{A}$, the regression-FORA in only one condition did not converge (sensibly or otherwise). This was for Observer 1 at $\mathcal{WT} = 4$, $\mathcal{T} = 400$ ms, $\mathcal{W} = 10$ Hz, and $-8$ dB (the lowest SNR).[18] The FORA based on $\mathcal{A}$ started off at $\mathcal{A} = 0.5067$, but it was extremely flat and hardly increased over six replications. It could not converge because $\mu = -0.3432$ (which was was greater than $-1$). This data

[18] These particular experimental parameters (including SNR) were the only ones for which FORA regression broke down so badly that *none* of the FORAs based on either $\mathcal{A}$, $d'$ or $\mathcal{D}_2$ converged.

resulted in a log-log plot with $r^2 = 0.1950$. The FORA based on $d'$ suffered the same problem as that for $\mathcal{A}$, with virtually the same $\mu$ and $r^2$ values.

These examples of poor convergence show two extremes that were part of a pattern for Observer 1. The most problematic results of the entire data set occurred for Observer 1 at $\mathcal{T} = 400$ ms, regardless of the bandwidth, SNR or measure of sensitivity. In contrast, results for Observer 2 when $\mathcal{T} = 400$ ms were sensible for all bandwidths, SNRs and sensitivity measures. FORAs for Observer 1 at $\mathcal{T} = 400$ ms at the higher SNRs showed appreciable improvement from one to six replications,[19] more so than for conditions based on shorter durations. This indicated that there was a large amount of unique noise affecting Observer 1 at the longest duration.

Observer 1 undoubtably had very high levels of unique noise, because the empirical GOC curve improved appreciably based on only six replications. Whether the underlying asymptotic values of $\mathcal{A}$ and $d'$ were in fact so high is inconclusive. The log-log plots were shallow and the data series showed signs of curving downwards. If so, the empirical FORA may have coverged on a lower sensitivity value if the data series was extended further. Running further replications probably would have helped to stabilise the FORA results for this observer and stimulus duration. FORA regression functions for this experiment generally provided such good fits to data, that it may be easy to forget that only six replications were run.

### FORAs based on the six-event measure, $\mathcal{D}_6$

Scurfield (Scurfield, 1995, 1996, 1998) extended the Theory of Signal Detectability to $n$-event discrimination tasks. He developed $n$-event ROC analysis within the context of information theory, and proposed using a new measure, $\mathcal{D}_n$, as a measure of overall performance in an $n$-event task. $\mathcal{D}_n$ specifies the amount of information about event-orderings that is contained in an observer's decisions. Like other measures of sensitivity, $\mathcal{D}_n$ is based on decision axis values when used in a theoretical context, and on ratings or mean-ratings when used in a practical context. The measure $\mathcal{D}_2$ in two-event discrimination tasks is a specific case of $\mathcal{D}_n$. In an $n$-event task, $\mathcal{D}_n$ can range between zero and $\log_b(n!)$, where the base $b$ defines the unit of information.

Lapsley Miller's (1999) experiments involved a two-event task, and were analysed using two-event ROC and GOC analyses and two-event measures of performance for each SNR, observer and experimental condition. The data could also be interpreted and analysed in terms of a six-event discrimination task (one $N$ event plus five $SN$-events, per observer and per condition). The *six*-event measure, $\mathcal{D}_6$, may be used to evaluate overall performance for each observer and condition, taken *across all SNRs*. There are two theoretical assumptions required in order for this analysis to hold, (1) that all evidence distributions (one per event)

[19] The value of $\mathcal{A}$ improved by 0.10–0.15 for the three highest SNRs.

fall onto the same unidimensional decision axis, and (2) that there is a strictly monotonic increasing transform between the decision axis (with six distributions on it) and the rating scale. That is to say, a psychophysical transfer function must be assumed. Given these assumptions, which also underlie two-event ROC analysis, then Lapsley Miller's analysis based on $\mathcal{D}_6$ holds.[20]

Experimental results based on $\mathcal{D}_6$ were affected by unique noise in much the same way as results based on $\mathcal{A}$, namely, performance was depressed and there was variability across replications. For a given set of replications, the sum of ratings was calculated per stimulus, and used as the basis for calculating $\mathcal{D}_6$. The number of replications per set was anywhere from from one to six, inclusive, depending on which combination of replications was analysed. A value of $\mathcal{D}_6$ was calculated for each combination, based on all six signal levels (including noise-alone), and the average $\mathcal{D}_6$ value at each combination-size defined a FORA. Data values were only combined within observers and not across observers.

Each of the FORAs based on $\mathcal{D}_6$ had a regression function fitted to it using Equation 6.5, and asymptotic $\mathcal{D}_6$ values were calculated. Across all observers and conditions, all but two of the 36 FORAs converged to sensible values, and those that did not were for Observer 1 when $\mathcal{T} = 400$ ms. Single-replication $\mathcal{D}_6$ values varied between 1.0 and 2.0 bits, depending on the condition and the observer. The asymptotic $\mathcal{D}_6$ values that converged sensibly ranged between 1.8 and 3.9 bits, and showed a definite improvement compared to their respective single-replication $\mathcal{D}_6$ values.

Comparisons between single-replication and asymptotic values of $\mathcal{D}_6$ showed that performance improved by 50–100% for Observer 1, and by 50–80% for Observer 2. This was fairly consistent across all conditions, and demonstrates the overall gains to be had from using GOC analysis across a wide range of stimulus parameters and performance levels. The effect of unique noise was to reduce by one third to one half the amount of information about event-ordering contained in observers' decisions. The information is lost in single-replication performance, but can be retrieved by applying GOC analysis to average out unique noise.

Across all conditions—even those that did not converge sensibly—FORA regression fits could hardly have been better. For Observer 1, seven of the 18 conditions had log-log plot correlation values[21] of $r = -1.0000$ (to four decimal places), another six conditions had $r = -0.9999$, and the other five conditions had $r$ between $-0.9966$ and $-0.9997$. For Observer 2, twelve of the 18 conditions had $r = -1.0000$, five conditions had $r = -0.9999$, and one condition had $r = -0.9998$.

The stability of log-log plots based on $\mathcal{D}_6$ was due, in part, to the large number of trials per condition (18 000 trials per FORA, with 3000 trials per replication). The use of $\mathcal{D}_6$ meant all of this data could be sensibly contribute towards a single overall measure

[20] Using base 2 logarithms, $\mathcal{D}_6$ can vary between 0 and 9.49 bits. The subscript "6" in $\mathcal{D}_6$ refers to the number of events and not to the number of replications run, which also happens to be 6.

[21] Original $r$-values are reported here, rather than $r^2$.

of performance. The very high correlations may also partly reflect the fact that only six replications were run. If the data series was extended further, say to 30 replications, it is not known if the log-log plots would curve or not.

There were problems with the FORAs based on $\mathcal{D}_6$ for Observer 1 at $\mathcal{T} = 400$ ms, which mirrored problems with two-event FORAs for the same observer at the same stimulus duration. $\mathcal{D}_6$ FORAs for two conditions with this duration converged, but not to sensible values. The $\mathcal{D}_6$ asymptotes for $\mathcal{W} = 5$ Hz and $\mathcal{W} = 10$ Hz were 11.1559 bits and 30.3367 bits respectively, which are nonsensical because the maximum possible value of $\mathcal{D}_6$ is 9.4919 bits. Like the two-event FORA results for Observer 1 for these experimental conditions, the $\mathcal{D}_6$ FORAs and log-log plots were all very shallow, and the log-log plots showed signs of curving downwards, even over five data points. This indicated that the regression-FORA would overestimate the asymptote, and that more replications may have been needed to obtain sensible results.

### 8.4.3 Summary

Lapsley Miller's (1999) amplitude discrimination experiments involved two observers running six replications each, in 18 experimental conditions, at five SNRs per condition. FORA regression was found to be extremely robust over 180 different combinations of observers, conditions and SNRs, in which performance ranged from close to chance to near-perfect. Log-log plots for two-event FORAs based on $\mathcal{A}$ typically had $r^2$ values greater than 0.995, and six-event FORAs based on $\mathcal{D}_6$ typically had $r^2$ values greater than 0.999. The experiments showed that it was possible in practice to estimate asymptotic performance and asymptotic psychometric functions from as few as six replications. The results also showed that a comparison of observers based on ROC performance *often* differs from a comparison based on estimated unique-noise-free performance. Relative-performance reversal across observers occurred in 22% of pairings of SNR and condition. The result is a consequence of individual differences in common noise and unique noise, and sampling variability of replications due to having a finite data set.

## 8.5 Summary of Chapter

The aim of the chapter was to show ACA and FORA regression for a variety of different experiments. These, and the FORA results from the preceding chapters, demonstrate the pervasive and robust nature of the FORA data pattern characterised by Equation 6.5. The pattern held across observers, experimental methodologies, types of stimuli, measures of performance, and levels of performance. All of the FORAs were increasing functions, which reflected the improvement in GOC performance as more replications were added.

In Section 8.1, FORA regression worked for both binary-decision and rating methodologies, and the estimated asymptotes were similar in both cases. Section 8.2 showed that the pattern held for different observers who performed at different levels. Section 8.2 also clearly showed that comparisons among observers within a group depended on what what is compared: relative single-replication performance does not necessarily translate into relative unique-noise-free performance. This type of result also occurred in Whitmore et al.'s (1993) experiment in Section 7.3.1, in the unpublished experiment in Section 8.3, and throughout Lapsley Miller's (1999) project. Performance reversal may sometimes occur due to sampling variability, for example, but some of the very firm patterns, such as those shown in Figure 8.6, argue against that as the sole reason.

Section 8.3 showed ACA and FORA results for an experiment that used multiple SNRs. The FORA data pattern was found to hold over a wide range of performance levels, including very high levels ($d' > 4$). Many similar experiments were run by Lapsley Miller (1999), which were described in Section 8.4. Estimated asymptotes in multiple-SNR experiments may be used to specify asymptotic, unique-noise-free psychometric functions. These are of interest in psychophysics, since they provide details about the potential performance of sensory systems that are unavailable through conventional ROC analysis.

# Chapter 9

# Summary and Conclusions

Human beings are inconsistent decision makers in discrimination tasks. Psychophysical results change from replication to replication of an experiment, even when identical stimulus sets are used across replications. This contributes to error in the task and decreases performance. There are practical implications of this decrease, namely the consequences of errors in any task. There are also theoretical implications, because theories tested against single-replication results will be in error to the extent that the replication is in error. Most psychophysical experiments employ only a single presentation of a stimulus set, and therefore are not designed to account for inconsistent decisions. Their results and findings would change if they did.

Studies that incorporate inconsistency, usually under the heading of internal noise, are often concerned with modelling the error and quantifying it, particularly in terms of a ratio of unique-to-common noise variances, $k$. Relatively little emphasis has been placed on removing the error. There are at least two experimental procedures that will reduce the error, multiple-presentation experiments and multiple-replication experiments. Both methods involve repeated presentations of a given stimulus. Multiple-presentation experiments, such as Swets et al. (1959), assume internal averaging of unique noise, prior to each decision, whereas multiple-replication experiments, such as Taylor et al. (1991), do not. Rather, decisions are made after each replication and averaged after the fact by the experimenter. Neither of these experimental designs are widely used, although they both address practical and theoretical problems.

This thesis is primarily concerned with the removal of unique noise resulting from observer inconsistency in multiple-replication experiments. The data in such experiments can be understood in the context of the Theory of Signal Detectability by the use of receiver operating characteristic analysis. There are two basic effects of observer inconsistency, (1) each replication of a multiple-replication experiment results in a different ROC curve, and a different performance value, and (2) the average performance level is lower compared to what it would be without inconsistency. It is possible to remove the extra error by

averaging across replications, but what is being averaged is crucial to the results. Mean ROC analysis involves averaging ROC curves, whereas GOC analysis involves averaging ratings or decisions on a per stimulus basis. A mean ROC curve indicates expected single-replication ROC performance, whereas an asymptotic GOC curve indicates unique-noise-free performance. It was seen throughout this thesis that mean ROC analysis does not remove the effects of unique noise, but rather it incorporates them.

**Part I**

The basic question that motivated the developments in Part I was "why does GOC analysis work?" Empirical transform-average GOC analysis, reported in Chapter 3, showed that GOC analysis worked under a large variety of transforms. *All* of the transforms applied in Chapter 3 resulted in GOC curves that were better than the mean ROC curve, and some were often as good as the sums-of-integer-ratings GOC curve. However GOC analysis worked, it seemed to be robust with respect to order-preserving transforms, even the more extreme ones.

The topic of strictly monotonic increasing (s.m.i.) transforms also arose in the theoretical relationship between a decision axis and a rating scale, namely, the transfer function. The question of how transform-average GOC analysis works may be rephrased as a question of how GOC analysis could work at all under any arbitrary ordinal scaling.

In a multiple-replication experiment, each stimulus is associated with a distribution of ratings, defined across replications. GOC analysis averages ratings on a per stimulus basis, and a GOC curve results from the ordering of a stimulus set according to mean rating. What is required to explain GOC analysis is a general statistical property under which the order of a stimulus set remains unaffected by s.m.i. transforms of a rating scale or a decision axis. The key statistical property that achieves the desired result is *stochastic ordering*. The distribution of ratings per stimulus extends to a family of distributions when an entire stimulus set is considered. If the family is stochastically ordered, then the order of mean ratings will always be the same for any strictly monotonic increasing transform. Since mean ratings form the basis of GOC analysis, then this suggests that GOC analysis will work under any arbitrary ordinal transform, including transfer functions that transform the decision axis into the rating scale. Assuming that stochastic ordering holds on the decision axis, then GOC analysis will work. If the family is not stochastically ordered, then GOC analysis may work, but it is specific to a given rating scale, transfer function and decision axis. With stochastic ordering, any specific scaling of an underlying decision axis or a rating scale does not need to be known in order for GOC analysis to work. The transfer function does not need to be known; the only requirement, in theory, is that the transfer function is strictly monotonic increasing.

An implication of the theory of GOC analysis is that there is no inherent or special scaling of a rating scale in a discrimination task. Rating distributions do not need to

approximate theoretical distributions in order for GOC analysis to work. No inherent meaning of rating categories is necessary for an ordinal rating scale to be useful in GOC analysis or discrimination tasks in general.

Many models of inconsistent observers implicitly or explicitly average unique-noise-affected values on a decision axis. The effect of scaling on the order of stimuli, according to mean value, is as much a problem on a decision axis as it is on a rating scale. Hence stochastic ordering is as relevant to such models as it is to GOC analysis. One of the findings of Chapter 5 was that if stochastic ordering does not hold, then the removal of unique noise on one decision axis is not the same the removal of unique noise on a second (s.m.i.-transformed) decision axis. If averaging unique noise on a decision axis is viewed as recovering a theoretical ROC curve on a decision axis, then the so-called theoretical curve may change under s.m.i. transforms of the decision axis when stochastic ordering does not hold.

Arbitrary ordinal scales have implications for models of unique noise. Unique noise is often modelled as having the same form for all common noise values on a decision axis. If unique noise is of the same form along one decision axis, however, then it will generally not be so after a non-linear s.m.i. transform of the decision axis. This may not matter if the focus of a theory is on a specific decision axis, but if a theory or model is used to model a data set of ratings, then arbitrary ordinal scales and axes are a problem for analyses that assume a particular scale, or axis.

The potentially arbitrary nature of ordinal scales, particularly decision axes, raises questions about measures that are scale-specific. The unique-to-common noise variance ratio, $k$, is one such measure. It is often used to specify or characterise observer inconsistency. Decision axes are arbitrary, in the sense that an unlimited number of different, s.m.i.-related axes produce the same ROC and GOC data. Therefore, $k$ calculated on one decision axis can be different from $k$ calculated on another decision axis, where each axis could account for, or result in, the same rating data.

The two main contributions of this thesis are *the theory of GOC analysis*, and the *FORA regression* procedure. There seems to be nothing in the theory of GOC analysis, however, to suggest that the FORA data pattern should emerge from GOC analysis, or from models of unique-noise-affected observers. FORAs do not follow in any obvious way from the theory of GOC analysis. Theoretical FORAs were noted in Chapter 6, but were not discussed. Some theoretical FORAs are of a similar form to the ones described here, whereas others are not. The theoretical FORAs are based on specific models, and are a starting point for a general theory. The main theoretical problem that remains unsolved is the theoretical basis for FORA regression, and its relationship to the theory of GOC analysis.

**Part II**

Data from a multiple-replication experiment can be combined in GOC analysis. As more replications are added together, unique noise is removed and GOC curves tends to an asymptotic form. As a consequence, a measure of performance tends toward an asymptotic value as a function of replications added (FORA). Replications may be combined in any order, and each order results in a different sample-FORA. Taylor (1984) used all combinations analysis (ACA) to reduce the variability and error associated with sample-FORAs. For a given data set of $m$ replications, the GOC curve can be calculated for each subset of $\xi$ replications, and a measure of performance can be calculated based on each GOC curve. The average performance value for all combinations of size $\xi$, plotted as a function of $\xi$, defines an average FORA (or just *FORA*).

FORAs generally follow the same form. Logarithms of FORA increments are approximately linear functions of the logarithm of the number of replications, which is seen when the increments are graphed as a *log-log plot*. A linear log-log plot implies a summed power series of the form

$$A_i \quad = \quad A_1 + \kappa \sum_{j=2}^{i} j^{\mu}, \qquad i \geqslant 2,$$

(Equation 6.5), where $A_i$ represents performance after $i$ replications are combined, $A_1$ denotes the initial value of the FORA, and $\kappa$ and $\mu$ determine how performance improves as as the number of replications in the data set increases. A linear regression to a log-log plot does not provide the best regression to a FORA. A non-linear least-squares regression procedure was derived instead, which fit Equation 6.5 to a FORA data series, rather than to a log-log plot. The regression function can be extrapolated to an infinite number of replications to obtain an estimate of asymptotic, unique-noise-free performance.

The FORA regression function is a three-parameter data model, and is not a theory of performance in multiple-replication experiments. Parameter values derive from data for any given FORA. The data model is non-parametric in the statistical sense, because it does not assume any distributions. FORA regression requires that a minimum of three replications are run. The resulting regression function may be extrapolated to any number of replications, and in the limit, it estimates asymptotic, unique-noise-free performance.

FORAs were plotted or described for a variety of different experiments, tasks, decision methodology, measures of performance, stimulus parameters, levels of performance, and individual observers. Linear log-log plots occurred in all of the experiments across all of these factors and showed that Equation 6.5 describes an extremely robust data pattern. The following results were found:

- The log-log plot was either linear or near linear for most data sets, with $r^2 > 0.99$ being the norm.
- FORA regression works over a very wide range of performance levels, including very high levels such as $d' > 4$.
- FORA regression works across observers as well as within observers.
- The ranking of members of a group according to their performance in a task can be affected by unique noise. Rankings based on ROC performance, GOC performance, and asymptotic performance do not necessarily agree with one another.
- FORA analysis showed that while GOC analysis removes unique noise effects from both continuous rating scale data and binary-decision data, the removal is more efficient when the number of rating categories is large.
- FORAs based on $P(C)$ were more variable than FORAs based on $\mathcal{A}$, $d'$, $\mathcal{D}_2$, or $\mathcal{D}_6$. This probably reflected the fact that $P(C)$ was calculated from a single ROC point, whereas the other measures were based on the area[1] under an entire ROC curve, $\mathcal{A}$.
- Given a large enough data set, it is possible to calculate sample statistics of estimated asymptotes, and put error bounds on the asymptote.
- ACA need not be computed in full for very large data sets. It is possible to achieve reasonable asymptotic estimates using partial-ACA and sampled-ACA.
- Stable estimates of the asymptote are possible with 8–10 replications, although this will generally depend on experimental conditions.
- FORA regression can be used to estimate unique-noise-free psychometric functions.

[1] Or from volumes under ROC hypersurfaces, in the case of $\mathcal{D}_6$.

# References


Aho, A. V., Hopcroft, J. E., & Ullman, J. D. (1983) *Data Structures and Algorithms.* Addison-Wesley, Reading, Massachusetts.

Ahumada, Jr., A. (1967) *Detection of tones masked by noise: A comparison of human observers with digital–computer–simulated energy detectors of varying bandwidths.* Ph.D. thesis, University of California, Los Angeles.

Ahumada, Jr., A. & Lovell, J. (1971) Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*(6), 1751–1756.

Ahumada, Jr., A., Marken, R., & Sandusky, A. (1975) Time and frequency analyses of auditory signal detection. *Journal of the Acoustical Society of America*, *57*(2), 385–390.

Anderson, C. & Whittle, L. (1971) Physiological noise and the missing 6 db. *Acustica*, *24*, 261–272.

Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Bell, D. W. & Nixon, J. C. (1971) Reliability of ratings in an auditory signal-detection experiment. *Journal of the Acoustical Society of America*, *49*(2), 435–439.

Berg, B. G. (1987) *Internal noise in auditory decision tasks.* Ph.D. thesis, Indiana University, Bloomington.

Berg, B. G. (1989) Analysis of weights in multiple observation tasks. *Journal of the Acoustical Society of America*, *86*(5), 1743–1746.

Berg, B. G. (1990) Observer efficiency and weights in a multiple observation task. *Journal of the Acoustical Society of America*, *88*(1), 149–158.

Borowski, E. J. & Borwein, J. M. (1989) *Dictionary of Mathematics.* Collins Reference. Collins, Glasgow.

Boven, R. (1976) *The use of multiple observers in signal detection theory: A method to remove the effect of unique noise from experimental data.* Master's thesis, Victoria University of Wellington, Wellington, New Zealand.

Burkill, J. & Burkill, H. (1970) *A Second Course in Mathematical Analysis.* Cambridge University Press, Cambridge.

Cargo, G. (1965) Comparable means and generalized convexity. *Journal of Mathematical Analysis and Applications*, *12*, 387–392.

Cargo, G. & Shisha, O. (1969) A metric space connected with generalized means. *Journal of Approximation Theory*, *2*, 207–222.

Clarke, F. R., Birdsall, T. G., & Tanner, Jr., W. P. (1959) Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America*, *31*, 629–630.

Clarke, L. E. (1975) *Random Variables*. Longman, New York.

Courant, R. (1937) *Differential and integral calculus*, Vol. 1 & 2. Interscience: A division of John Wiley & Sons, London.

de Boer, E. (1966) Intensity discrimination of fluctuating signals. *Journal of the Acoustical Society of America*, *40*(3), 552–560.

Dorfman, D. D. & Berbaum, K. S. (1986) RSCORE–J: Pooled rating–method data: A computer program for analyzing pooled ROC curves. *Behavior Research Methods, Instruments, and Computers*, *18*(5), 452–462.

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

Durlach, N. I., Braida, L. D., & Ito, Y. (1986) Towards a model of discrimination of broadband signals. *Journal of the Acoustical Society of America*, *80*, 63–72.

Egan, J. P. (1975) *Signal Detection Theory and ROC Analysis*. Academic Press, New York.

Egan, J. P. & Clarke, F. R. (1966) Psychophysics and signal detection. In Sidowski, J. B. (Ed.), *Experimental methods and instrumentation in psychology*, chap. 5. McGraw Hill Book Company.

Egan, J. P., Schulman, A. I., & Greenberg, G. Z. (1959) Operating characteristics determined by binary decisions and by ratings. In Swets, J. A. (Ed.), *Signal detection and recognition by human observers: Contemporary readings* (1964 edition)., chap. 7. John Wiley & Sons, New York.

Elliot, P. B. (1964) Tables of $d'$. In Swets, J. A. (Ed.), *Signal detection and recognition by human observers: Contemporary readings* (1964 edition)., chap. Appendix 1. John Wiley & Sons, New York.

Emmerich, D. S. (1968a) Receiver–operating characteristics determined under several interaural conditions of listening. *Journal of the Acoustical Society of America*, *43*(2), 298–307.

Emmerich, D. S. (1968b) ROCs obtained with two signal intensities presented in random order, and a comparison between yes-no and rating ROCs. *Perception & Psychophysics*, *3*(1), 35–40.

Evans, II, G. W., Wallace, G. F., & Sutherland, G. L. (1967) *Simulation Using Digital Computers*, chap. B. 6 Pseudo–random number generators, pp. 187–189. Prentice–Hall, Englewood Cliffs, N.J.

Fatt, P. & Katz, B. (1950) Some observations on biological noise. *Nature*, *166*(4223), 597–598.

Findlay, M. C. & Whitmore, G. A. (Eds.). (1978) *Stochastic Dominance: An Approach to Decision Making Under Risk*. D. C. Heath, Lexington, Massachusetts.

Fishburn, P. C. & Vickson, R. G. (1978) Theoretical foundations of stochastic dominance. In Findlay, M. C. & Whitmore, G. A. (Eds.), *Stochastic Dominance: An Approach to Decision Making Under Risk* (1978 edition)., chap. 2. D. C. Heath, Lexington, Massachusetts.

Friedman, D. & Massaro, D. (1998) Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, *5*(3), 370–389.

Friedman, M. P. & Carterette, E. C. (1964) Detection of Markovian sequences of signals. *Journal of the Acoustical Society of America*, *36*(12), 2334–2339.

Galvin, S. J. (1988) *The Theory of Type II ROC Analysis*. Master's thesis, Victoria University of Wellington, Wellington, New Zealand.

Galvin, S. J., Podd, J., Drga, V., & Whitmore, J. K. (1998) *Extending the Theory of Signal Detectability to Discrimination Between Correct and Incorrect Decisions*. Submitted to the Journal of Mathematical Psychology.

Gibson, J. J. (1960) The concept of the stimulus in psychology. *The American Psychologist*, *11*, 694–703.

Gilkey, R. H. (1981) *Molecular psychophysics and models of auditory signal detectability*. Ph.D. thesis, Indiana University, Bloomington.

Gilkey, R. H. & Robinson, D. E. (1986) Models of auditory masking: A molecular psychophysical approach. *Journal of the Acoustical Society of America*, *79*(5), 1499–1510.

Gilkey, R. H., Robinson, D. E., & Hanna, T. E. (1985) Effects of masker waveform and signal-to-masker phase relation on diotic and dichotic masking by reproducible noise. *Journal of the Acoustical Society of America*, *78*(4), 1207–1219.

Gradshteyn, I. & Ryzhik, I. (1965) *Table of Integrals, Series, and Products*. Academic Press, New York.

Green, D. M. (1960a) Auditory detection of a noise signal. *Journal of the Acoustical Society of America*, *32*(1), 121–131.

Green, D. M. (1960b) Psychoacoustics and detection theory. *Journal of the Acoustical Society of America*, *32*(10), 1189–1203.

Green, D. M. (1964) Consistency of auditory detection judgements. *Psychological Review*, *71*(5), 392–407.

Green, D. M., Birdsall, T. G., & Tanner, Jr., W. P. (1957) Signal detection as a function of signal intensity and duration. *Journal of the Acoustical Society of America*, *29*(4), 523–531.

Green, D. M. & McGill, W. J. (1970) On the equivalence of detection probabilities and well-known statistical quantities. *Psychological Review*, *77*(4), 294–301.

Green, D. M. & Swets, J. A. (1974) *Signal Detection Theory and Psychophysics*. Robert E. Krieger Publishing Co., Huntington, New York.

Green, D. M. & Moses, F. L. (1966) On the equivalence of two recognition measures of short–term memory. *Psychological Bulletin*, *66*(3), 228–234.

Hanley, J. A. (1988) The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making*, *8*, 197–203.

Hanley, J. A. & McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Diagnostic Radiology*, *143*(1), 29–36.

Hautas, M. J. (1995) Corrections for extreme proportions and their biasing effects on estimated values of $d'$. *Behavior Research Methods, Instruments, and Computers*, *27*(1), 46–51.

Hsieh, F. & Turnbull, B. (1996) Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, *24*(1), 25–40.

Hutchinson, T. P. (1981) A review of some unusual applications of signal detection theory. *Quality and Quantity*, *15*, 71–98.

Isabelle, S. K. & Colburn, H. S. (1991) Detection of tones in reproducible narrow–band noise. *Journal of the Acoustical Society of America*, *89*(1), 352–359.

Jeffress, L. A. (1964) A stimulus–oriented approach to detection theory. *Journal of the Acoustical Society of America*, *36*, 766–774.

Jeffress, L. A. (1967) Stimulus–oriented approach to detection re-examined. *Journal of the Acoustical Society of America*, *41*(2), 480–488.

Jeffress, L. A. (1970) Masking. In Tobias, J. V. (Ed.), *Foundations of modern auditory theory*, Vol. 1, pp. 87–114. Academic Press, New York.

Jeffress, L. A. & Robinson, D. E. (1962) Binaural analysis as a function of physiological masking. *Journal of the Acoustical Society of America*, *34*(10), 1658–1659.

Jesteadt, W. & Sims, S. L. (1975) Decision processes in frequency discrimination. *Journal of the Acoustical Society of America*, *57*(5), 1161–1168.

Kroll, Y. & Levy, H. (1980) Stochastic dominance: A review and some new evidence. *Research in Finance*, *2*, 163–227.

Lakey, J. R. (1976) Temporal masking–level differences: The effect of mask duration. *Journal of the Acoustical Society of America*, *59*(6), 1434–1442.

Lapsley Miller, J. A. (1996) *Common Noise Sampling Variability.* Unpublished simulation results.

Lapsley Miller, J. A. (1999) *The role of the bandwidth–duration product $\mathcal{WT}$ in the detectability of diotic signals.* Ph.D. thesis, Victoria University of Wellington, Wellington, New Zealand.

Lapsley Miller, J. A., Scurfield, B. K., Drga, V. F., Galvin, S. J., & Whitmore, J. K. (1998) *Single–interval and two–interval forced–choice tasks in the theory of signal detectability.* In preparation.

Leshowitz, B. (1969) Comparison of ROC curves from one– and two–interval rating–scale procedures. *Journal of the Acoustical Society of America*, *46*(2(2)), 399–402.

Licklider, J. C. R. & Dzendolet, E. (1948) Oscillographic scatterplots illustrating various degrees of correlation. *Science*, *107*, 121–124.

Lindsey, J. W. & Soderquist, D. R. (1972) Binaural analysis as a function of physiological masking. *Journal of the Acoustical Society of America*, *52*, 170(A).

Luce, R. D. (1997) Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology*, *41*, 79–87.

Mackworth, J. F. (1970) *Vigilance and attention: A signal detection approach.* Penguin Books, London.

Macmillan, N. A. & Kaplan, H. L. (1985) Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*(1), 185–199.

Mantel, N. (1969) Functional averages of a variable. *The American Statistician*, *23*(1), 21–22.

Marill, T. (1956) *Detection theory and psychophysics.* Tech. rep. MIT Technical Report 319, MIT Research Laboratory of Electronics.

Markowitz, J. & Swets, J. A. (1967) Factors affecting the slope of empirical ROC curves: Comparison of binary and rating responses. *Perception & Psychophysics*, *2*, 91–97.

McAulay, K. (1978) *A temporal model of aural frequency discrimination.* Ph.D. thesis, Victoria University of Wellington, Wellington, New Zealand.

McFadden, D. (1968) Masking-level differences determined with and without interaural disparities in masker intensity. *Journal of the Acoustical Society of America*, *44*(1), 212–223.

McGill, W. J. (1967) Neural counting mechanisms and energy detection in audition. *Journal of Mathematical Psychology*, *4*, 351–376.

McGill, W. J. & Teich, M. C. (1991) Auditory signal detection and amplification in a neural transmission network. In Commons, M. C., Nevin, J. A., & Davison, M. C. (Eds.), *Signal detection: Mechanisms, models and applications*, chap. 1. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

McKinley, R. C. & Weber, D. L. (1994) Detection and recognition of repeated tones and tonal patterns. *Journal of the Acoustical Society of America*, *95*(5), 2642–2651.

McNemar, Q. (1955) *Psychological Statistics.* John Wiley & Sons, New York.

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

Metz, C. E. & Shen, J. (1992) Gains from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis. *Medical Decision Making*, *12*, 60–75.

Moise, A., Clement, B., Ducimetiere, P., & Bourassa, M. (1985) Comparison of receiver operating curves derived from the same population: A bootstrapping approach. *Computers and Biomedical Research*, *18*, 125–131.

Nachmias, J. (1968) Effects of presentation probability and number of response alternatives on simple visual detection. *Perception & Psychophysics*, *3*(2B), 151–155.

Nash, J. (1979) *Compact Numerical Methods for Computers: Linear algebra and function minimisation.* Adam Hilger Ltd, Bristol.

Norris, N. (1976) General means and statistical theory. *The American Statistician*, *30*(1), 8–12.

Parzen, E. (1960) *Modern probability theory and its applications.* John Wiley & Sons, NY.

Pfafflin, S. M. (1968) Detection of auditory signal in restricted sets of reproducible noise. *Journal of the Acoustical Society of America*, *43*(3), 487–490.

Pfafflin, S. M. & Mathews, M. V. (1966) Detection of auditory signals in reproducible noise. *Journal of the Acoustical Society of America*, *39*(2), 340–345.

Podd, J. V. (1975) *Type I and Type II ROC Analysis of Change in Human Decision Axis.* Master's thesis, Victoria University of Wellington, Wellington, New Zealand.

Pollack, I. & Hsieh, R. (1969) Sampling variability of the area under the ROC–curve and of $d'_e$. *Psychological Bulletin*, *71*(3), 161–173.

Pollack, I. & Norman, D. (1964) A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*, 125–126.

Rabiner, L. R. & Gold, B. (1975) *Theory and application of digital signal processing.* Prentice–Hall.

Regan, D. (1972) *Evoked Potentials in Psychology, Sensory Physiology and Clinical Medicine.* Chapman and Hall, London.

Richards, V. M. & Zhu, S. (1994) Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *Journal of the Acoustical Society of America*, *95*(1), 423–434.

Robinson, D. E. & Watson, C. S. (1970) Psychophysical methods in modern psychoacoustics. In Tobias, J. V. (Ed.), *Foundations of modern auditory theory*, Vol. 2, pp. 101–131. Academic Press, New York.

Rockette, H. E., Gur, D., & Metz, C. E. (1992) The use of continuous and discrete judgements in ROC studies of diagnostic imaging techniques. *Investigative Radiology*, *27*, 169–172.

Ronken, D. A. (1969) Intensity discrimination of Rayleigh noise. *Journal of the Acoustical Society of America*, *45*(1), 54–57.

Rudin, W. (1976) *Principles of Mathematical Analysis*. McGraw-Hill Kogakusha Ltd, Tokyo.

Sakitt, B. (1973) Indices of discriminability. *Nature*, *241*, 133–134.

Schacknow, P. N. & Raab, D. H. (1976) Noise–intensity discrimination: Effects of bandwidth conditions and mode of masker presentation. *Journal of the Acoustical Society of America*, *60*(4), 893–905.

Schulman, A. I. & Greenberg, G. (1970) Operating characteristics and a priori probability of the signal. *Perception & Psychophysics*, *8*(5), 317–320.

Schulman, A. I. & Mitchell, R. R. (1966) Operating characteristics from yes–no and forced–choice procedures. *Journal of the Acoustical Society of America*, *40*(2), 473–477.

Scurfield, B. K. (1995) *Discrimination among events by neural networks*. Ph.D. thesis, Victoria University of Wellington, Wellington, New Zealand.

Scurfield, B. K. (1996) Multiple–event forced–choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, *40*(3), 253–269.

Scurfield, B. K. (1998) Generalization of the theory of signal detectability to $m$–dimensional $n$–event forced–choice tasks. *Journal of Mathematical Psychology*, *42*(1), 5–31.

Shaw, E. A. G. & Piercy, J. E. (1962) Physiological noise in relation to audiometry. *Journal of the Acoustical Society of America*, *34*, 745(A).

Shohat, J. (1930) Stieltjes integrals in mathematical statistics. *Annals of Mathematical Statistics*, *1*, 73–94.

Siegel, R. A. (1979) *Internal and External Noise in Auditory Detection*. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Siegel, R. A. & Colburn, H. S. (1983) Internal and external noise in binaural detection. *Hearing Research*, *11*, 117–123.

Siegel, R. A. & Colburn, H. S. (1989) Binaural processing of noisy stimuli: Internal/external noise ratios for diotic and dichotic stimuli. *Journal of the Acoustical Society of America*, *86*(6), 2122–2128.

Simpson, A. J. & Fitter, M. J. (1973) What is the best index of discriminability. *Psychological Bulletin*, *80*(6), 481–488.

Smith, M. & Wilson, E. (1953) A model of the auditory threshold and its application to the problem of the multiple observer. *Psychological Monographs: General and Applied*, *67*(9), 1–35.

Smith, W. D. (1995) Clarification of sensitivity measure $A'$. *Journal of Mathematical Psychology*, *39*, 82–89.

Soderquist, D. R. & Lindsey, J. W. (1972) Physiological noise as a masker of low frequencies: The cardiac cycle. *Journal of the Acoustical Society of America*, *52*(4), 1216–1220.

Somoza, E. & Mossman, D. (1991) ROC curves and the binormal assumption. *Diagnostic Testing in Neuropsychiatry*, *3*(4), 436–439.

Sorkin, R. D. & Dai, H. (1994) Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes*, *60*, 1–13.

Speeth, S. D. & Mathews, M. V. (1960) Sequential effects in the signal detection situation. *Journal of the Acoustical Society of America*, *32*, 932(A).

Spiegel, M. F. & Green, D. M. (1981) Two procedures for estimating internal noise. *Journal of the Acoustical Society of America*, *70*(1), 69–73.

Stearns, S. D. (1975) *Digital Signal Analysis*. Hayden Book Company, Rochelle Park, New Jersey.

Swets, J. A. (Ed.). (1964) *Signal detection and recognition by human observers: Contemporary readings* (1964 edition). John Wiley & Sons, New York.

Swets, J. A., Shipley, E. F., McKey, M. J., & Green, D. M. (1959) Multiple observations of signals in noise. *Journal of the Acoustical Society of America*, *31*(4), 514–521.

Swets, J. A., Tanner, Jr., W. P., & Birdsall, T. G. (1961) Decision processes in perception. In Swets, J. A. (Ed.), *Signal detection and recognition by human observers: Contemporary readings* (1964 edition)., chap. 1. John Wiley & Sons, New York.

Swets, J. A. (1986) Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

Tanner, W. P. & Sorkin, R. D. (1970) The theory of signal detectability. In Tobias, J. V. (Ed.), *Foundations of modern auditory theory*, Vol. 2, pp. 65–98. Academic Press, New York.

Tanner, Jr., W. P. (1961) Physiological implications of psychophysical data. In Swets, J. A. (Ed.), *Signal detection and recognition by human observers: Contemporary readings* (1964 edition)., chap. 16. John Wiley & Sons, New York.

Tanner, Jr., W. P. & Birdsall, T. G. (1958) Definitions of $d'$ and $\eta$ as psychophysical measures. *Journal of the Acoustical Society of America*, *30*(10), 922–928.

Taylor, A. J. (1984) *Auditory psychophysics in birds: The effects of unique noise on sensitivity*. Ph.D. thesis, Victoria University of Wellington, Wellington, New Zealand.

Taylor, A. J., Boven, R., & Whitmore, J. K. (1991) Reduction of unique noise in the psychophysics of hearing by group operating characteristic analysis. *Psychological Bulletin*, *109*(1), 133–146.

Thijssen, J. & Vendrik, A. (1968) Internal noise and transducer function in sensory detection experiments: Evaluation of psychometric curves and of ROC curves. *Perception & Psychophysics*, *3*(5), 387–400.

Triesman, M. & Faulkner, A. (1984) The effect of signal probability on the slope of the receiver operating characteristic given by the rating procedure. *British Journal of Mathematical and Statistical Psychology*, *37*, 199–215.

Tucker, A., Evans, R. B., & Jeffress, L. A. (1967) ROC curves for multiple–signal levels in a detection task. *Journal of the Acoustical Society of America*, *41*, 1611(A).

Watson, C. S. (1963) *Signal detection and certain physical characteristics of the stimulus during the observation interval*. Ph.D. thesis, Indiana University, Bloomington.

Watson, C. S., Franks, J. R., & Hood, D. C. (1972) Detection of tones in the absence of external masking noise. I. effects of signal intensity and signal frequency. *Journal of the Acoustical Society of America*, *52*(2), 633–643.

Watson, C. S., Kellogg, S. C., Kawanishi, D. T., & Lucas, P. A. (1973) The *uncertain* response in detection–oriented psychophysics. *Journal of Experimental Psychology*, *99*(2), 180–185.

Watson, C. S., Rilling, M. E., & Bourbon, W. T. (1964) Receiver–operating characteristics determined by a mechanical analog to the rating scale. *Journal of the Acoustical Society of America*, *36*(2), 283–288.

Whitmore, J. K., Drga, V., & Taylor, A. (1993) A diotic amplitude discrimination experiment replicated 100 times..

Whitt, W. (1988) Stochastic ordering. In Kotz, S. & Johnson, N. (Eds.), *Encyclopedia of Statistical Sciences* (1988 edition)., Vol. 8. John Wiley & Sons, New York.

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

Wilcox, G. W. (1968) *Inter-observer agreement and models of monaural auditory processing in detection tasks*. Ph.D. thesis, University of Michigan.

Yerushalmy, J. (1969) The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. *Radiologic Clinics of North America*, *7*(3), 381–392.

Yost, W. A. (1988) The masking-level difference and overall masker level: Restating the internal noise hypothesis. *Journal of the Acoustical Society of America*, *83*(4), 1517–1521.

# Appendix A

# Arcsin-averaging

Two results are shown in this appendix: (1) the equation for a general linear transform of the $\sin(x)$ function, and (2) that the *transform-average* based on $\arcsin(y)$ is identical to that based on $\arcsin(\sqrt{y})$.[1]

### The general linear transform of the sine function

The central, strictly monotonic increasing, sigmoidal section of the sine function is defined for values of the argument between $\frac{-\pi}{2}$ and $\frac{\pi}{2}$ radians. This section may be transformed linearly (in both directions) onto the regions defined by the intervals $[a, b]$ and $[c, d]$ as sketched in Figure A.1. There are no restrictions on $a$, $b$, $c$ and $d$ other than that they are real-valued and finite, and that $a < b$ and $c < d$. The general equation for the sketched function is

$$g(x) = \left(\frac{d-c}{2}\right)\left[\sin\left(\left(\frac{x-a}{b-a}\right)\pi - \frac{\pi}{2}\right) + 1\right] + c \tag{A.1}$$

for $x : a \leqslant x \leqslant b$, which results in $g(x) : c \leqslant g(x) \leqslant d$. Since it is not immediately obvious that Equation A.1 relates to Figure A.1, a brief description of the equation is given below.

Let the argument of the sine function Equation A.1 be

$$x' = \left(\frac{x-a}{b-a}\right)\pi - \frac{\pi}{2}, \tag{A.2}$$

where $x'$ is a linear function of $x$. Equation A.1 then simplifies to

$$g(x) = \left(\frac{d-c}{2}\right)(\sin(x') + 1) + c. \tag{A.3}$$

[1] Notationally, $x$ and $y$ are just general variables in this appendix and do not refer to values on any decision axis.

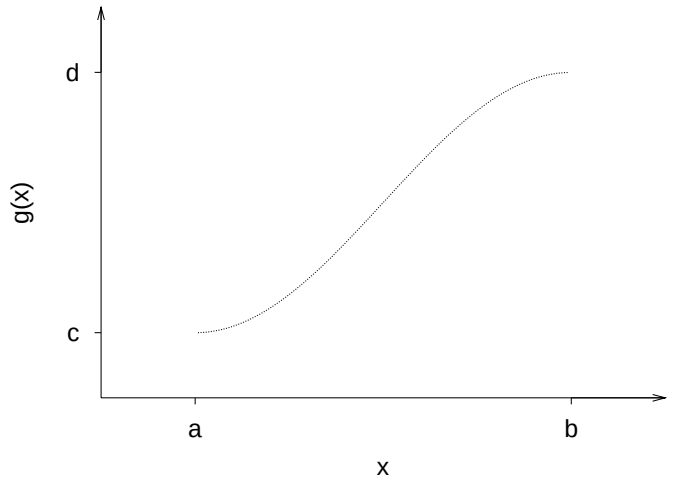


FIGURE A.1: General increasing linear transform of a section of the sine function.

If $a \leqslant x \leqslant b$, then

$$\frac{-\pi}{2} \leqslant x' \leqslant \frac{\pi}{2},$$

which can be seen by substituting $a$ and $b$ in turn for $x$ in Equation A.2. The result is as desired since it is the section of the sine function between $\frac{-\pi}{2}$ and $\frac{\pi}{2}$ that is of interest. Since $a \leqslant x \leqslant b$, then

$$-1 \leqslant \sin(x') \leqslant 1.$$

Substituting $\sin(x') = -1$ and $\sin(x') = 1$ in turn in Equation A.3 means that

$$c \leqslant \left(\frac{d-c}{2}\right)(\sin(x') + 1) + c \leqslant d, \tag{A.4}$$

which is just a linear function of $\sin(x')$.

Together, Equations A.1 to A.4 mean that if $a \leqslant x \leqslant b$ then $c \leqslant g(x) \leqslant d$, and hence that Equation A.1 is the correct description of the function sketched in Figure A.1. Starting with the sine function defined bewteen $\frac{-\pi}{2}$ and $\frac{\pi}{2}$, Equation A.2 is a linear rescaling in the horizontal direction while Equation A.4 is a linear rescaling in the vertical direction.

### Application to transform-averaging

If Equation A.1 gives $y = g(x)$, then the inverse function, $x = g^{-1}(y)$, is given by

$$g^{-1}(y) = \left(\frac{b-a}{2}\right)\left[\frac{2}{\pi}\arcsin\left(2\left(\frac{y-c}{d-c}\right) - 1\right) + 1\right] + a, \tag{A.5}$$

for $c \leqslant y \leqslant d$.

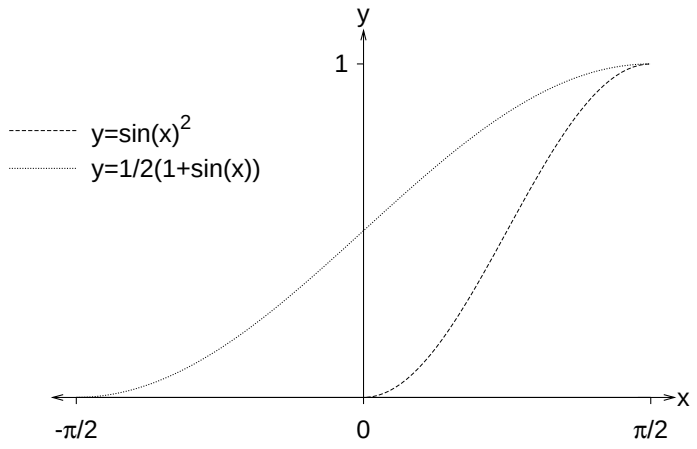


FIGURE A.2: The functions $y = \frac{1}{2}(1 + \sin(x))$ and $y = \sin^2(x)$.

Equations A.1 and A.5 can be used to calculate the transform-average mean of a given set of values. In transform-average GOC analysis (Chapter 3), the equations would be applied to ratings. Either of the two equations could be defined as the forward transform, and the other as the reverse transform, although the resulting GOC curve will depend on which is used as which. Figure 3.1(c) shows the arcsin-mean GOC curve, where the transform-average (Equation 3.1) is Equation A.5 (with parameters $a = 0$, $b = 1$, $c = 1$ and $d = 36$, and with ratings $r_{ji}$ substituted for $y$). Figure 3.1(e) shows the sine-mean GOC curve, where the transform-average (Equation 3.1) is Equation A.1 (with parameters $a = 1$, $b = 36$, $c = 0$ and $d = 1$, and with ratings $r_{ji}$ substituted for $x$).

Equations A.1 and A.5 can also be used to calculate arcsin-averaged proportions (or probabilities), as in Section 2.3. Proportions lie between zero and one inclusive, so have $c = 0$ and $d = 1$. In Figure A.1, the proportion values, say $y_i$, contributing to the mean lie on the vertical axis between $c$ and $d$. They are converted to values $x_i$ on the horizontal axis via the relationship $x_i = g^{-1}(y_i)$. The arithmetic mean $x$-value, $\overline{x}$, is calculated and transformed back as $\overline{y} = g(\overline{x})$. The mean ROC curves presented throughout the earlier chapters were calculated using Equations A.1 and A.5 in the manner described above, specifically with $a = \frac{-\pi}{2}$, $b = \frac{\pi}{2}$, $c = 0$ and $d = 1$. This means the transform functions are

$$y = \frac{1}{2}(1 + \sin(x)) \tag{A.6}$$

and

$$x = \arcsin(2y - 1), \tag{A.7}$$

with the range $y \in [0, 1]$ being mapped onto the domain $x \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ (Figure A.2).

The rationale for transform-averaging like this is discussed in Section 2.3. It is similar to converting to $z$-scores using the inverse of the standard $N(0,1)$ Gaussian cumulative distribution function, $z = \Phi^{-1}(y)$, averaging $z$-scores and converting back using $y = \Phi(z)$. Using Equations A.6 and A.7 avoids the problem of using $\Phi^{-1}(y)$ when any of the $y_i$ equal either zero or one.

It also has been suggested (McNicol, 1972; Macmillan & Kaplan, 1985) that the transform

$$x = \arcsin\left(\sqrt{y}\right) \tag{A.8}$$

and its inverse

$$y = \sin^2(x) \tag{A.9}$$

be used in place of $x = \Phi^{-1}(y)$ and $y = \Phi(x)$ respectively. It is shown that using Equations A.8 and A.9 results in the same transform-average mean as when using Equations A.6 and A.7.

**The relationship between $\frac{1}{2}\arcsin(2y-1)$ and $\arcsin\left(\sqrt{y}\right)$**

It is easiest to show the equivalence of means using Equations A.1 and A.9. From well known trigonometric formulae,

$$\begin{aligned} y &= \sin^2(x) && \text{(A.10)} \\ &= \sin(x)\sin(x) \\ &= \frac{1}{2}\left(\cos(x-x) - \cos(x+x)\right) \\ &= \frac{1}{2}\left(1 - \cos(2x)\right) \\ &= \frac{1}{2}\left(1 + \sin\left(2x - \frac{\pi}{2}\right)\right), && \text{(A.11)} \end{aligned}$$

which is the same as Equation A.1 with $a = 0$, $b = \frac{\pi}{2}$, $c = 0$ and $d = 1$. In other words, $y = \sin^2(x)$ is a sine function that has been linearly transformed in both directions. The inverse transforms of Equations A.10 and A.11 are, respectively,

$$\begin{aligned} x &= \arcsin\left(\sqrt{y}\right) \\ &= \frac{1}{2}\left(\arcsin(2y-1) + \frac{\pi}{2}\right). && \text{(A.12)} \end{aligned}$$

The range of $y$ for Equation A.12, $y \in [0,1]$, is the same as that for Equations A.6 and A.7. Equation A.12 maps $y$-values onto the domain $x \in [0, \frac{\pi}{2}]$ whereas Equation A.7 maps $y$-values onto the domain $x \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ (Figure A.2). This difference is reflected in the argument of the sine functions in Equations A.11 and A.6 respectively. The relationship

between the two domains is a simple linear transform, namely for any $x_1 \in [\frac{-\pi}{2}, \frac{\pi}{2}]$, there is an $x_2 \in [0, \frac{\pi}{2}]$,[2] such that

$$x_2 = \frac{1}{2}\left(x_1 + \frac{\pi}{2}\right) \tag{A.13}$$

A given set of proportions ($y$ values) could be transformed into either a set of $x_1$ values (via Equation A.7), or into a set of $x_2$ values (via Equation A.12, or via Equation A.7 followed by A.13). The means of the two sets of $x$-values are related by Equation A.13, which is to say that $\overline{x_2} = \frac{1}{2}\left(\overline{x_1} + \frac{\pi}{2}\right)$. Applying either Equation A.6 to $\overline{x_1}$ or Equation A.12 (or A.9) to $\overline{x_2}$ results in the same (arcsin-averaged) $y$-value, regardless of whether the $x$-values were originally derived via Equation A.6 or Equation A.8.

[2]Here, the subscript "1" refers to transforms based on Equations A.6 and A.7, and the subscript "2" refers to transforms based on Equations A.8 and A.9.

# Appendix B

# Linear transforms of function domains

The purpose of this appendix is to provide some results that are used in Appendix C. These results show what happens when a linear increasing transform, $s = h(t) = kt + c$, is applied to the domain of a general Riemann-integrable function $\lambda_1(t)$, and its effect on the integral of the transformed function. It is important to note that the linear transform is applied to the *domain* of $\lambda_1(t)$ and not to its *range*. Once the primary result has been derived, it is then applied to a more complex transform that is a composite of sections of linear increasing functions. Each linear function is defined on a separate (disjoint) interval of the real number line, $\mathbb{R}$, where all of the functions have the same slope, but each has a different intercept. How these results relate to Appendix C is described at the end of this appendix.

For a given function, $\lambda_1(t)$, define a second function, $\lambda_2(s) = \lambda_1(t)\ \forall\ s, t : s = h(t) = kt + c$, so that

$$\lambda_2(s) \quad = \quad \lambda_1\left(\frac{s-c}{k}\right),$$

where $t = h^{-1}(s) = \frac{s-c}{k}$ is the inverse linear transform ($k$ and $c$ are real-valued constants, $k > 0$). Consider an interval[1] $t \in [a, b)$ mapped via $h$ onto $s \in [ka + c, kb + c)$. The integral

[1] Square brackets denote inclusion, and round brackets denote exclusion. The choice of inclusion and exclusion reflects the usage of these results in Appendix C.

of $\lambda_2$ over the transformed interval is

$$\begin{aligned}
\int_{s=h(a)}^{h(b)} \lambda_2(s)\, ds &= \int_{s=ka+c}^{kb+c} \lambda_2(s)\, ds \\
&= \int_{s=ka+c}^{kb+c} \lambda_1\left(h^{-1}(s)\right)\, ds \\
&= \int_{s=ka+c}^{kb+c} \lambda_1\left(\frac{s-c}{k}\right)\, ds \\
&= k\int_{t=a}^{b} \lambda_1(t)\, dt, \qquad \text{(B.1)}
\end{aligned}$$

which is achieved by substitution using $t = \frac{s-c}{k}$. The scalar value $k$ in Equation B.1 comes from the fact that $ds = k\, dt$.

The result shows that when a linear transform $h$ of slope $k > 0$ is applied to the domain $[a, b)$ of a Riemann-integrable function $\lambda_1$, then the integral of the resulting function $\lambda_2$ over its related domain $[h(a), h(b))$ is equal to $k$ times the integral of $\lambda_1$ taken over $[a, b)$. Another interpretation of this result is that if the domain of a function is either stretched or compressed linearly by a factor of $k$, then the integral of the resulting function (taken over the new domain) is scaled by $k$. Note that this result is depends only on the slope, $k$, and is independent of the intercept $c$.

**Extension to transforms over unions of disjoint domains.** Let $I_1, I_2, I_3, \ldots$ be a sequence of disjoint, non-zero intervals in $\mathbb{R}$, where the boundaries of the $i^{th}$ interval are $a_i$ and $b_i$ $(a_i < b_i)$. Let $h_1, h_2, h_3, \ldots$ be a sequence of linear transforms associated with these intervals, where the $i^{th}$ transform, $h_i$, is defined only on the $i^{th}$ domain $I_i$. Let $J_i$ denote the *range* of the $i^{th}$ transform (where $J_i$ has boundary values at $h_i(a_i)$ and $h_i(b_i)$). Furthermore, let $h_i(t) = kt + c_i$, where the slope $k$ is identical for all $i$, and where the intercepts are arbitrary, subject to the constraint that all of the ranges $J_1, J_2, J_3, \ldots$ are disjoint. Let $I = \cup_i I_i$ and $J = \cup_i J_i$ be the unions of domains and ranges, respectively. Let the transform $h$ (without subscript) be defined as $h(t) = h_i(t)\ \forall\ t \in I_i$, which is the composite of all of the linear functions over their respective domains. (Note that $h(t)$ does not need to be defined for values of $t$ that lie outside of $I$).

For any given Riemann-integrable function, $\lambda_1(t)$, that is defined on $I$, let $\lambda_2(s) = \lambda_1(t)\ \forall\ s, t : s = h(t)$. The transform $h(t)$ provides a one-to-one mapping of $I$ onto $J$, and

therefore $s = h(t)$ has a unique inverse function. Equation B.1 shows that

$$\begin{aligned} \int_{J_i} \lambda_2(s)\, ds &= \int_{s=h(a_i)}^{h(b_i)} \lambda_2(s)\, ds \\ &= k \int_{t=a_i}^{b_i} \lambda_1(t)\, dt \\ &= k \int_{I_i} \lambda_1(t)\, dt \end{aligned} \tag{B.2}$$

holds for all $i$, because $h$ has the same slope, $k$, over each interval, $I_i$, and because Equation B.1 is independent of the intercept, $c_i$, for the $i^{th}$ interval.

Since the intervals $I_i$ are disjoint, and the intervals $J_i$ are disjoint, then

$$\begin{aligned} \int_{J} \lambda_2(s)\, ds &= \sum_{all\ i} \int_{J_i} \lambda_2(s)\, ds \\ &= \sum_{all\ i} k \int_{I_i} \lambda_1(t)\, dt \\ &= k \int_{I} \lambda_1(t)\, dt. \end{aligned} \tag{B.3}$$

Equation B.3 is applied only in Appendix C. Using some notation from Appendix C, the function $G(t)$ is substituted for $\lambda_1$, and $G^*(s)$ is substituted for $\lambda_2$. The unions of intervals—denoted as $I$ and $J$ here—are respectively used to represent either $I^+$ and $J^+$, $I^-$ and $J^-$, or $I^0$ and $J^0$, all of which are defined in Appendix C.

# Appendix C

# Stochastic ordering

The topic of this appendix is the *stochastic ordering* of random variables. Stochastic ordering is also known as *stochastic dominance*, particularly when applied to models involving utility and risk functions. Stochastic ordering is not new in statistics, and thorough bibliographies on the topic can be found in Findlay and Whitmore (1978), Kroll and Levy (1980) and Whitt (1988). There are various types of stochastic ordering. The type that is most relevant to the theory of GOC analysis is called *first-degree* stochastic ordering. There is also an important distinction between *strict* and *non-strict* stochastic ordering (which is explained in detail in Section C.1). Most of the literature deals with non-strict stochastic ordering, but both types are crucial to the theory of GOC analysis. Fishburn and Vickson (1978) provide a series of proofs that are the most similar to the material in this appendix. The context and assumptions underlying Fishburn and Vickson's work are not entirely suitable for a general theory of GOC analysis, and readers are referred to Fishburn and Vickson (1978, Appendices 2A and 2B in particular) for a comparison with the proofs presented here.

A variety of definitions, theorems and corollaries about stochastic ordering are presented in this appendix. The results are applied in Chapter 5 to the model of a unique-noise-affected ideal observer shown in Figure 5.2 and form the core of the theory of GOC analysis. The appendix is divided into three sections. Section C.1 provides definitions of stochastic ordering and deals with the effect of continuous, s.m.i. transforms of stochastically ordered random variables. Section C.2 deals with s.m.i. transforms of random variables that are not stochastically ordered. Section C.3 deals with stochastically ordered random variables that are transformed using a (non-strict) monotonic increasing step function.

Although this appendix primarily deals with the stochastic ordering of random variables, stochastic ordering can also be applied to *sample sets* of values, where each set is sampled from a different random variable. In that case, *sample* cumulative distribution functions take the place of the cumulative distribution functions in the derivations and

results in Sections C.1, C.2 and C.3. Due to sampling variability, stochastic ordering of a set of random variables does not guarantee stochastic ordering of sample sets based on the random variables, nor vice versa.

The notation used in this appendix follows the notation used in Chapter 5. Unless stated otherwise, the random variables described in this appendix can be either discrete, mixed or continuous. (Random variables that are singular continuous (Parzen, 1960), either in whole or in part, are excluded from consideration, because the integrals that are used in the derivations would be undefined.)

Much use is made here of Riemann-Stieltjes integrals, but in the derivations, these mostly result in the more familiar Riemann integrals and are, at worst, piece-wise integrable. Background material on the calculus of Riemann-Stieltjes integrals is available in Clarke (1975) and Rudin (1976), for example. Summaries of how Riemann-Stieltjes integrals can be applied to probability and statistics are provided by Clarke (1975), Parzen (1960) and Shohat (1930).[1]

## C.1 The effect of strictly monotonic increasing transforms of random variables on stochastic ordering, and on the ordering of expected values

This section provides a definition of stochastic ordering, and describes properties that are held by stochastically ordered sets of random variables. This is followed by Theorem 1 and its corollaries, which show the effect of strictly monotonic increasing transforms of stochastically ordered random variables. The effect of such transforms on expected values of random variables is also described.

**Definition 1** *Consider any two random variables, $Y_1$ and $Y_2$, which are either continuous, or discrete, or mixed, and which have respective cumulative distribution functions $F_{Y_1}$ and $F_{Y_2}$. $Y_1$ is stochastically less than $Y_2$ (denoted $Y_1 \overset{st}{<} Y_2$) if and only if $F_{Y_1}(t) \geqslant F_{Y_2}(t)\ \forall\, t \in \mathbb{R}$ and if $F_{Y_1}(t) > F_{Y_2}(t)$ for some non-zero interval on the real number line $\mathbb{R}$. The converse is $Y_1 \overset{st}{\not<} Y_2$, which means that $Y_1 \overset{st}{<} Y_2$ is not true.*

**Definition 2** *For two random variables, $Y_1$ and $Y_2$, $Y_1$ is stochastically less than or equal to $Y_2$ (denoted $Y_1 \overset{st}{\leqslant} Y_2$) if and only if $F_{Y_1}(t) \geqslant F_{Y_2}(t)\ \forall\, t \in \mathbb{R}$. The converse is $Y_1 \overset{st}{\not\leqslant} Y_2$, which means that $Y_1 \overset{st}{\leqslant} Y_2$ is not true.*

[1] Although readers should beware of typographical errors in Shohat (1930).

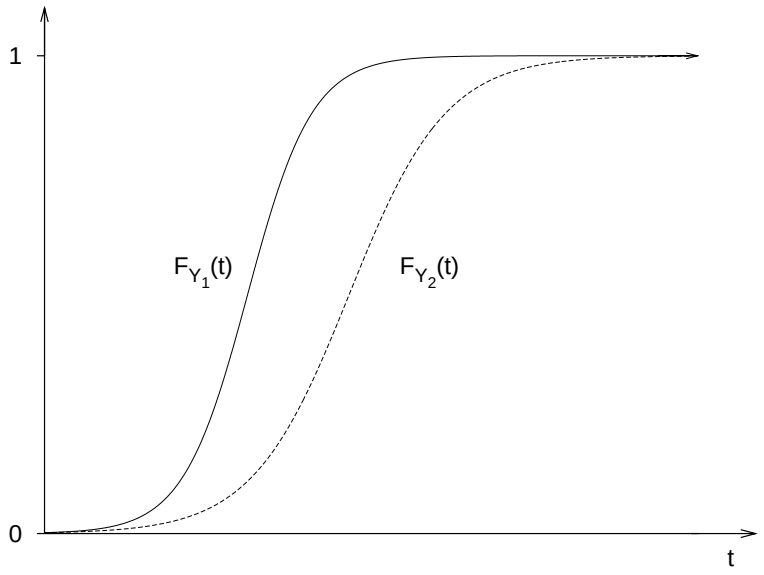


FIGURE C.1:

Cumulative distribution functions, $F_{Y_1}(t)$ and $F_{Y_2}(t)$, of two stochastically ordered random variables, $Y_1$ and $Y_2$, where $Y_1 \overset{st}{<} Y_2$. In this example, $Y_1$ and $Y_2$ are both continuous.

The difference between Definitions 1 and 2 is that $Y_1 = Y_2$ is a possibility in Definition 2 (i.e. the two cumulative distribution functions may be identical), but not in Definition 1. The relationship $Y_1 \overset{st}{\leqslant} Y_2$ implies that either $Y_1 \overset{st}{<} Y_2$ or $Y_1 = Y_2$ applies, but not both. Figure C.1 shows an example of the cumulative distribution functions of $Y_1$ and $Y_2$, where $Y_1 \overset{st}{<} Y_2$.

The stochastic ordering of random variables *sometimes* follows the same rules as the numerical ordering of quantities, but not always. For example, if $Y_1 \overset{st}{\leqslant} Y_2$ *and* $Y_1 \overset{st}{<} Y_2$ then $Y_1 \neq Y_2$. Similarly, if $Y_1 \overset{st}{\leqslant} Y_2$ *and* $Y_1 \overset{st}{\not<} Y_2$ then $Y_1 = Y_2$. Also, note that $Y_1 \overset{st}{<} Y_2$ implies $Y_1 \overset{st}{\leqslant} Y_2$, but not vice versa, and that $Y_1 = Y_2$ implies $Y_1 \overset{st}{\leqslant} Y_2$, but not vice versa. However, $Y_1 \overset{st}{\not<} Y_2$ *does not* imply that $Y_1 \overset{st}{\geqslant} Y_2$, and $Y_1 \overset{st}{\not\leqslant} Y_2$ *does not* imply that $Y_1 \overset{st}{>} Y_2$. Furthermore, $Y_1 \neq Y_2$ does not imply that either $Y_1 \overset{st}{<} Y_2$ or $Y_2 \overset{st}{<} Y_1$ holds. It is possible to have $Y_1$ and $Y_2$ such that $Y_1 \neq Y_2$, $Y_1 \overset{st}{\not<} Y_2$ and $Y_1 \overset{st}{\not>} Y_2$ all hold simultaneously. An example of this case is used in the proof of Theorem 2.

**Corollary 1** *(Transitivity) Let* $Y_1$, $Y_2$ *and* $Y_3$ *be any three random variables. If* $Y_1 \overset{st}{<} Y_2$ *and* $Y_2 \overset{st}{<} Y_3$, *then* $Y_1 \overset{st}{<} Y_3$. *If* $Y_1$, $Y_2$ *and* $Y_3$ *are such that either* $Y_1 \overset{st}{<} Y_2 \overset{st}{\leqslant} Y_3$ *or* $Y_1 \overset{st}{\leqslant} Y_2 \overset{st}{<} Y_3$, *then* $Y_1 \overset{st}{<} Y_3$. *If* $Y_1 \overset{st}{\leqslant} Y_2 \overset{st}{\leqslant} Y_3$, *then* $Y_1 \overset{st}{\leqslant} Y_3$.

**Definition 3** *The domain of a random variable is the set of values of* $\mathbb{R}$ *for which the probability mass* or *density is non-zero. The* mutual domain *of a set of random variables is the smallest continuous interval on* $\mathbb{R}$ *containing all values of the union of the domains of the random variables.*

The *mutual domain* is not a standard mathematical concept, but is introduced here for later convenience. The mutual domain of a set of random variables is not necessarily the same as the union of the domains of the random variables, although it may be. An example of where they are not the same is where $Y_1$ and $Y_2$ are two continuous uniform random variables whose domains do not overlap and are separated by a gap. The mutual domain of $Y_1$ and $Y_2$ is the union of the domains of each random variable *and* the interval of the gap. If $Y_1$ and $Y_2$ were discrete random variables, the mutual domain includes the points at which probability is massed, as well as all of the intervals between these points. The concept is convenient when applying the same transform to a multitude of random variables simultaneously. The mutual domain may be the entire real number line, although it need not be.

**Definition 4** *In terms of a Riemann-Stieltjes integral, the expectation of a random variable* $Y$ *is*

$$E(Y) = \int_{t=-\infty}^{\infty} t \, dF_Y(t) \tag{C.1}$$

*(Clarke, 1975; Parzen, 1960).*

**Definition 5** $E(Y)$ *exists and is finite if and only if* $\int_{t=-\infty}^{\infty} t \, dF_Y(t)$ *is absolutely integrable, that is, when* $\int_{t=-\infty}^{\infty} |t| \, dF_Y(t) < \infty$ *(Clarke, 1975; Parzen, 1960).*

Computationally, Definition 5 is equivalent to determining that the Riemann integrals

$$\int_{t=0}^{\infty} (1 - F_Y(t)) \, dt \qquad \text{and} \qquad \int_{t=-\infty}^{0} F_Y(t) \, dt$$

both converge and are finite (Clarke, 1975), in which case

$$E(Y) = \int_{t=0}^{\infty} (1 - F_Y(t)) \, dt - \int_{t=-\infty}^{0} F_Y(t) \, dt \tag{C.2}$$

(Clarke, 1975; Parzen, 1960). Equation C.2 may be interpreted geometrically (Parzen, 1960, pp. 211-212). Say the cumulative distribution function $F_Y(t)$ is as graphed in Figure C.2. Let $A_Y$ be the area of the region below $F_Y(t)$ and to the left of $t = 0$, and let $B_Y$ be the area of the region above $F_Y(t)$ and below one that lies to the right of $t = 0$. The expected value, $E(Y)$, is then $B_Y$ minus $A_Y$. Equations C.1 and C.2 are useful because of their generality. The expectation is described entirely in terms of the cumulative

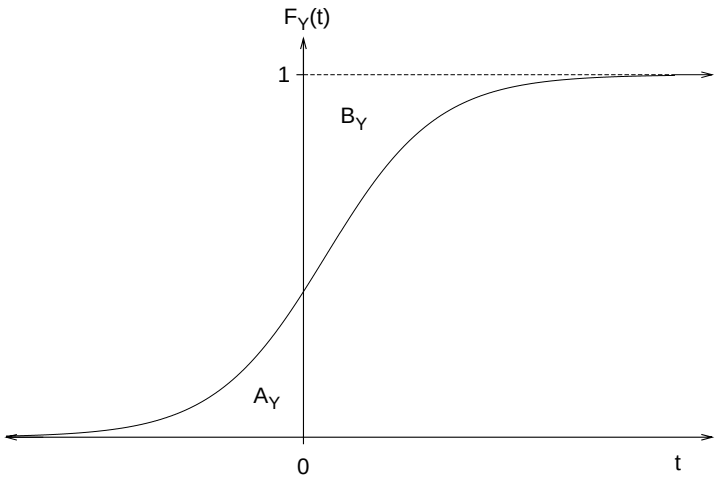


FIGURE C.2:
Regions defined by the cumulative distribution function, $F_Y(t)$, of a random variable, $Y$, relating to the expected value, $E(Y)$ (after Parzen, 1960, Figure 2A).

distribution function $F_Y(t)$, and it applies equally well to discrete, continuous, or mixed random variables. If $Y$ is either discrete or mixed, then the integrals are evaluated in a piecewise manner.

**Theorem 1** *If $Y_1$ and $Y_2$ are any two continuous, mixed or discrete random variables whose expectations exist and are finite, then $Y_1 \overset{st}{<} Y_2$ implies that $E(Y_1) < E(Y_2)$. Furthermore, for any strictly monotonic increasing transform, $h$, defined over the mutual domain of $Y_1$ and $Y_2$, then $Y_1 \overset{st}{<} Y_2$ implies that $h(Y_1) \overset{st}{<} h(Y_2)$ and, consequently, that $E(h(Y_1)) < E(h(Y_2))$, if the expectations exist and are finite.*

**Proof.** Let $Y_1$ and $Y_2$ be any two continuous, mixed or discrete random variables whose expectations exist and are finite. In that case, from Definition 5, they may be expressed as the Riemann-Stieltjes integrals

$$E(Y_1) = \int_{t=-\infty}^{\infty} t \, dF_{Y_1}(t)$$

and

$$E(Y_2) = \int_{t=-\infty}^{\infty} t \, dF_{Y_2}(t). \tag{C.3}$$

The difference between the means is

$$\begin{aligned} E(Y_1) - E(Y_2) &= \int_{t=-\infty}^{\infty} t\, dF_{Y_1}(t) - \int_{t=-\infty}^{\infty} t\, dF_{Y_2}(t) \\ &= \int_{t=-\infty}^{\infty} t\, dG(t) \end{aligned} \tag{C.4}$$

where $G(t)$ is defined as

$$G(t) \stackrel{def}{=} F_{Y_1}(t) - F_{Y_2}(t) \quad \forall\, t \in \mathbb{R}. \tag{C.5}$$

Figures C.3 and C.4 sketch $F_{Y_1}(t)$, $F_{Y_2}(t)$ and $G(t)$ for hypothetical continuous and discrete examples.

Integrating Equation C.4 by parts[2] gives

$$E(Y_1) - E(Y_2) = \left[t\, G(t)\right]_{t=-\infty}^{\infty} - \int_{t=-\infty}^{\infty} G(t)\, dt, \tag{C.6}$$

The first term on the right-hand side of Equation C.6 vanishes to zero because

$$\begin{aligned} \left[t\, G(t)\right]_{t=-\infty}^{\infty} &= \lim_{\substack{a\to-\infty \\ b\to\infty}} \left[t\, G(t)\right]_{t=a}^{b} \\ &= \lim_{\substack{a\to-\infty \\ b\to\infty}} \left[t\, \left(F_{Y_1}(t) - F_{Y_2}(t)\right)\right]_{t=a}^{b} \\ &= \lim_{a\to-\infty} \left[t\, \left(F_{Y_1}(t) - F_{Y_2}(t)\right)\right]_{t=a}^{0} + \lim_{b\to\infty} \left[t\, \left(1 - F_{Y_2}(t)\right) - t\, \left(1 - F_{Y_1}(t)\right)\right]_{t=0}^{b} \\ &= 0, \end{aligned}$$

since

$$\begin{aligned} \lim_{a\to-\infty} aF(a) &= -1 \times \left[\lim_{a\to\infty} aF(-a)\right] && \text{(C.7)} \\ &= -1 \times 0 \\ &= 0 && \text{(C.8)} \end{aligned}$$

and

$$\lim_{b\to\infty} b\,(1 - F(b)) = 0 \tag{C.9}$$

for the cumulative distribution function $F$ of any random variable whose expectations

[2]The expression in Equation C.4 is a meaningful Riemann-Stieltjes integral, even though $G(t)$ is not monotonic increasing, and hence integration by parts is permissable. This is because $G(t)$ is of bounded variation by virtue of its composition. See Clarke (1975), Rudin (1976) or Borowski and Borwein (1989) with respect to functions of bounded variation and their use in Riemann-Stieltjes integrals.

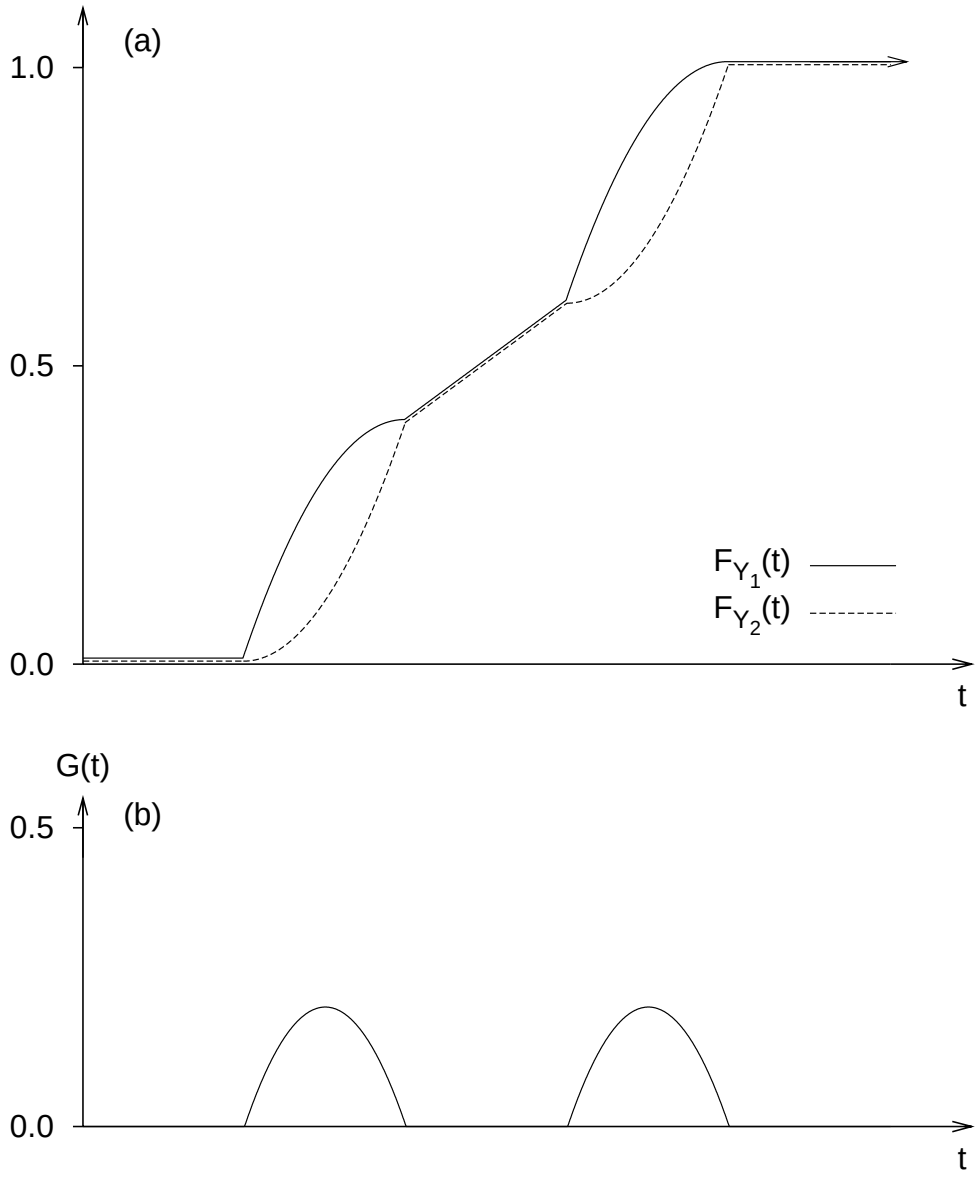


FIGURE C.3:
(a) Cumulative distribution functions, $F_{Y_1}(t)$ and $F_{Y_2}(t)$, of two stochastically ordered continuous random variables, $Y_1$ and $Y_2$, where $Y_1 \overset{st}{<} Y_2$. (The functions are offset slightly where they are equal.) (b) The difference function $G(t) = F_{Y_1}(t) - F_{Y_2}(t)$.

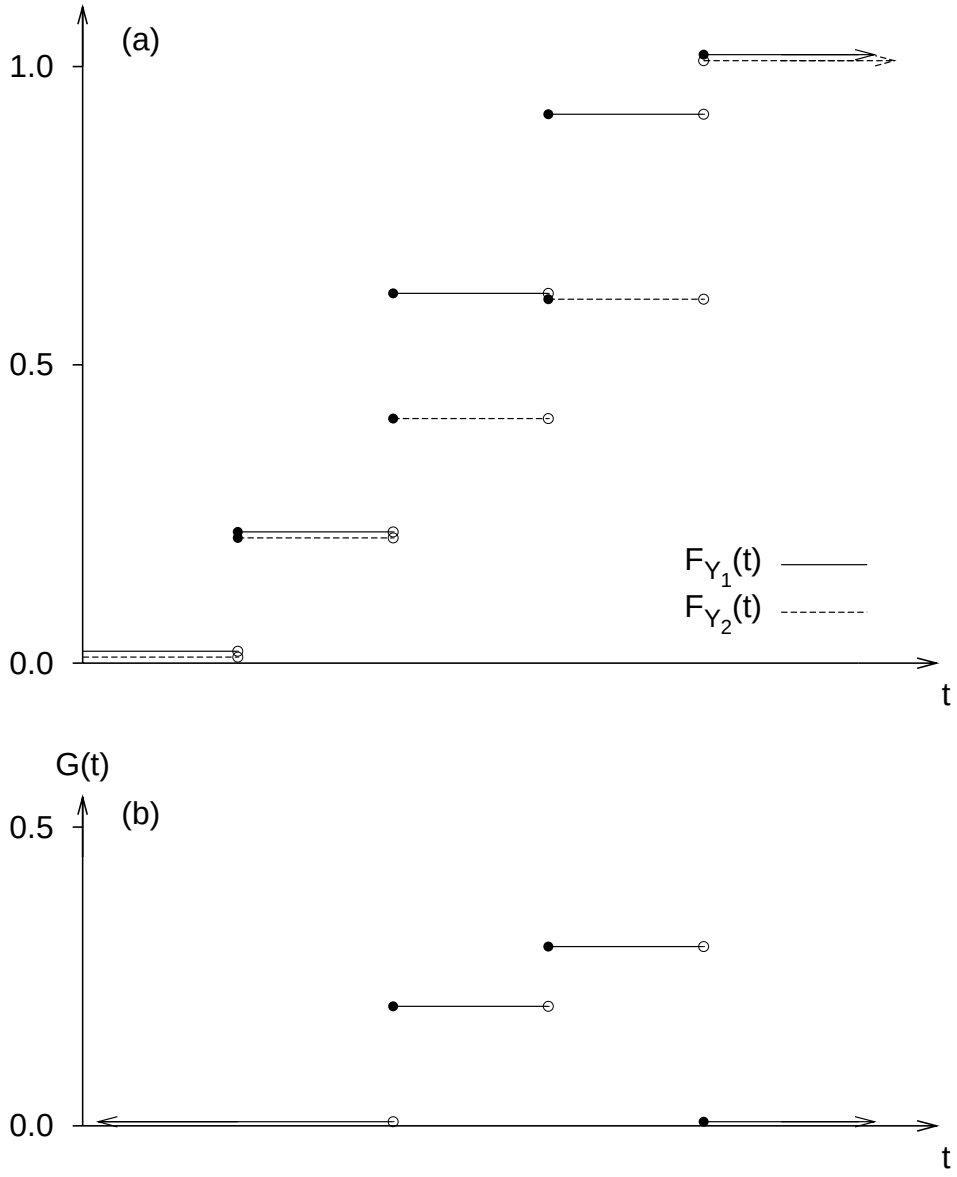


FIGURE C.4:
(a) Cumulative distribution functions, $F_{Y_1}(t)$ and $F_{Y_2}(t)$, of two stochastically ordered discrete random variables, $Y_1$ and $Y_2$, where $Y_1 \overset{st}{<} Y_2$. (Filled circles denote inclusion. Empty circles denote exclusion. The functions are offset slightly where they are equal.) (b) The difference function $G(t) = F_{Y_1}(t) - F_{Y_2}(t)$.

exist and are finite (Clarke, 1975).[3] Equation C.6 then simplifies to

$$E(Y_1) - E(Y_2) \quad = \quad -\int_{t=-\infty}^{\infty} G(t)\, dt. \tag{C.10}$$

The integral in Equation C.10 is a Riemann integral, regardless of whether $Y_1$ and $Y_2$ are continuous, discrete, or mixed. If either of $Y_1$ and $Y_2$ is either discrete or mixed, then $G(t)$ is not continuous over the entire real number line, and Equation C.10 is evaluated in a piecewise manner, as is needed in Figure C.4, for example.

A second way of arriving at the same result implicitly involves the same vanishing of terms as given above, which is required in order for the integrals in Equation C.2 to converge (i.e. when expectations exist and are finite). From Equation C.2,

$$E(Y_2) \quad = \quad \int_{t=0}^{\infty} \left(1 - F_{Y_2}(t)\right)\, dt - \int_{t=-\infty}^{0} F_{Y_2}(t)\, dt,$$

and so

$$\begin{aligned} E(Y_1) - E(Y_2) \quad &= \quad \int_{t=0}^{\infty} \left[\left(1 - F_{Y_1}(t)\right) - \left(1 - F_{Y_2}(t)\right)\right] dt \ldots \\ &\qquad \ldots - \int_{t=-\infty}^{0} \left(F_{Y_1}(t) - F_{Y_2}(t)\right) dt \qquad (C.11) \\ &= \quad -\int_{t=-\infty}^{\infty} \left(F_{Y_1}(t) - F_{Y_2}(t)\right)\, dt \\ &= \quad -\int_{t=-\infty}^{\infty} G(t)\, dt. \end{aligned}$$

(The integrals involved in Equation C.11 can be related back to Figure C.2, in terms of differences in the areas of the regions $A_Y$ and $B_Y$, applied separately to $Y_1$ and $Y_2$. In this context, Equation C.11 states that $E(Y_1) - E(Y_2) = (B_{Y_1} - B_{Y_2}) - (A_{Y_1} - A_{Y_2})$. This equation is not pursued further, but is noted as a geometrical interpretation of $E(Y_1) - E(Y_2)$.)

Definition 1 and Equation C.5 together imply that if $Y_1 \overset{st}{<} Y_2$, then $G(t) \geqslant 0 \;\; \forall\, t \in \mathbb{R}$ and $G(t) > 0$ for some non-zero intervals in $\mathbb{R}$. This implies that if $Y_1 \overset{st}{<} Y_2$, then the integral of $G(t)$ is positive, that is

$$\int_{t=-\infty}^{\infty} G(t)\, dt \quad > \quad 0. \tag{C.12}$$

Taken together with Equation C.10, this implies that $E(Y_1) - E(Y_2) < 0$, and so $E(Y_1) < E(Y_2)$. This completes the proof of the first part of Theorem 1.

[3] Equations C.8 and C.9 hold (and the expectation exists) for many, but not all, random variables. For example, the mean of a Cauchy random variable does not exist, precisely because the limits given in Equations C.7 and C.9 are infinite.

A trivial corollary of this result is that if $Y_1 \overset{st}{>} Y_2$, then $E(Y_1) > E(Y_2)$.

If $Y_1 \neq Y_2$ and if neither $Y_1 \overset{st}{<} Y_2$ nor $Y_1 \overset{st}{>} Y_2$ holds, then $F_{Y_1}(t)$ and $F_{Y_2}(t)$ must cross at least once. This means there is some non-zero interval in $\mathbb{R}$ for which $F_{Y_1}(t) > F_{Y_2}(t)$ holds true, and some interval for which $F_{Y_1}(t) < F_{Y_2}(t)$ holds true. Equivalently, there is some non-zero interval for which $G(t) > 0$ holds true, and some interval for which $G(t) < 0$ holds true. Hence no general statement may be made about the sign of the definite integral in Equation C.10, nor about the orderings of the $E(Y_1)$ and $E(Y_2)$. Equation C.10 would have to be specifically evaluated to determine the order of means.

Let $R_1 = h(Y_1)$, $R_2 = h(Y_2)$, where $h$ is any continuous s.m.i. transform defined over the mutual domain of $Y_1$ and $Y_2$ for which $E(R_1)$ and $E(R_2)$ exist and are finite. Since $h$ is s.m.i. over the domain of $Y_1$, then $\forall\, s, t : s = h(t)$,

$$\begin{aligned} F_{R_1}(s) &= P(R_1 \leqslant s) \\ &= P(R_1 \leqslant h(t)) \\ &= P(h^{-1}(R_1) \leqslant t) \\ &= P(Y_1 \leqslant t) \\ &= F_{Y_1}(t). \end{aligned}$$

Similarly, $F_{R_2}(s) = F_{Y_2}(t) \quad \forall\, s, t : s = h(t)$. Next, let $G^*(s)$ be defined as

$$G^*(s) \overset{def}{=} F_{R_1}(s) - F_{R_2}(s).$$

It follows from this that $\forall\, s, t : s = h(t)$,

$$G^*(s) = G(t). \tag{C.13}$$

The function $G^*(s)$, defined for $R_1$ and $R_2$, corresponds to $G(t)$, defined for $Y_1$ and $Y_2$. The equivalences that lead to Equation C.13 imply that if $F_{Y_1}(t) \geqslant F_{Y_2}(t) \quad \forall\, t \in \mathbb{R}$ and $F_{Y_1}(t) > F_{Y_2}(t)$ for some non-zero interval in $\mathbb{R}$, then $F_{R_1}(s) \geqslant F_{R_2}(s) \quad \forall\, s \in \mathbb{R}$ and $F_{R_1}(s) > F_{R_2}(s)$ for some non-zero interval in $\mathbb{R}$, *and vice versa*. In brief,

$$Y_1 \overset{st}{<} Y_2 \quad \Leftrightarrow \quad R_1 \overset{st}{<} R_2. \tag{C.14}$$

The order of expectations of the transformed random variables can also be determined by specifically calculating

$$E(R_1) - E(R_2) = -\int_{s=-\infty}^{\infty} G^*(s)\, ds, \tag{C.15}$$

which follows from Equation C.10 when applied to $R_1$, $R_2$ and $G^*(s)$. As with $Y_1$ and $Y_2$,

if $R_1 \overset{st}{<} R_2$, then the integral of $G^*(s)$ is positive, that is

$$\int_{s=-\infty}^{\infty} G^*(s)\, ds \quad > \quad 0. \tag{C.16}$$

Taken together with Equation C.15, this implies that $E(R_1) < E(R_2)$. This result holds for *any* s.m.i. transform $h$, where $R_1 = h(Y_1)$ and $R_2 = h(Y_2)$, as long as $E(R_1)$ and $E(R_2)$ exist and are finite. So if $Y_1 \overset{st}{<} Y_2$ then $R_1 \overset{st}{<} R_2$, which implies that $E(R_1) < E(R_2)$. Consequently, $Y_1 \overset{st}{<} Y_2$ implies that $E(h(Y_1)) < E(h(Y_2))$. Q.E.D.

Note that while stochastic ordering of random variables implies an ordering of the means, the converse is not true. An ordering of the means does not imply that stochastic ordering holds.

**Continuous random variables.** If $Y_1$ and $Y_2$ are both continuous random variables, with respective probability density functions $f_{Y_1}$ and $f_{Y_2}$, then $E(Y_1) = \int_{t=-\infty}^{\infty} t\, f_{Y_1}(t)\, dt$, and $E(Y_2) = \int_{t=-\infty}^{\infty} t\, f_{Y_2}(t)\, dt$. Equation C.4 can be evaluated using

$$\begin{aligned} E(Y_1) - E(Y_2) &= \int_{t=-\infty}^{\infty} t\, dG(t) \\ &= \int_{t=-\infty}^{\infty} t \left( \frac{d}{dt} G(t) \right) dt \\ &= \int_{t=-\infty}^{\infty} t \left( f_{Y_1}(t) - f_{Y_2}(t) \right) dt. \end{aligned}$$

**Discrete random variables.** If $Y_1$ and $Y_2$ are both discrete random variables, then their cumulative distribution functions are piecewise continuous, and so are $G(t)$ and $G^*(s)$. Figure C.4 sketched what $G(t)$ may look like in the discrete case. Integrals of such functions, and Equations C.10 and C.15, are calculated as Riemann integrals, evaluated in a piecewise manner. In practice, this requires a series summation (over possibly an infinite number) of definite Riemann integrals, each one defined over a finite interval, where the limits of each integral are neighbouring points at which probability is massed. If the expectations exist and are finite, then the sum of the definite integrals converges to $E(Y_1) - E(Y_2)$. The sum converges because $E(Y_1)$ and $E(Y_2)$ are both finite, and the convergence is absolute (Definition 5).

Discrete random variables are typically (but not necessarily) defined on an integer domain. In general, a discrete random variable may be defined over any countable set of values, $\{\ldots, c_0, c_1, c_2, \ldots\}$. Only real-valued discrete random variables whose values, $c_i$, are s.m.i. with their integer-valued index values $i$ are considered here.[4]

[4]Other real-valued discrete variables are possible whose values either do not, or cannot, form an s.m.i. mapping onto the integers. It is difficult to base ROC analysis on such variables and so they are excluded from consideration here.

If $c_{j_1}$ and $c_{j_2}$ are any two points at which probability is massed, then the integral of $G(t)$ evaluated between these points is

$$\int_{t=c_{j_1}}^{c_{j_2}} G(t)\, dt \;=\; \sum_{j=j_1}^{j_2-1} (c_{j+1} - c_j)\; G(c_j).$$

For integer-valued discrete random variables, $c_{j+1} - c_j = 1$, and so

$$\begin{aligned}\int_{t=c_{j_1}}^{c_{j_2}} G(t)dt &= \sum_{j=j_1}^{j_2-1} G(c_j) \\ &= \sum_{j=j_1}^{j_2-1} \left(F_{Y_1}(c_j) - F_{Y_2}(c_j)\right).\end{aligned}$$

**Corollary 2** *Let $Y_1$ and $Y_2$ be any two random variables whose expectations exist and are finite. Let $h_1$, $h_2$, $h_3$ ... be continuous, strictly monotonic increasing functions, where $h_1$ is defined on the mutual domain of $Y_1$ and $Y_2$, $h_2$ is defined on the mutual domain of $h_1(Y_1)$ and $h_1(Y_2)$, $h_3$ is defined on the mutual domain of $h_2(h_1(Y_1))$ and $h_2(h_1(Y_2))$, and so on. If $Y_1 \overset{st}{<} Y_2$, then*

$$E[\ldots h_3(h_2(h_1(Y_1)))] \;<\; E[\ldots h_3(h_2(h_1(Y_2)))],$$

*if the expectations exist and are finite.*

**Proof.** Consider first the case of two transforms, $h_1$ and $h_2$. Corollary 2 results then from the double application of Theorem 1. From Theorem 1, if $Y_1 \overset{st}{<} Y_2$, then $h_1(Y_1) \overset{st}{<} h_1(Y_2)$. And again from Theorem 1, if $h_1(Y_1) \overset{st}{<} h_1(Y_2)$ then $h_2(h_1(Y_1)) \overset{st}{<} h_2(h_1(Y_2))$, which implies that $E[h_2(h_1(Y_1))] < E[h_2(h_1(Y_2))]$. By extension, and by repeated application of Theorem 1, any number of s.m.i. transforms can be nested to produce the desired result. Another way of viewing this is that the s.m.i. transform of an s.m.i. transform is itself an s.m.i. transform (i.e. $h_0(t) = h_2(h_1(t))$ is an s.m.i. transform). The series can be extended so that $h_0(t) = \ldots h_3(h_2(h_1(t))$, which is an s.m.i. transform, and Theorem 1 applied to $h_0$ also gives the desired result.

**Corollary 3** *If there is a stochastically ordered set of random variables, $\{Y_1, Y_2, Y_3 \ldots\}$ such that $Y_1 \overset{st}{<} Y_2 \overset{st}{<} Y_3 \overset{st}{<} \ldots$, then $h(Y_1) \overset{st}{<} h(Y_2) \overset{st}{<} h(Y_3) \overset{st}{<} \ldots$ holds for any continuous s.m.i. transform $h$ defined over the mutual domain of all of the $Y_j$.*

**Corollary 4** *For any $\{Y_1, Y_2, Y_3 \ldots\}$, defined and stochastically ordered as in Corollary 3, and for any continuous s.m.i. transform $h$ defined over the mutual domain of all the $Y_j$, $E(Y_1) < E(Y_2) < E(Y_3) \ldots$ and $E(h(Y_1)) < E(h(Y_2)) < E(h(Y_3)) \ldots$ The ordering of the expected values of both the untransformed and the transformed random variables follows the stochastic ordering of the Y-variables, if the expectations exist and are finite.*

Corollaries 3 and 4 follow directly from the application of Equation C.14 and Theorem 1 to successive overlapping pairs of the $Y$-variables ($Y_1$ and $Y_2$, then $Y_2$ and $Y_3$, etc.).

## C.2 Strictly monotonic increasing transforms of random variables that are not stochastically ordered

The results and derivations presented in the preceding section showed what happens when an s.m.i. function is used to transform random variables that are stochastically ordered. In contrast, this section shows what happens under such transforms when the random variables are *not* stochastically ordered. Before the main result (Theorem 2) is derived, some of the integrals involved in the preceding section are reformulated, in order to simplify later proofs.

**Notation.** The symbols $\begin{matrix}\geqq\\<\end{matrix}$ and $\begin{matrix}\leqq\\>\end{matrix}$ are used to succinctly provide three relationships in one statement. For example, "$A \begin{matrix}\geqq\\<\end{matrix} B \Rightarrow C \begin{matrix}\leqq\\>\end{matrix} D$" should be read as "if $A > B$ then $C < D$; if $A = B$ then $C = D$; and if $A < B$ then $C > D$." In derivations that use these stacked symbols, it may be helpful to follow only the top inequality. If that holds, then the bottom inequality holds because it is the converse of the top inequality, and the middle equality also holds.

Let $I^+$, $I^-$ and $I^0$ respectively be the union of intervals in $\mathbb{R}$ over which $G(t) > 0$, $G(t) < 0$ and $G(t) = 0$, or equivalently, the union of intervals in $\mathbb{R}$ over which $F_{Y_1}(t) > F_{Y_2}(t)$, $F_{Y_1}(t) < F_{Y_2}(t)$ and $F_{Y_1}(t) = F_{Y_2}(t)$. Figure C.5 shows how $G(t)$ relates to $F_{Y_1}(t)$ and $F_{Y_2}(t)$, and how $I^+$, $I^-$ and $I^0$ also relate to these functions. Let $\int_{I^+} G(t)\,dt$, $\int_{I^-} G(t)\,dt$ and $\int_{I^0} G(t)\,dt$ denote the integral of $G(t)$, with respect to $t$, taken over the intervals that make up $I^+$, $I^-$ and $I^0$, respectively. Note that $I^+ \cup I^- \cup I^0 = \mathbb{R}$, and also that $\int_{I^0} G(t)\,dt = 0$ by the definition of $I^0$. Given these definitions, then

$$\begin{aligned}\int_{t=-\infty}^{\infty} G(t)\,dt &= \int_{I^+} G(t)\,dt + \int_{I^-} G(t)\,dt + \int_{I^0} G(t)\,dt \\ &= \int_{I^+} G(t)\,dt + \int_{I^-} G(t)\,dt \\ &= \int_{I^+} G(t)\,dt - \left|\int_{I^-} G(t)\,dt\right|, \end{aligned} \tag{C.17}$$

where $\int_{I^+} G(t)\,dt > 0$ and $\int_{I^-} G(t)\,dt < 0$ by the definitions of $I^+$ and $I^-$.

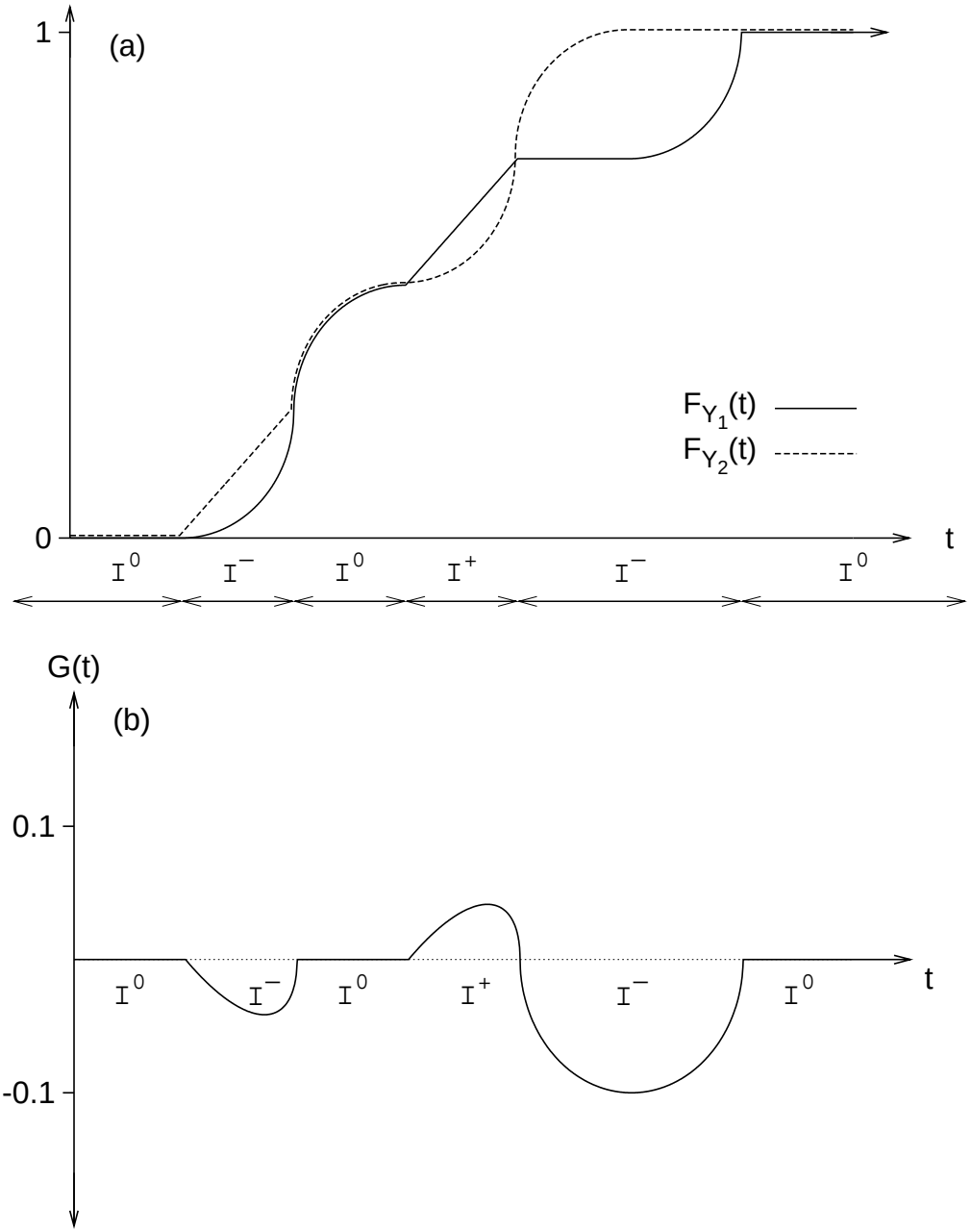


FIGURE C.5:

(a) Cumulative distribution functions, $F_{Y_1}(t)$ and $F_{Y_2}(t)$, of two random variables, $Y_1$ and $Y_2$, that are not stochastically ordered ($Y_1 \overset{st}{\not<} Y_2$). (The functions are offset slightly where they are equal.) (b) The difference function $G(t) = F_{Y_1}(t) - F_{Y_2}(t)$. Intervals defined as $I^+$, $I^-$ and $I^0$ in the text are also shown.

**Lemma 1**

$$\int_{I^+} G(t)\,dt \gtreqless \left|\int_{I^-} G(t)\,dt\right| \quad \Leftrightarrow \quad \int_{t=-\infty}^{\infty} G(t)\,dt \gtreqless 0$$

$$\Leftrightarrow \quad E(Y_1) \lesseqgtr E(Y_2)\,.$$

(Note that the direction of the inequalities changes from one line to the next in Lemma 1.) Which of $\int_{I^+} G(t)\,dt$ and $\int_{I^-} G(t)\,dt$ has the greatest *absolute* value determines the order of $E(Y_1)$ and $E(Y_2)$. The integrals could also be of equal magnitude but of opposite sign, which implies that that $E(Y_1) = E(Y_2)$. Lemma 1 holds for *any* $Y_1$ and $Y_2$ whose expectations exist and are finite. It follows directly from Equations C.10 and C.17, and is true regardless of whether $Y_1$ and $Y_2$ are stochastically ordered or not.

Let $h$, $R_1 = h(Y_1)$, $R_2 = h(Y_2)$ and $G^*(s)$ be defined as for Theorem 1. Considering the mutual domain of $R_1$ and $R_2$, let $J^+$, $J^-$ and $J^0$ respectively denote the unions of intervals over which $G^*(s)$ is greater than, less than or equal to zero. Let $\int_{J^+} G^*(s)\,ds$ denote the integral of $G^*(s)$ with respect to $s$ taken over the intervals that make up $J^+$, and similarly for integrals over $J^-$ and $J^0$. Since $h$ is an s.m.i. transform, then there is a simple relationship between $I^+$ and $J^+$, between $I^-$ and $J^-$, and between $I^0$ and $J^0$: the boundaries of a $J$ interval are defined by the transform $h$ being applied to the boundaries of a $I$ interval.

**Lemma 2**

$$\int_{s=-\infty}^{\infty} G^*(s)\,ds \gtreqless 0 \qquad \Leftrightarrow \qquad E(R_1) \lesseqgtr E(R_2)$$

$$\textit{i.e. that} \quad E(h(Y_1)) \lesseqgtr E(h(Y_2))$$

(Like in Lemma 1, note the change of direction of the inequalities.) Lemma 2 follows from Lemma 1 except applied to $R_1$, $R_2$ and $G^*(s)$, rather than $Y_1$, $Y_2$ and $G(t)$. It is true regardless of whether $R_1$ and $R_2$ are stochastically ordered or not.

**Theorem 2** *Let $h$, $Y_1$ and $Y_2$ be defined as for Theorem 1, such that $E(h(Y_1))$ and $E(h(Y_2))$ both exist and are finite. If $Y_1 \overset{st}{\not\leq} Y_2$, $Y_1 \overset{st}{\not\geq} Y_2$ and $Y_1 \neq Y_2$, then regardless of the order of $E(Y_1)$ and $E(Y_2)$, it is* always *possible to choose a strictly monotonic increasing transform, $h$, such that $E(h(Y_1))$ is either less than, greater than, or equal to $E(h(Y_2))$.*

**Proof.** Let $Y_1$ and $Y_2$ be such that $Y_1 \overset{st}{\not<} Y_2$, $Y_2 \overset{st}{\not<} Y_1$ and $Y_1 \neq Y_2$, meaning that $F_{Y_1}(t) > F_{Y_2}(t)$ for some intervals in $\mathbb{R}$ while $F_{Y_1}(t) < F_{Y_2}(t)$ for other intervals. Consequently, $G(t) > 0$ for some intervals and $G(t) < 0$ for others, like in Figure C.5(b).

The proof of Theorem 2 relies on describing an s.m.i. transform, $h$, that will either maintain or reverse (as desired) the order of $E(R_1)$ and $E(R_2)$ relative to the order of $E(Y_1)$ and $E(Y_2)$. It is assumed that the expectations involved in the proof exist and are finite, meaning that the integrals that are involved converge. The transform that is proposed here is only one possible transform that will suffice to prove the theorem. Many other transforms will also achieve the same end.

Choose a continuous s.m.i. transform $h(t)$ that is composed entirely of joined line segments, where different segments can have different slopes and where all of the slopes are positive (since $h$ is an increasing function). The segments are defined so they only join at the values of $t$ defined by boundary points between $I^+$, $I^-$ and $I^0$. Let the slope of all segments comprising $I^+$ be set to $k^+$, let the slope of all segments comprising $I^-$ be set to $k^-$, and let the slope of all segments comprising $I^0$ be set to $k^0$. The slopes, $k^+$, $k^-$ and $k^0$, are all positive constants, since $h(t)$ is s.m.i. Their values are chosen later, and it is important to note that $k^+$, $k^-$ and $k^0$ are independent of $Y_1$ and $Y_2$.

From Equations C.10 and C.17, and from Lemma 1 (and with attention to the direction of the inequalities), it follows that

$$E(Y_1) \underset{>}{\overset{<}{\equiv}} E(Y_2) \quad \Leftrightarrow \quad E(Y_1) - E(Y_2) \underset{>}{\overset{<}{\equiv}} 0 \tag{C.18}$$

$$\Leftrightarrow \quad -\int_{t=-\infty}^{\infty} G(t)\,dt \underset{>}{\overset{<}{\equiv}} 0 \tag{C.19}$$

$$\Leftrightarrow \quad \int_{t=-\infty}^{\infty} G(t)\,dt \underset{<}{\overset{>}{\equiv}} 0$$

$$\Leftrightarrow \quad \int_{I^+} G(t)\,dt - \left|\int_{I^-} G(t)\,dt\right| \underset{<}{\overset{>}{\equiv}} 0$$

$$\Leftrightarrow \quad \int_{I^+} G(t)\,dt \underset{<}{\overset{>}{\equiv}} \left|\int_{I^-} G(t)\,dt\right|$$

$$\Leftrightarrow \quad \frac{\int_{I^+} G(t)\,dt}{\left|\int_{I^-} G(t)\,dt\right|} \underset{<}{\overset{>}{\equiv}} 1. \tag{C.20}$$

Similarly, from Lemma 2

$$E(R_1) \underset{>}{\overset{<}{\equiv}} E(R_2) \quad \Leftrightarrow \quad \frac{\int_{J^+} G^*(s)\,ds}{\left|\int_{J^-} G^*(s)\,ds\right|} \underset{<}{\overset{>}{\equiv}} 1. \tag{C.21}$$

Equations C.20 and C.21 are clearly similar in form. Since $G^*(s) = G(t)\ \forall\, s,t : s = h(t)$

(Equation C.13), then the main difference between the related integrals in the two equations is the domain of integration. It is shown in Appendix B that an increasing piecewise-linear rescaling of a function *domain* by a transform such as $h$ results in an integral that is scaled by the slope of the line segments in the transform. In the current example, the transform only applies to each of $I^+$, $I^-$ and $I^0$ separately, which are regions over which the slope of the transform has the same value. The possibility that any of these domains of integration is the union of disjoint subsets of $\mathbb{R}$ is taken into account in Appendix B. Specifically, the results in Appendix B show that

$$\int_{J^+} G^*(s)\,ds = k^+ \int_{I^+} G(t)\,dt \tag{C.22}$$

and

$$\left|\int_{J^-} G^*(s)\,ds\right| = k^- \left|\int_{I^-} G(t)\,dt\right|, \tag{C.23}$$

with

$$\begin{aligned}\int_{J^0} G^*(s)\,ds &= k^0 \int_{I^0} G(t)\,dt \\ &= 0,\end{aligned} \tag{C.24}$$

where Equation C.24 holds by the definitions of $I^0$ and $J^0$. Since the value of $k^0$ does not affect Equation C.24, $k^0$ can be set arbitrarily to one, and then ignored. Equations C.22 and C.23 together imply that

$$\frac{\int_{J^+} G^*(s)\,ds}{\left|\int_{J^-} G^*(s)\,ds\right|} = \frac{k^+ \int_{I^+} G(t)\,dt}{k^- \left|\int_{I^-} G(t)\,dt\right|}. \tag{C.25}$$

Substituting Equation C.25 into Equation C.21 implies that

$$E(R_1) \lessgtr E(R_2) \quad \Leftrightarrow \quad \frac{k^+ \int_{I^+} G(t)\,dt}{k^- \left|\int_{I^-} G(t)\,dt\right|} \gtreqless 1 \tag{C.26}$$

$$\Leftrightarrow \quad \frac{k^-}{k^+} \lessgtr \gamma, \tag{C.27}$$

where

$$\gamma \stackrel{def}{=} \frac{\int_{I^+} G(t)\,dt}{\left|\int_{I^-} G(t)\,dt\right|}$$

is defined here for convenience only. (Note that the ratio of $k$-values is $\frac{k^+}{k^-}$ in Equation C.26, which changes to $\frac{k^-}{k^+}$ in Equation C.27.) The assumptions given for Theorem 2, that $Y_1$ and $Y_2$ are not stochastically ordered and that the integrals in Equation C.25 converge

(because $E(Y_1)$ and $E(Y_2)$ exist), imply that $\gamma$ is a fixed *finite*, positive value for any given pair of random variables, $Y_1$ and $Y_2$.

In order to prove Theorem 2, separate conditions need to be shown under which $E(h(Y_1)) < E(h(Y_2))$, $E(h(Y_1)) = E(h(Y_2))$, and $E(h(Y_1)) > E(h(Y_2))$, regardless of the order of $E(Y_1)$ and $E(Y_2)$. Equation C.27 does just that. With $\gamma$ being a fixed positive value, Equation C.27 states that if $k^+$ and $k^-$ are *chosen* such that $k^-/k^+ < \gamma$, then $E(h(Y_1)) < E(h(Y_2))$; if $k^-/k^+ > \gamma$, then $E(h(Y_1)) > E(h(Y_2))$; and if $k^-/k^+ = \gamma$ then $E(h(Y_1)) = E(h(Y_2))$. In this way, $k^+$ and $k^-$ can be chosen so the order of $E(h(Y_1))$ and $E(h(Y_2))$ is either the same as or opposite to the order of $E(Y_1)$ and $E(Y_2)$. This completes the proof of Theorem 2. Q.E.D.

Several points follow directly from the proof of Theorem 2.

- Theorem 2 only applies when $Y_1$ and $Y_2$ are *not* stochastically ordered. If either $Y_1 \overset{st}{<} Y_2$ holds, or $Y_1 \overset{st}{>} Y_2$ holds, then Theorem 1 applies instead of Theorem 2.
- The value of $k^0$ does not matter at all, only the ratio of $k^-$ and $k^+$ is important.
- The results of Theorem 2 hold for an unlimited number of transform functions. These functions can be either linear or non-linear, as long as Equation C.21 is satisfied. The results are a consequence of defining $h$ such that the $I^+$ and $I^-$ regions of $\mathbb{R}$ are sufficiently stretched or compressed (as desired) into $J^+$ and $J^-$ respectively. This was achieved in the proof by setting $k^+$ and $k^-$.

### C.2.1 Non-strict inequalities in stochastic ordering

The results presented so far were stated for the case of *strict* stochastic ordering. They also apply, with modification, to *non-strict* stochastic ordering (Definition 2). In Theorem 1, strict stochastic and numerical inequalities (e.g. "$\overset{st}{<}$" and "$<$") may be replaced by non-strict inequalities (e.g. "$\overset{st}{\leqslant}$" and "$\leqslant$") and the modified theorem will still hold.

Assuming that the expectations involved exist, then Equations C.1 to C.10, C.13 and C.15 all hold *regardless* of whether $Y_1$ and $Y_2$ are stochastically ordered or not, (This includes non-strict stochastic ordering, so these equations do not need modification.) If $Y_1 \overset{st}{\leqslant} Y_2$ in the statement of Theorem 1, then Equation C.12 (and its following text), Equation C.14 (both inequalities), and Equation C.16 (and its following text) can all be changed from strict inequalities to non-strict inequalities. The non-strict inequalities follow through to the conclusion of the proof on p. 268.

Non-strict ordering may also be applied to Corollaries 2, 3 and 4, and leads to the following conclusion:

**Corollary 5** *Corollaries 3 and 4 hold for any arbitrary combination of strict and non-strict inequalities in the ordering sequence, where the ordering is stochastic for the random variables and numerical for their expected values. For example, if* $Y_1 \overset{st}{<} Y_2 \overset{st}{\leqslant} Y_3 \overset{st}{<} \ldots$, *then* $h(Y_1) \overset{st}{<} h(Y_2) \overset{st}{\leqslant} h(Y_3) \overset{st}{<} \ldots$, *and consequently,* $E(Y_1) < E(Y_2) \leqslant E(Y_3) < \ldots$ *and* $E(h(Y_1)) < E(h(Y_2)) \leqslant E(h(Y_3)) < \ldots$

The combination of strict and non-strict ordering of more than two random variables was first dealt with in Corollary 1. Corollary 5 results from Corollary 1, and by application of both strict and non-strict versions of Theorem 1 to successive overlapping pairs of the random variables involved.

In Theorem 2, the conditions that $Y_1 \overset{st}{\not<} Y_2$ and $Y_1 \overset{st}{\not>} Y_2$ (along with $Y_1 \neq Y_2$) can be replaced by their non-strict counterparts, $Y_1 \overset{st}{\not\leqslant} Y_2$ and $Y_1 \overset{st}{\not\geqslant} Y_2$, without changing the statement of the theorem. This is because $Y_1 \overset{st}{\not\leqslant} Y_2$ implies that $Y_1 \overset{st}{\not<} Y_2$ and $Y_1 \neq Y_2$, and $Y_1 \overset{st}{\not\geqslant} Y_2$ implies that $Y_1 \overset{st}{\not>} Y_2$ and $Y_1 \neq Y_2$.

## C.3 The effect of step function transforms of random variables on stochastic ordering, and on the ordering of expected values

The preceding section dealt with s.m.i. transforms of random variables, their effect on stochastic ordering, and subsequent effects on the ordering of expected values. This section deals with the effect of *step function transforms* on stochastic ordering, and on the ordering of expected values.

Let $R_j$ be a continuous, discrete, or mixed random variable. Let $\Lambda$ be a monotonic increasing *step function* defined on a partition $\Psi_R$ of the real number line, $\mathbb{R}$, and let $Q_j = \Lambda(R_j)$ be a new random variable. $Q_j$ is always discrete, regardless of the form of $R_j$. The domain of $Q_j$ can be any countable set of real numbers, $\Omega_Q = \{\ldots, q_0, q_1, q_2, \ldots\}$, where $q_{\ell-1} < q_\ell$. $\Omega_Q$ can be either finite or infinite, but it has at least two members (since it is defined from a step function). The partition $\Psi_R$ is a countable set of real-valued numbers, $\Psi_R = \{\ldots, r_0, r_1, r_2, \ldots\}$, $r_{\ell-1} < r_\ell$, where the number of values in $\Omega_Q$ is one more than the number of members in $\Psi_R$. Each interval[5] $(r_{\ell-1}, r_\ell]$ defined by the partition is assigned a value in $\Omega_Q$ via the left-continuous step function

$$\Lambda(r) \quad = \quad q_\ell, \quad \text{for } r: \; r_{\ell-1} < r \leqslant r_\ell, \tag{C.28}$$

An example of $\Lambda(r)$ is sketched in Figure C.6. If $\Psi_R$ is bounded below, then take $r_{\ell-1}$ to be $-\infty$ for the smallest value of $\ell$. Similarly, if $\Psi_R$ is bounded above, then take $r_\ell$ to be $\infty$ for the largest value of $\ell$.

[5]Square brackets denote inclusion, and round brackets denote exclusion.

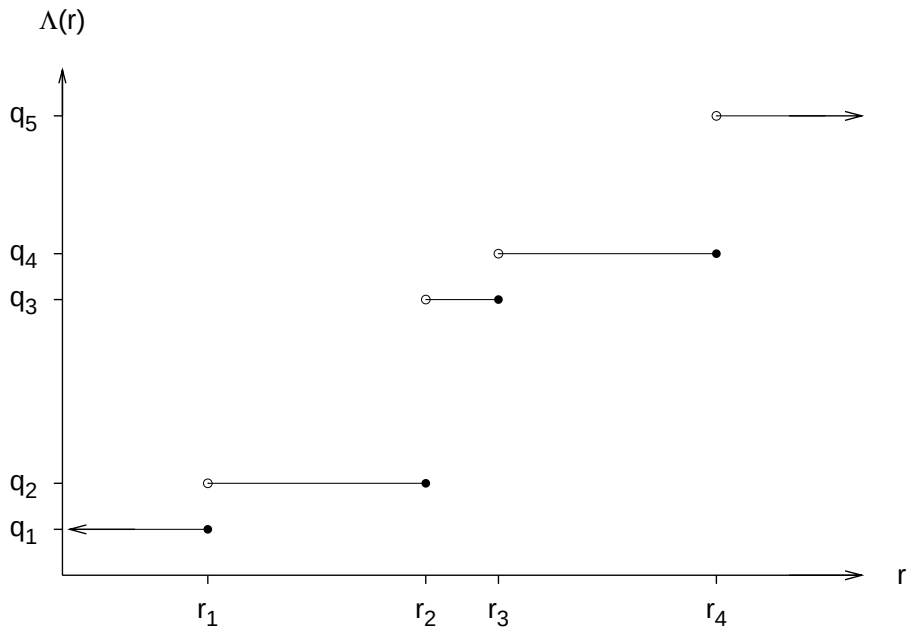


FIGURE C.6:
Example of a monotonic increasing step function $\Lambda(r)$ that maps intervals defined by a partition $\Psi_R = \{\dots, r_0, r_1, r_2, \dots\}$ onto discrete values lying in the set $\Omega_Q = \{\dots, q_0, q_1, q_2, \dots\}$. (Filled circles denote inclusion. Empty circles denote exclusion.) Note that $\Lambda(r)$ is *not* a cumulative distribution function.

For the $\ell^{th}$ interval, the effect of $\Lambda$ is to mass the probability of $R_j$ lying in the interval $(r_{\ell-1}, r_\ell]$ and assign it to $Q_j$ at the value $q_\ell$. In terms of Riemann-Stieltjes integrals, the probability mass function of $Q_j$ is

$$\begin{aligned} P(Q_j = q_\ell) &= P(r_{\ell-1} < R_j \leqslant r_\ell) \\ &= \int_{r=r_{\ell-1}}^{r_\ell} dF_{R_j}(r), \end{aligned}$$

for any $R_j$ that is either continuous, discrete or mixed. If $R_j$ is continuous and its probability density function is $f_{R_j}(r) = \frac{d}{dr} F_{R_j}(r)$, then

$$P(Q_j = q_\ell) = \int_{r=r_{\ell-1}}^{r_\ell} f_{R_j}(r)\, dr.$$

If $R_j$ is discrete and its probability mass function is $P(R_j = r)$, then

$$P(Q_j = q_\ell) = \sum_{\substack{all\ r: \\ r_{\ell-1} < r \leqslant r_\ell}} P(R_j = r).$$

The cumulative distribution function of $Q_j$ is

$$\begin{aligned} F_{Q_j}(q) &= P(Q_j \leqslant q) \\ &= \sum_{\substack{all\ \ell: \\ q_\ell \leqslant q}} P(Q_j = q_\ell). \end{aligned} \tag{C.29}$$

Equation C.29 defines $F_{Q_j}(q)$ for all $q \in \mathbb{R}$. When $q = q_\ell$, the value of the cumulative distribution function of $Q_j$ matches that of the cumulative distribution function of $R_j$, that is,

$$\begin{aligned} F_{Q_j}(q_\ell) &= P(Q_j \leqslant q_\ell) \\ &= P(R_j \leqslant r_\ell) \\ &= F_{R_j}(r_\ell), \end{aligned} \tag{C.30}$$

(noting that $r_\ell$ for the maximum $\ell$ is taken to be $\infty$.) If $R_j$ is continuous, then

$$F_{Q_j}(q_\ell) = \int_{r=-\infty}^{r_\ell} f_{R_j}(r)\, dr,$$

and if $R_j$ is discrete, then

$$F_{Q_j}(q_\ell) = \sum_{\substack{all\ r: \\ r \leqslant r_\ell}} P(R_j = r).$$

Next, consider $R_j$ to be one of a family of continuous, discrete, or mixed random variables, $\{R_1, R_2, R_3 \dots\}$, with $\Lambda$ defined on the mutual domain of all of them. The partition $\Psi_R$ is not just specific to the $j^{th}$ $R$-variable, but may be applied equally to any such $R$-variables lying on the same axis.

**Theorem 3** *Let $R_j$ and $R_k$ be any two of the $R$-variables in a family of random variables $\{R_1, R_2, R_3 \dots\}$. Let the step function $\Lambda$, and its related partition $\Psi_R$, be defined as above for Equation C.28, and let $Q_j = \Lambda(R_j)$ and $Q_k = \Lambda(R_k)$. If $R_j \overset{st}{<} R_k$, then $Q_j \overset{st}{\leqslant} Q_k$. Whether $Q_j \overset{st}{<} Q_k$ holds, or $Q_j = Q_k$ holds, depends on how $\Psi_R$ partitions the mutual domain of $R_j$ and $R_k$. If $F_{R_j}(r_\ell) > F_{R_k}(r_\ell)$ for any $r_\ell \in \Psi_R$, then $Q_j \overset{st}{<} Q_k$, otherwise $Q_j = Q_k$.*

**Proof.** For the random variables, $R_k$ and $Q_k$,

$$F_{Q_k}(q_\ell) \quad = \quad F_{R_k}(r_\ell) \tag{C.31}$$

from Equation C.30. If $R_j \overset{st}{<} R_k$, then

$$F_{R_j}(r) \quad \geqslant \quad F_{R_k}(r), \quad \forall\, r \in \mathbb{R}, \tag{C.32}$$

with a strict inequality holding over some non-zero interval in $\mathbb{R}$ (because *strict* stochastic ordering is involved). Together, Equations C.30, C.31 and C.32 imply that if $R_j \overset{st}{<} R_k$, then

$$F_{Q_j}(q_\ell) \quad \geqslant \quad F_{Q_k}(q_\ell)$$

for all $q_\ell \in \Omega_Q$. Since $Q_j$ and $Q_k$ are discrete, then

$$F_{Q_j}(q) \quad \geqslant \quad F_{Q_k}(q) \tag{C.33}$$

holds for all $q \in \mathbb{R}$ (including all $q_\ell \in \Omega_Q$). By Definition 2, Equation C.33 implies that $Q_j \overset{st}{\leqslant} Q_k$, which proves the main conclusion of Theorem 3.

**Conditions for strict stochastic ordering of $Q_j$ and $Q_k$**

The condition $Q_j \overset{st}{\leqslant} Q_k$ implies that either $Q_j \overset{st}{<} Q_k$ or $Q_j = Q_k$. The condition $Q_j \overset{st}{<} Q_k$ holds if there is a non-zero interval in $\mathbb{R}$ in which $F_{Q_j}(q) > F_{Q_k}(q)$. Such an interval may or may not exist, depending on the specifics of $F_{R_j}(q)$, $F_{R_k}(q)$ and $\Psi_R$, in particular, how the partition $\Psi_R$ cuts across any intervals for which the *strict* inequality $F_{R_j}(r) > F_{R_k}(r)$ holds.

In order to demonstrate this claim, let $\alpha$ be a value (not necessarily in $\Psi_R$) where $F_{R_j}$ and $F_{R_k}$ change from being equal to being non-equal, and let $\beta$ be the next largest value at which $F_{R_j}$ and $F_{R_k}$ change back from being unequal to being equal (where $\alpha$ and $\beta$ need not be finite). For reference purposes, any non-zero interval between successive points at which $F_{R_j}(r) \neq F_{R_k}(r)$, such as $[\alpha, \beta)$ or $(\alpha, \beta)$, is referred to as an *unbroken interval*, an example of which is sketched in Figure C.7. By definition, $F_{R_j}(r) \neq F_{R_k}(r)$ for all $r \in [\alpha, \beta)$. To remove any need to treat continuous random variables separately from discrete or mixed random variables, the point $r = \alpha$ is included in the unbroken interval in the case when $R_j$ and $R_k$ are both continuous random variables (even though $F_{R_j}(\alpha) = F_{R_k}(\alpha)$). This is done for convenience, and does not affect the conclusions.

Consider a case where $\alpha$ and $\beta$ are both finite and $R_j$ and $R_k$ are such that there is only *one* unbroken interval $[\alpha, \beta)$ in $\mathbb{R}$. If $\Psi_R$ is such that $[\alpha, \beta)$ falls entirely within the *same single interval* defined via $\Psi_R$, then $Q_j = Q_k$. If $[\alpha, \beta)$ falls across more than one

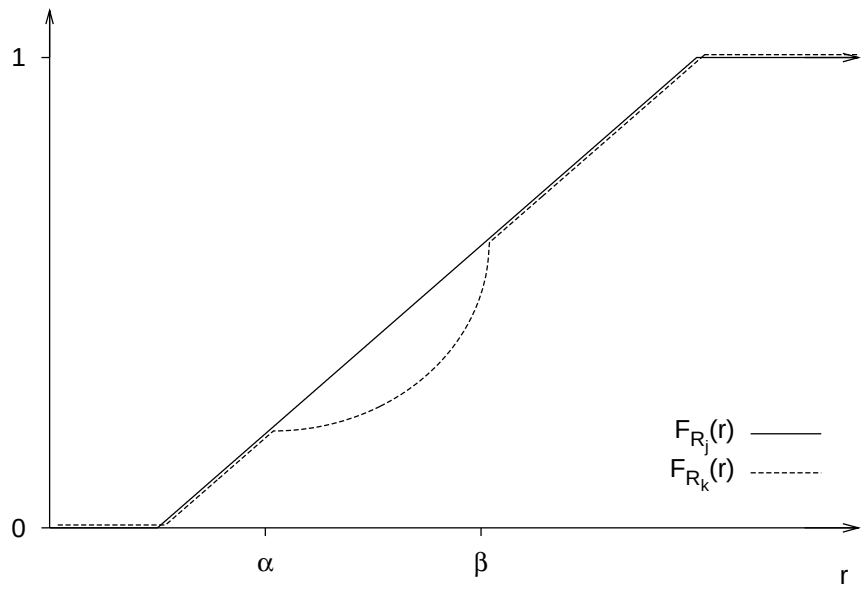


FIGURE C.7:
Cumulative distribution functions, $F_{R_j}(r)$ and $F_{R_k}(r)$, for two stochastically ordered random variables, $R_j$ and $R_k$, where $R_j \overset{st}{<} R_k$, showing boundary points, $\alpha$ and $\beta$, of the unbroken interval over which $F_{R_j}(r)$ and $F_{R_k}(r)$ differ. (The functions are offset slightly where they are equal.)

interval defined via $\Psi_R$, then $Q_j \overset{st}{<} Q_k$.

*First*, consider if $[\alpha, \beta)$ falls entirely within the same interval, say the $\ell^{th}$ interval, $(r_{\ell-1}, r_\ell]$. In that case

$$r_{\ell-1} < \alpha < \beta \leqslant r_\ell, \tag{C.34}$$

This is a trivial case, because the step function masses all of the probability of $R_j$ and $R_k$ at the same point, namely at $q = q_\ell$. Specifically, $P(Q_j = q_\ell) = P(Q_k = q_\ell) = 1$, which implies that $Q_j = Q_k$.

*Second*, consider the case where $[\alpha, \beta)$ falls across more than one interval defined via $\Psi_R$, for example, the $\ell^{th}$ interval, $(r_{\ell-1}, r_\ell]$, so that

$$r_{\ell-1} < \alpha < r_\ell < \beta. \tag{C.35}$$

As before, $F_{R_j}(r) = F_{R_k}(r)$ for all $r < \alpha$, including for $r < r_{\ell-1}$. In this case, however, $r_\ell$ falls within $[\alpha, \beta)$, and so by the definition of $[\alpha, \beta)$, $F_{R_j}(r_\ell) > F_{R_k}(r_\ell)$. By Equations C.30 and C.31, this implies that $F_{Q_j}(q_\ell) > F_{Q_k}(q_\ell)$ holds for at least one value, $q_\ell \in \Omega_Q$. Since $F_{Q_j}$ and $F_{Q_k}$ are right-continuous cumulative distribution functions, then

$F_{Q_j}(q) > F_{Q_k}(q)$ holds for all $q \in [q_\ell, q_{\ell+1})$, which, along with $Q_j \overset{st}{\leqslant} Q_k$ (from Equation C.33), implies that $Q_j \overset{st}{<} Q_k$.

These results show that whether or not $Q_j$ and $Q_k$ follow a strict stochastic ordering depends on where the intervals of $\Psi_R$ fall with respect to an interval over which $F_{R_j}$ and $F_{R_k}$ are different.

The results can be extended to an arbitrary number of unbroken intervals. By Definition 1, if $R_j \overset{st}{<} R_k$, then there must be at least one unbroken interval for which $F_{R_j}(r) \neq F_{R_k}(r)$, such as $[\alpha, \beta)$, and in general, there may be more. Let the set of such unbroken intervals be $\{[\alpha_1, \beta_1), [\alpha_2, \beta_2), \dots\}$. The unbroken intervals are defined *independently of* the partition intervals $\{ \dots (r_1, r_2], (r_2, r_3], (r_3, r_4] \dots \}$, because the unbroken intervals result from the specific distributions of $R_j$ and $R_k$, whereas the partition intervals are defined by the partition $\Psi_R$.

When $R_j \overset{st}{<} R_k$, $Q_j = Q_k$ occurs if and only if *both* boundary points ($\alpha_i$ and $\beta_i$) of *each and every* unbroken interval lie entirely within a partition interval (but different unbroken intervals can lie within different partition intervals). That is to say, for every index value $i$ there is an index value $\ell$ such that

$$[\alpha_i, \beta_i) \in (r_{\ell-1}, r_\ell],$$

(i.e. that Equation C.34 holds for each $[\alpha_i, \beta_i)$ ). If these conditions hold, then strict stochastic ordering of $R_j$ and $R_k$ results in the *non*-strict stochastic ordering of $Q_j$ and $Q_k$.

If, *for any* such unbroken interval, $[\alpha_i, \beta_i)$, there is a single value $r_\ell \in \Psi_R$ (for any index value $\ell$) such that $\alpha_i < r_\ell < \beta_i$, then $F_{Q_j}(q_\ell) > F_{Q_k}(q_\ell)$ over a non-zero interval, $[q_\ell, q_{\ell+1})$, which implies that $Q_j \overset{st}{<} Q_k$, rather than $Q_j \overset{st}{\leqslant} Q_k$.

This completes the proof of Theorem 3.

**In summary,** if $R_j \overset{st}{<} R_k$, then (at worst) $Q_j \overset{st}{\leqslant} Q_k$ holds and so

$$R_j \overset{st}{<} R_k \quad \Rightarrow \quad Q_j \overset{st}{\leqslant} Q_k \tag{C.36}$$

$$\Rightarrow \quad E(Q_j) \leqslant E(Q_k). \tag{C.37}$$

Equation C.37 follows from Equation C.36, by applying Theorem 1 (modified for non-strict stochastic ordering according to Section C.2.1) to $Q_j$ and $Q_k$.

If the partition $\Psi_R$ is such that *at least one* value $r_\ell \in \Psi_R$ falls *within* an unbroken

interval $[\alpha_i, \beta_i)$, then

$$R_j \overset{st}{<} R_k \quad \Rightarrow \quad Q_j \overset{st}{<} Q_k \tag{C.38}$$

$$\Rightarrow \quad E(Q_j) < E(Q_k), \tag{C.39}$$

where Equation C.39 follows from Equation C.38 by Theorem 1.

**Special cases.** If $R_j$ and $R_k$ are such that $F_{R_j}(r) > F_{R_k}(r)$ holds over the entire real number line (for example if $R_j$ and $R_k$ were both Gaussian of equal variance but different means), then Equations C.38 and C.39 *must* apply. Similar results hold for random variables for which $F_{R_j}(r) > F_{R_k}(r)$ applies over their entire mutual domain only (which could just be a subset of $\mathbb{R}$), as long as $\Psi_R$ and $\Lambda$ involve a partition of the mutual domain of $R_j$ and $R_k$. When $\Psi_R$ was first defined (p. 276), it was described as a partition of the entire real number line, $\mathbb{R}$. In order to apply Theorem 3 or any consequent corollaries, $\Psi_R$ need only be defined as a partition of the mutual domain of any $R$-variables that are involved. If the entire mutual domain falls within a single partition interval defined by $\Psi_R$ (such as $(r_{\ell-1}, r_\ell]$ ), then all of the resulting $Q$-variables are singular because all probability for each one is massed at a single point.

*In practical terms*, Theorem 3 shows that if $R_j \overset{st}{<} R_k$, then the finer the partition, $\Psi_R$, the less likely it is that $Q_j = Q_k$ and the more likely it is that $Q_j \overset{st}{<} Q_k$. For an entire family of $R$-variables, as mentioned in Theorem 3, the more variables there are in the family, the less likely it would seem that a single partition applied to their mutual domain would result in of all of the $Q$-variables being equal to each other.

**Corollary 6**

$$R_j \overset{st}{\leqslant} R_k \quad \Rightarrow \quad Q_j \overset{st}{\leqslant} Q_k$$

$$\Rightarrow \quad E(Q_j) \leqslant E(Q_k).$$

Theorem 3 holds for when the random variables, $R_j$ and $R_k$, are *strictly* stochastically ordered. Corollary 6 says that the same results apply when $R_j$ and $R_k$ follow non-strict stochastic ordering ($R_j \overset{st}{\leqslant} R_k$). This can be seen by also considering when $R_j = R_k$, in which case $Q_j = Q_k$ and so $E(Q_j) = E(Q_k)$. That, together with Theorem 3, prove Corollary 6.

**Corollary 7** *Let $R_j$ and $R_k$ be any two random variables whose expectations exist and are finite. Let $\Lambda_1$, $\Lambda_2$, $\Lambda_3$ ... be left-continuous, monotonic increasing step functions, where $\Lambda_1$ is defined on the mutual domain of $R_j$ and $R_k$, $\Lambda_2$ is defined on the mutual domain of $\Lambda_1(R_j)$ and $\Lambda_1(R_k)$, $\Lambda_3$ is defined on the mutual domain of $\Lambda_2(\Lambda_1(R_j))$ and $\Lambda_2(\Lambda_1(R_k))$,*

*and so on. If either* $R_j \overset{st}{<} R_k$ *or* $R_j \overset{st}{\leqslant} R_k$*, then*

$$E[\ldots\Lambda_3(\Lambda_2(\Lambda_1(R_j)))] \quad \leqslant \quad E[\ldots\Lambda_3(\Lambda_2(\Lambda_1(R_k)))],$$

*if the expectations exist and are finite.*

**Proof.** Consider first the case of just two transforms, $\Lambda_1$ and $\Lambda_2$. Corollary 7 results then from the application of Theorem 3 (if $R_j \overset{st}{<} R_k$) or Corollary 6 (if $R_j \overset{st}{\leqslant} R_k$), followed by the application of Corollary 6 to the results, in either case. In the first application (using just $\Lambda_1$), where $R_j$ and $R_k$ *may* be strictly stochastically ordered, the resulting variables, $\Lambda_1(R_j)$ and $\Lambda_1(R_k)$, are ordered, but are not necessarily strictly ordered. Again from Corollary 6, if $\Lambda_1(R_j) \overset{st}{\leqslant} \Lambda_1(R_k)$ then $\Lambda_2(\Lambda_1(R_j)) \overset{st}{\leqslant} \Lambda_2(\Lambda_1(R_k))$, which implies that $E[\Lambda_2(\Lambda_1(R_j))] \leqslant E[\Lambda_2(\Lambda_1(R_k))]$. By extension, and by repeated application of Corollary 6, any number of monotonic increasing step function transforms can be nested to produce the desired result. Another way of viewing this is that a monotonic increasing step function of a monotonic increasing step function (i.e. $\Lambda_0(r) = \Lambda_2(\Lambda_1(r))$ is itself a monotonic increasing step function. The series can be extended to any number of step functions, $\Lambda_0(r) = \ldots\Lambda_3(\Lambda_2(\Lambda_1(r))$. Corollary 6 applied to $Q_j = \Lambda_0(R_j)$ and $Q_k = \Lambda_0(R_k)$ gives the desired result.

Even if $R_j \overset{st}{<} R_k$ held to start with in Corollary 7 (rather than $R_j \overset{st}{\leqslant} R_k$), Equation C.36 shows that strict ordering is not guaranteed to hold after the first transform, $\Lambda_1$, is applied, although it may hold, depending on the conditions outlined earlier in the proof of Theorem 3. In a series of nested transforms, $\ldots\Lambda_3(\Lambda_2(\Lambda_1(r))$, applied to $R_j$ and $R_k$ such that $R_j \overset{st}{<} R_k$, once strict stochastic order has been lost in the series, *it cannot be recovered* by further transformations.

**Corollary 8** *Assume there is a family of stochastically ordered R-variables,* $R_1, R_2, R_3 \ldots$*, in which any* $R_j$ *may be either continuous, discrete, or mixed. Let* $\Lambda$*,* $\Psi_R$ *and* $\Omega_Q$ *be defined as for Theorem 3, and let* $Q_j = \Lambda(R_j)$ *define a family of discrete Q-variables,* $Q_1, Q_2, Q_3 \ldots$ *If the R-variables are such that* $R_1 \overset{st}{<} R_2 \overset{st}{<} R_3 \overset{st}{<} \ldots$*, then the Q-variables are such that* $Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} Q_3 \overset{st}{\leqslant} \ldots$ *Strict stochastic ordering of the Q-variables is possible, but not guaranteed, in accordance with the conditions given in Theorem 3.*

Corollary 8 follows from the repeated application of Theorem 3 (and Equation C.36 in particular) to successive overlapping pairs of $R$-variables ($R_1$ and $R_2$, then $R_2$ and $R_3$, etc.). Note that the partition $\Psi_R$ and the step function $\Lambda$ are independent of the $R$-variables, and hence are the same for all pairings $R_j$ and $R_k$.

**Corollary 9** *For a family of R-variables and Q-variables that are defined and stochastically ordered as in Corollary 8, the ordering of the expected values of the Q-variables follows the stochastic ordering of the R-variables, if the expectations exist and are finite, so* $E(Q_1) \leqslant E(Q_2) \leqslant E(Q_3) \ldots$ *Strict numerical ordering of these expectations is possible, but not guaranteed, in accordance with the conditions given in Theorem 3.*

Corollary 9 follows from the repeated application of Theorem 3, and Equation C.37 in particular, to successive overlapping pairs of $R$-variables ($R_1$ and $R_2$, then $R_2$ and $R_3$, etc.).

The premises in Corollaries 8 and 9, that the sequence $R_1 \overset{st}{<} R_2 \overset{st}{<} R_3 \overset{st}{<} \ldots$ involves strict ordering, may be modified according to Corollary 5. So "$\overset{st}{<}$" may be arbitrarily replaced by "$\overset{st}{\leqslant}$" in the premises about the $R$-variables without altering the conclusions that $Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} Q_3 \overset{st}{\leqslant} \ldots$ and that $E(Q_1) \leqslant E(Q_2) \leqslant E(Q_3) \ldots$

## Summary

Theorem 1 and Corollaries 2, 3 and 4, about strict stochastic ordering following continuous s.m.i. transforms of $Y$-variables have their respective equivalents in Theorem 3 and Corollaries 7, 8 and 9, about non-strict stochastic ordering following monotonic increasing step function transforms of $R$-variables.

## C.4 Summary

Section C.1 showed that: (1) a set of stochastically ordered random variables results in a set of numerically ordered means, where the ordering of means follows the stochastic ordering of the random variables, (2) stochastic ordering is unaffected by s.m.i. transforms of random variables, and (3) the order of means of the transformed stochastically ordered random variables follows the order of means of the original random variables. Theorem 1 and the corollaries in Section C.1 hold for strict stochastic ordering, and with modification, also to non-strict stochastic ordering.

Section C.2 showed that s.m.i. transforms of a pair of random variables that are not stochastically ordered results in new pair of random variables that are also not stochastically ordered. Given the original pair of random variables, it is always possible to invent an s.m.i. transform that will put the means of the transformed random variables in either order, or set them to equal each other. This is true, regardless of the order of the means of the original random variables. (It is of particular interest in Chapter 5 to consider cases when the order of means following the transform is *opposite* to that of the means prior to the transform.)

Section C.3 dealt with monotonic increasing step function transforms of stochastically ordered random variables. A step function is associated with a partition of the real number line, and the transform effectively masses the probability associated with each partition interval onto a single value that lies in the range of the step function. Step function transforms of stochastically ordered random variables maintain stochastic ordering. They *may* preserve strict stochastic ordering, but unlike s.m.i. transforms, step function transforms are not *guaranteed* to do so. Whether or not strict ordering is maintained depends on how the points of discontinuity in the step function lie with respect to intervals over which cumulative distribution functions differ across random variables.

The relationships among the $Y$, $R$ and $Q$-variables and their expectations may be summarised as

$$\begin{array}{ccccc} Y_1 \overset{st}{<} Y_2 \overset{st}{<} \ldots & \Leftrightarrow & R_1 \overset{st}{<} R_2 \overset{st}{<} \ldots & \Rightarrow & Q_1 \overset{st}{\leqslant} Q_2 \overset{st}{\leqslant} \ldots \\ \Downarrow & & \Downarrow & & \Downarrow \\ E(Y_1) < E(Y_2) < \ldots & & E(R_1) < E(R_2) < \ldots & & E(Q_1) \leqslant E(Q_2) \leqslant \ldots \end{array}$$

where the arrows show the direction of implication. Note that the ordering of expected values does not imply anything about stochastic ordering, nor does it imply anything about about the expected values of any transformed random variables.

Once strict stochastic ordering has been lost due to partitioning, or transforming by a step function, it cannot be recovered by further transformations. *In practical terms*, the finer the partition $\Psi_R$, the less likely it is that strict stochastic ordering will be lost, and the more likely it is that strict stochastic ordering will be maintained. This is of benefit when a *strict* ordering of transformed means is desirable.

# Appendix D

# Weighted sums across stochastically ordered sets of random variables

The aim of this appendix is to show that if there are multiple sets of stochastically ordered random variables, where each set follows the same stochastic ordering, then weighted sums of random variables taken across the sets are also stochastically ordered. The stochastic ordering of the weighted sums is the same as the ordering within each of the sets. This result is unaffected by any s.m.i. transforms that may be applied separately to each set. The result is modified when monotonic increasing step function transforms are applied separately to each set, because strict stochastic ordering may be lost.

Like in Appendix C, the results in this appendix primarily deal with stochastic ordering of random variables, but they can also be applied to *sample sets* of values, where each set is sampled from a different random variable. In that case, *sample* cumulative distribution functions take the place of the cumulative distribution functions in the derivations and results.

The material here follows on from the contents of Appendix C, and subsumes and incorporates the results that are presented there. The notation and interpretation is consistent with what is in Appendix C, but is extended here to cover multiple sets of random variables. The results in this appendix are applied in Chapter 5 to the theory of GOC analysis, particularly when it is extended to cover different observers who may use different decision axes and may have different transfer functions.

The main result in this appendix, Theorem 4, is about weighted sums of stochastically ordered random variables, the proof of which takes up most of Section D.1. Section D.2 presents corollaries that are general extentions to Theorem 4. Section D.3 extends the results to include weighted sums of s.m.i.-transformed random variables.

## D.1 Weighted sums of stochastically ordered random variables

Let there be two sets of random variables, $\{U_{1,1},\ U_{1,2},\ U_{1,3}\ \dots\ \}$ and $\{U_{2,1}, U_{2,2}, U_{2,3}\dots\}$, two sets of $x$-values, $\{x_{1,1}, x_{1,2}, x_{1,3}\dots\}$ and $\{x_{2,1}, x_{2,2}, x_{2,3}\dots\}$, and two types of mixture processes, $\oplus_1$ and $\oplus_2$. There are two groupings of variables, values and functions, each of which is referred to as a *division*. In the context of GOC analysis, each division is associated with an individual observer, where, for the same stimulus set, each observer has his or her own decision axis, unique noise distributions and form of unique and common noise mixing.

The $U$-variables may or may not be identically distributed, both within sets and across divisions. The $x$-values are real-valued and are not random variables, although they may be interpreted as sample values from some random variable ($X$). The $x$-values are generally different from each other within a set, and may or may not be identical across divisions (for the same index values). The two mixture processes may or may not be the same as each other.

Let $\{Y_{1,1}, Y_{1,2}, Y_{1,3}\dots\}$ and $\{Y_{2,1}, Y_{2,2}, Y_{2,3}\dots\}$ be two sets of $Y$-variables that are derived from the $x$-values, $U$-variables and mixture processes according to

$$Y_{\eta,\varepsilon} \quad = \quad x_{\eta,\varepsilon} \oplus_\eta U_{\eta,\varepsilon}, \tag{D.1}$$

where $\eta = 1$ or 2 denotes the division, and $\varepsilon = 1, 2, 3\dots$ denotes the index-value within each set. The $j^{th}$ and the $k^{th}$ $Y$-variables for the first division are $Y_{1,j} = x_{1,j} \oplus_1 U_{1,j}$ and $Y_{1,k} = x_{1,k} \oplus_1 U_{1,k}$, while those for the second division are $Y_{2,j} = x_{2,j} \oplus_2 U_{2,j}$ and $Y_{2,k} = x_{2,k} \oplus_2 U_{2,k}$. The $Y$-variables may or may not be identically distributed *across divisions* for the same second index value $\varepsilon$. They may also be identically distributed within divisions, although they generally would not be. The $Y$-variables need not be independent of each other either.

**Theorem 4** *Let $a_1$ and $a_2$ be any two positive constants. If $Y_{1,j} \overset{st}{<} Y_{1,k}$ and $Y_{2,j} \overset{st}{<} Y_{2,k}$, then $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{<} (a_1 Y_{1,k} + a_2 Y_{2,k})$.*

**Proof.** Let

$$\Theta_j \quad = \quad a_1 Y_{1,j} + a_2 Y_{2,j} \tag{D.2}$$

be the $j^{th}$ weighted sum of random variables, where the summation is taken across divisions for a fixed set-index value, $j$. Similarly, let

$$\Theta_k \quad = \quad a_1 Y_{1,k} + a_2 Y_{2,k}$$

be the equivalent random variable for the set-index value $k$. The conclusion of Theorem 4 is that $\Theta_j \overset{st}{<} \Theta_k$.

The cumulative distribution function for $\Theta_j$ is

$$\begin{aligned}
F_{\Theta_j}(\theta) &= P\left(a_1 Y_{1,j} + a_2 Y_{2,j} \leqslant \theta\right) && \text{(D.3)} \\
&= P\left(Y_{2,j} \leqslant -\frac{a_1}{a_2} Y_{1,j} + \frac{\theta}{a_2}\right) \\
&= \int_{y=-\infty}^{\infty} \int_{y_2=-\infty}^{-\frac{a_1}{a_2}y+\frac{\theta}{a_2}} dF_{Y_{2,j}}(y_2)\, dF_{Y_{1,j}}(y) \\
&= \int_{y=-\infty}^{\infty} F_{Y_{2,j}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)\, dF_{Y_{1,j}}(y) && \text{(D.4)} \\
&\geqslant \int_{y=-\infty}^{\infty} F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)\, dF_{Y_{1,j}}(y) && \text{(D.5)} \\
&= \left[F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right) F_{Y_{1,j}}(y)\right]_{y=-\infty}^{\infty} - \int_{y=-\infty}^{\infty} F_{Y_{1,j}}(y)\, dF_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right) \\
& && \text{(D.6)} \\
&= [(0 \times 1) - (1 \times 0)] + \int_{y=-\infty}^{\infty} F_{Y_{1,j}}(y)\, d\left(-F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)\right) && \text{(D.7)} \\
&= \int_{y=-\infty}^{\infty} F_{Y_{1,j}}(y)\, dH_{Y_{2,k}}(y), && \text{(D.8)}
\end{aligned}$$

where

$$H_{Y_{2,k}}(y) \quad = \quad -F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right) \qquad \text{(D.9)}$$

is a left-continuous, bounded, *monotonic increasing* function of $y$. It takes on values between $-1$ and 0 inclusive, and $\lim_{y\to-\infty} H_{Y_{2,k}}(y) = -1$ and $\lim_{y\to\infty} H_{Y_{2,k}}(y) = 0$. Continuing

from Equation D.8, then

$$
\begin{aligned}
F_{\Theta_j}(\theta) &\geqslant \int_{y=-\infty}^{\infty} F_{Y_{1,j}}(y)\, dH_{Y_{2,k}}(y) && \text{(D.10)}\\
&\geqslant \int_{y=-\infty}^{\infty} F_{Y_{1,k}}(y)\, dH_{Y_{2,k}}(y) && \text{(D.11)}\\
&= \Big[F_{Y_{1,k}}(y) H_{Y_{2,k}}(y)\Big]_{y=-\infty}^{\infty} - \int_{y=-\infty}^{\infty} H_{Y_{2,k}}(y)\, dF_{Y_{1,k}}(y) && \text{(D.12)}\\
&= [(1\times 0) - (0\times -1)] + \int_{y=-\infty}^{\infty} F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)\, dF_{Y_{1,k}}(y) && \text{(D.13)}\\
&= P\left(a_1 Y_{1,k} + a_2 Y_{2,k} \leqslant \theta\right) && \text{(D.14)}\\
&= F_{\Theta_k}(\theta).
\end{aligned}
$$

This shows that $F_{\Theta_j}(\theta) \geqslant F_{\Theta_k}(\theta)$ holds for all $\theta \in \mathbb{R}$, which implies that $\Theta_j \overset{st}{\leqslant} \Theta_k$.

Equation D.14 follows from Equation D.13 by noting that Equation D.13 is identical in form to Equation D.4, except that the subscript $k$ has replaced the subscript $j$ in both the integrand and the integrator.[1] Working forwards from Equation D.13 to Equation D.14 is essentially the same as working backwards from Equation D.4 to Equation D.3.

The function $H_{Y_{2,k}}(y)$ was introduced in Equation D.8, although it could also have been used in Equations D.6 and D.7. Using $H_{Y_{2,k}}(y)$ in place of $-F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)$ is not absolutely necessary, but it may help to show that Equations D.6 and D.7 are non-negative in value. Let $a = \frac{a_1}{a_2} > 0$ and $b = \frac{\theta}{a_2}$, for brevity of notation. Dropping the limits, the subtraction of $\int F_{Y_{1,j}}(y)\, dF_{Y_{2,k}}(-ay+b)$ in Equation D.6 is equivalent to the addition of $\int F_{Y_{1,j}}(y)\, d(-F_{Y_{2,k}}(-ay+b))$ in Equation D.7, because the outer negative sign may be carried through to the integrator. The sign of a Riemann-Stieltjes integral is determined by the sign of the integrand and the *directionality* of the integrator. The directionality refers to whether the integrator is monotonic increasing or monotonic decreasing. The integrand in Equations D.7 and D.8, $F_{Y_{1,j}}(y)$, is a non-negative (sometimes positive) function and the *directionality* of the integrator, $H_{Y_{2,k}}(y) = -F_{Y_{2,k}}(-ay+b)$, is also non-negative (sometimes positive). This is because, by the properties of cumulative distribution functions, $F_{Y_{2,k}}(-ay+b)$ is non-increasing (and sometimes decreasing), and so $H_{Y_{2,k}}(y)$ is non-decreasing (and sometimes increasing).

If the random variable associated with either the integrand or the integrator (or both) in Equations D.4 to D.13 is continuous, then the integrand or integrator is a continuous function, and since both the integrand and the integrator are monotonic and bounded,

[1] For a Riemann-Stieltjes integral of the form $\int F(y)\, dG(y)$, $F(y)$ is the integrand and $G(y)$ is the integrator.

then the Riemann-Stieltjes integrals involved are well-defined (Rudin, 1976, Thms. 6.8 and 6.9).

If all of the random variables involved in Theorem 4 are discrete, then both the integrand and integrator functions are piecewise continuous, meaning they have discontinuities in $\mathbb{R}$. Nevertheless the proof of Theorem 4 uses well-defined Riemann-Stieltjes integrals. If the integrand *and* the integrator functions were *both only* left-continuous or *both only* right-continuous at shared points of discontinuity, then the upper and lower sums that (in the limit) define the integral would have different values (which is to say, the Riemann-Stieltjes integral would not exist). However, in the integrals used in Equations D.4 through to D.13, one function is always left-continuous and the other is always right-continuous, and so the integrals are well-defined, precisely because they share no common left and no common right discontinuities (Burkill & Burkill, 1970, Thm. 6.24). This is also a necessary condition in order for integration by parts to hold in Equations D.6 and D.12 (Burkill & Burkill, 1970, Thm. 6.25),

Considering general continuous, discrete, or mixed random variables once again, the inequality between Equations D.4 and D.5 holds because $Y_{2,j} \overset{st}{<} Y_{2,k}$ by assumption, which is to say that $F_{Y_{2,j}}(y) \geqslant F_{Y_{2,k}}(y)$ holds for all values of the argument, and so $F_{Y_{2,j}}(-ay+b) \geqslant F_{Y_{2,k}}(-ay+b)$ as well. The fact that the direction of the arguments[2] of $F_{Y_{2,j}}$ and $F_{Y_{2,k}}$ in Equations D.4 and D.5 are opposite to the direction of the variable $y$ means only that the integrands are a non-increasing functions, not that they are negative-valued. By definition, $F_{Y_{2,j}}$ and $F_{Y_{2,k}}$ are both non-negative (sometimes positive) functions that have positive value over non-zero intervals in $\mathbb{R}$. If an inequality exists between Equations D.4 and D.5, then it must be in the direction indicated above. The same argument can also be put forward for the inequality between Equations D.10 and D.11, because the integrator, $H_{Y_{2,k}}(y)$, is a monotonic increasing function and because $F_{Y_{1,j}}(y) \geqslant F_{Y_{1,k}}(y)$.

**Summary.** $F_{\Theta_j}(\theta) \geqslant F_{\Theta_k}(\theta)$ for all $\theta \in \mathbb{R}$, so $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{\leqslant} (a_1 Y_{1,k} + a_2 Y_{2,k})$. What remains to be shown is that a strict inequality, $F_{\Theta_j}(\theta) > F_{\Theta_k}(\theta)$, holds for all values of $\theta$ over some non-zero interval, implying that the weighted sums of random variables follow a *strict* stochastic ordering. Conditions under which $F_{\Theta_j}(\theta) = F_{\Theta_k}(\theta)$ are given below, which lead to a description of conditions under which $F_{\Theta_j}(\theta) > F_{\Theta_k}(\theta)$ holds.

### Equality versus strict inequality between Equations D.4 and D.5

There may be values of $\theta$ for which $F_{\Theta_j}(\theta) = F_{\Theta_k}(\theta)$. In order for that to happen, then the non-strict inequalities in Equations D.5 and D.11 must both be equalities. Equality would occur in Equation D.5 if $Y_{2,j} = Y_{2,k}$, and would occur in Equation D.11 if $Y_{1,j} = Y_{1,k}$, but

[2] If arguments of c.d.f.'s have been dropped, then this implictly means that relationships are considered for all possible values of the argument over a given domain, including a linear rescaling of the argument (i.e. $-ay+b$). If no domain is specified, then the domain is implictly the entire real number line.

these possibilities are excluded by assumption since $Y_{1,j} \overset{st}{<} Y_{1,k}$ and $Y_{2,j} \overset{st}{<} Y_{2,k}$. However, it is also possible that Equations D.4 and D.5 are equal, and that Equations D.10 and D.11 are equal, even when the *integrands* from one line to the next follow a *strict* inequality over a non-zero interval. This can occur because the result of the integrals depend on the integrator as well as the integrand.

Consider just the relationship between Equations D.4 and D.5 for any given value of $\theta$. $F_{Y_{2,j}}(y) > F_{Y_{2,k}}(y)$ holds, by assumption, for $y$ values over some non-zero interval, regardless of the form of $Y_{2,j}$ and $Y_{2,k}$. Consider a simplified version of Equations D.4 and D.5, in which the strict inequality, $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$, holds only over a single interval of $y$ values, denoted $C^+$ (where $a = \frac{a_1}{a_2} > 0$ and $b = \frac{\theta}{a_2}$). Let the bounds of the interval be $c_1$ and $c_2$, where $c_1 < c_2$ and where $c_1$ and $c_2$ could be finite or infinite. In that case, $C^+ = (c_1, c_2]$ if at least one of $Y_{2,j}$ and $Y_{2,k}$ is discrete, and $C^+ = (c_1, c_2)$ if both $Y_{2,j}$ and $Y_{2,k}$ are continuous. There is inclusion at the *end* of the interval in the discrete case because, although $F_{Y_{2,j}}(y)$ and $F_{Y_{2,k}}(y)$ are right-continuous functions, $F_{Y_{2,j}}(-ay+b)$ and $F_{Y_{2,k}}(-ay+b)$ in the integrals are left-continuous functions, meaning the start of the interval $C^+$ is excluded while the end is included. If either of $Y_{2,j}$ or $Y_{2,k}$ is a mixed random variable, then if the interval of $y$ values over which $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$ ends with a discontinuity, then $c_2$ is included, otherwise it is excluded. The complement to $C^+$ is denoted as $C^0$, which is the union of intervals over which $y$ is such that $F_{Y_{2,j}}(-ay+b) = F_{Y_{2,k}}(-ay+b)$. Since $C^+ \cup C^0 = \mathbb{R}$ and $C^+ \cap C^0 = \emptyset$, then Equations D.4 and D.5 may be reformulated as

$$\int_{y \in C^+ \cup C^0} F_{Y_{2,j}}(-ay+b)\, dF_{Y_{1,j}}(y) \quad \geqslant \quad \int_{y \in C^+ \cup C^0} F_{Y_{2,k}}(-ay+b)\, dF_{Y_{1,j}}(y). \quad \text{(D.15)}$$

By definition, $F_{Y_{2,j}}(-ay+b) = F_{Y_{2,k}}(-ay+b)$ for all $y \in C^0$, which implies that

$$\int_{y \in C^0} F_{Y_{2,j}}(-ay+b)\, dF_{Y_{1,j}}(y) \quad = \quad \int_{y \in C^0} F_{Y_{2,k}}(-ay+b)\, dF_{Y_{1,j}}(y). \quad \text{(D.16)}$$

Subtracting the integrals in Equation D.16 from those in Equation D.15 leaves

$$\int_{y \in C^+} F_{Y_{2,j}}(-ay+b)\, dF_{Y_{1,j}}(y) \quad \geqslant \quad \int_{y \in C^+} F_{Y_{2,k}}(-ay+b)\, dF_{Y_{1,j}}(y). \quad \text{(D.17)}$$

The non-strict inequalities in Equations D.15 and D.17 hold in general. For any specific case, however, the integrals in Equation D.15 are either equal, or there is a strict-inequality between them. Which of these two possibilities holds depends only on the integrals taken over the interval $C^+$, and not those taken over $C^0$. This is because the integrals in Equation D.16 are always equal (by the definition of $C^0$).

Since $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$ for all $y \in C^+$, then whether or not there is an equality in Equation D.17 *does not depend on the integrands*, $F_{Y_{2,j}}$ and $F_{Y_{2,k}}$. Rather, *it*

*depends on the form of the integrator*, $F_{Y_{1,j}}(y)$, in the interval $C^+$. (Note that the integrator function remains the same throughout this development.) The integrator, $F_{Y_{1,j}}(y)$, is a monotonic increasing function, and it either increases over the interval $C^+$ or it remains at a constant value over all of $C^+$. Each case is considered in turn.

If $F_{Y_{1,j}}(y)$ is a constant for all $y \in C^+$, then

$$\begin{aligned} \int_{y \in C^+} F_{Y_{2,j}}(-ay+b)\, dF_{Y_{1,j}}(y) &= \int_{y \in C^+} F_{Y_{2,k}}(-ay+b)\, dF_{Y_{1,j}}(y) \\ &= 0 \end{aligned} \tag{D.18}$$

(Clarke, 1975, Thm. 1(ii), p. 166). This implies that if the interval $C^+$ falls entirely within an interval over which the integrator, $F_{Y_{1,j}}(y)$, is constant, then there is *equality* in Equation D.17, which in turn implies equality in Equation D.15, and hence equality between Equations D.4 and D.5 for any given value of $\theta$. (In such a case, the value that Equations D.4 and D.5 take on depends on the value of Equation D.16, which may be zero or positive.)

On the other hand, if $F_{Y_{1,j}}(y)$ is not a constant over the entire interval $C^+$, then $F_{Y_{1,j}}(y)$ *increases* within $C^+$. This, and the fact that $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$ for all $y \in C^+$, is sufficient for

$$\int_{y \in C^+} F_{Y_{2,j}}(-ay+b)\, dF_{Y_{1,j}}(y) > \int_{y \in C^+} F_{Y_{2,k}}(-ay+b)\, dF_{Y_{1,j}}(y) \tag{D.19}$$

to hold, as is shown below. Whenever Equation D.19 holds, then there is a strict inequality in Equation D.15, and hence between Equations D.4 and D.5.

If $F_{Y_{1,j}}(y)$ increases in $C^+$, then either it has at least one discontinuity with a positive jump (change in function value), or it has an increasing continuous section, or both. If $F_{Y_{1,j}}(y)$ increases continuously[3] in $C^+$, then Equation D.19 becomes

$$\int_{y \in C^+} F_{Y_{2,j}}(-ay+b) f_{Y_{1,j}}(y)\, dy > \int_{y \in C^+} F_{Y_{2,k}}(-ay+b) f_{Y_{1,j}}(y)\, dy, \tag{D.20}$$

where $f$ denotes a probability density function, which is non-negative, and must be positive over the portion or portions of $C^+$ over which $F_{Y_{1,j}}(y)$ is increasing. Since $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$ for all $y \in C^+$, then there is a strict inequality in Equation D.20, and so Equation D.19 holds for increasing $F_{Y_{1,j}}$ when $Y_{1,j}$ is *continuous*.

If $F_{Y_{1,j}}(y)$ has a discontinuity at $y = c$ within $C^+$, then the jump value is positive and equals $P(Y_{1,j} = c) > 0$. The contribution from this jump to the *left-hand side* of

[3] $F_{Y_{1,j}}(y)$ may be increasing over all of $C^+ = (c_1, c_2)$, or only over parts of $C^+$. The result is the same in either case.

Equation D.19 is

$$\lim_{\delta \to 0} F_{Y_{2,j}}(-ac+b)\left(F_{Y_{1,j}}(c) - F_{Y_{1,j}}(c-\delta)\right) = F_{Y_{2,j}}(-ac+b)P(Y_{1,j}=c)$$
$$> F_{Y_{2,k}}(-ac+b)P(Y_{1,j}=c), \quad \text{(D.21)}$$

since $F_{Y_{2,j}}(-ac+b) > F_{Y_{2,k}}(-ac+b)$ for any $c \in C^+$. Note that Equation D.21 is the contribution from the jump at $c$ to the *right-hand side* of Equation D.19. The same *strict inequality* holds for any and all jumps in $C^+$, and so Equation D.19 holds for increasing $F_{Y_{1,j}}$ when $Y_{1,j}$ is *discrete*.

Equation D.19 also holds for increasing $F_{Y_{1,j}}$ when $Y_{1,j}$ is a *mixed* random variable, in which case, the Riemann-Stieltjes integral in Equation D.19 may be evaluated separately for continuous sections and discontinuous jumps within $C^+$.

**Summary.** A derivation showing that $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{\leqslant} (a_1 Y_{1,k} + a_2 Y_{2,k})$ was given, and the existence of the Riemann-Stieltjes integrals in the derivation was discussed, particularly for the discrete case. A non-strict inequality between Equations D.4 and D.5 holds in general, but for any particular case, it is either a strict inequality, or just an equality. A simplified situation was introduced in which the *integrands*, $F_{Y_{2,j}}$ and $F_{Y_{2,k}}$, are such that $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$ holds true over a *single* interval of $y$ values, $C^+$. In that case, whether Equations D.4 and D.5 are equal in value or whether they follow a strict inequality depends on the nature of the *integrator*, $F_{Y_{1,j}}(y)$, over the interval $C^+$. If the integrator is constant over the entire interval $C^+$, then Equations D.4 and D.5 are equal. If the integrator increases somewhere in the interval $C^+$, then there is a strict inequality between Equations D.4 and D.5 for any given value of $\theta$. This result holds regardless of the form of the random variables involved.

Continuing with the proof, it remains to extend the results beyond the simplified situation, then to extend the results to also cover Equations D.10 and D.11, and finally to show why $\Theta_j \overset{st}{<} \Theta_k$ rather than $\Theta_j \overset{st}{\leqslant} \Theta_k$.

### Generalising from the simplified situation

The interval $C^+$ was originally defined (p. 291) based on a simplified situation where $F_{Y_{2,j}}(-ay+b) > F_{Y_{2,k}}(-ay+b)$ held only over single interval of $y$ values, $C^+$. In general, there may be any number of such intervals, say $C_1^+, C_2^+$, etc. Like before, define $C^0$ to be the complement in $\mathbb{R}$ of the union $C_1^+ \cup C_2^+ \cup \ldots$ Note that the integrals in Equations D.4 to D.13, taken from $y = -\infty$ to $\infty$, may be restated as integrals over $y \in C^0 \cup C_1^+ \cup C_2^+ \cup \ldots$ It can be seen that Equation D.16 still holds, by the definition of $C^0$. This implies that whether Equations D.4 and D.5 are equal or whether they follow a strict inequality is determined by integrals like those in Equation D.17, which are calculated separately over

$C_1^+, C_2^+$, etc. For the $\ell^{th}$ such interval, the relationship to consider is

$$\int_{y \in C_\ell^+} F_{Y_{2,j}}(-ay+b)\, dF_{Y_{1,j}}(y) \quad \geqslant \quad \int_{y \in C_\ell^+} F_{Y_{2,k}}(-ay+b)\, dF_{Y_{1,j}}(y). \qquad \text{(D.22)}$$

Like before, it is the form of the integrator, $F_{Y_{1,j}}(y)$, over $C_\ell^+$ that determines the contribution of Equation D.22 to Equations D.4 and D.5. Across $C^+$ intervals, if $F_{Y_{1,j}}(y)$ remains constant over $C_\ell^+$ for *every* $\ell$,[4] then there is an equality in Equation D.22 for every $\ell$ (where the integrals are all equal to 0, like in Equation D.18). In that case, there is an equality between Equations D.4 and D.5. On the other hand, if there is *even a single interval*, $C_\ell^+$, for which $F_{Y_{1,j}}(y)$ increases in $C_\ell^+$, then there is a strict inequality in Equation D.22, which implies a strict inequality between Equations D.4 and D.5, and hence $F_{\Theta_j}(\theta) > F_{\Theta_k}(\theta)$ holds for the given value of $\theta$.

### The inequality between Equations D.10 and D.11

The emphasis so far has been on the relationship between Equations D.4 and D.5. Identical arguments also hold for the relationship between Equations D.10 and D.11. The main change in the proof for those equations has to do with the definition of $C^+$. The integrands in Equations D.4 and D.5 are left-continuous functions, so in the discrete case, $C^+ = (c_1, c_2]$ included the upper cutoff in the interval $C^+$. The integrands in Equations D.10 and D.11 are right-continuous functions, and hence, in the discrete case, $C^+ = [c_1, c_2)$ should include the *lower* cutoff in the interval $C^+$. (The definition of $C^+$ should also be changed accordingly if the integrands are mixed random variables.) Note that $C^+ = (c_1, c_2)$ is still used in the continuous case and that the continuity or otherwise of the integrator in Equations D.10 and D.11 does not affect the definition of $C^+$. Equations analogous to Equations D.15 to D.21 can be constructed based on Equations D.10 and D.11 rather than on Equations D.4 and D.5. As before, any number of intervals $C_1^+, C_2^+ \ldots$ can be defined and if the integrator, $H_{Y_{2,k}}(y)$, is constant over *all* such intervals, then there is an equality between Equations D.10 and D.11, otherwise there is a strict inequality.

In order for $F_{\Theta_j}(\theta) > F_{\Theta_k}(\theta)$ to hold for a given value of $\theta$, only one of Equations D.5 and D.11 need have a strict inequality, while the other equation may have either an equality or a strict inequality.[5] For this reason, Equations D.10 and D.11 are left for now, although they are important later in Corollary 14.

### Strict inequalities over non-zero intervals

The effect of $\theta$ in Equations D.4 to D.13 is to change the relative positions of the integrand and the integrator functions, for example $F_{Y_{2,k}}$ and $F_{Y_{1,j}}$ respectively in Equation D.5. The

[4] Although $F_{Y_{1,j}}(y)$ may take on a different value for different $C_\ell^+$.

[5] The intervals over which there is a strict inequality between Equations D.4 and D.5 are not necessarily the same as the intervals over which there is a strict inequality between Equations D.10 and D.11.

scaling of the argument of $F_{Y_{2,k}}$ by the constants $a_1$ and $a_2$ does not matter here, only that $\theta$ shifts the relative locations of the functions.[6]

Although $F_{\Theta_j}(\theta) = F_{\Theta_k}(\theta)$ may hold for some $\theta$ (as shown earlier), it cannot hold for all $\theta$. Consider Equations D.4 and D.5. As $\theta$ systematically increases from $-\infty$, then the monotonic decreasing integrands, $F_{Y_{2,j}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)$ and $F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right)$, systematically shift together from left to right while the monotonic increasing integrator, $F_{Y_{1,j}}(y)$, remains fixed in place (because $\theta$ is not a parameter of the integrator). As the integrands shift to the right, so too does any non-zero interval of $y$ values over which

$$F_{Y_{2,j}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right) \quad > \quad F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right), \tag{D.23}$$

and there must be at least one such interval, because $Y_{2,j} \overset{st}{<} Y_{2,k}$ by assumption.

Equation D.23 may be rearranged as

$$F_{Y_{2,j}}\left(-\frac{a_1}{a_2}(y-\delta)\right) \quad > \quad F_{Y_{2,k}}\left(-\frac{a_1}{a_2}(y-\delta)\right),$$

where $\delta = \frac{\theta}{a_1}$ is the horizontal shift of the integrand functions for a given parameter value, $\theta$. Choose any non-zero interval, $D_0^+ = (c_1, c_2)$, over which

$$F_{Y_{2,j}}\left(-\frac{a_1}{a_2}y\right) \quad > \quad F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y\right) \tag{D.24}$$

holds for all $y \in (c_1, c_2)$ (where Equation D.24 is Equation D.23 when $\theta = 0$). Let $D_\theta^+ = (c_1 - \delta, c_2 - \delta) = (c_1 - \frac{\theta}{a_1}, c_2 - \frac{\theta}{a_1})$ be the horizontal translation of $D_0^+$. Although the location of $D_\theta^+$ depends on $\theta$, its length is always $|D_\theta^+| = c_2 - c_1$, which does not depend on $\theta$.

The *integrator* in Equations D.4 and D.5, $F_{Y_{1,j}}$, must increase somewhere within $\mathbb{R}$. Choose any interval of increase, say $I_\beta = [\beta_1, \beta_2]$, where $-\infty < \beta_1 \leqslant \beta_2 < \infty$, and possibly $\beta_1 = \beta_2$ if $F_{Y_{1,j}}(y)$ is discontinuous at $\beta_1$. For some small enough value, $\theta_1$, $D_{\theta_1}^+$ is all *just to the left* of $I_\beta$. The right-hand side of $D_0^+$ is at $y = c_2$ and the left-hand side of $I_\beta$ is at $y = \beta_1$. To shift $D_0^+$ so that its right-hand side matches the left-hand side of $I_\beta$, then $\delta$ must be such that $c_2 + \delta = \beta_1$ (i.e. $\delta = \beta_1 - c_2$), and so $\theta_1 = a_1\delta = a_1(\beta_1 - c_2)$. Similarly, for some large enough value, $\theta_2$, $D_{\theta_2}^+$ is all *just to the right* of $I_\beta$. The left-hand side of $D_0^+$ is at $y = c_1$ and the right-hand side of $I_\beta$ is at $y = \beta_2$. To shift $D_0^+$ so that its left-hand side matches the right-hand side of $I_\beta$, then $\delta$ must be such that $c_1 + \delta = \beta_2$ (i.e. $\delta = \beta_2 - c_1$), and so $\theta_2 = a_1\delta = a_1(\beta_2 - c_1)$.

As $\theta$ systematically increases from $\theta_1$ to $\theta_2$, the general interval, $D_\theta^+$ systematically shifts right from $D_{\theta_1}^+$ to $D_{\theta_2}^+$, and all the while, $D_\theta^+$ overlaps with $I_\beta$, at least partly if

[6] The parameter $\theta$ also does this implicitly in Equations D.10 and D.11 (which is made explicit by the definition of $H_{Y_{2,k}}$, given in Equation D.9).

not in full. From the definitions of $D_\theta^+$ and of $I_\beta$, and from the argument given for Equation D.19, then

$$\int_{y \in D_\theta^+ \cap I_\beta} F_{Y_{2,j}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right) dF_{Y_{1,j}}(y) \;>\; \int_{y \in D_\theta^+ \cap I_\beta} F_{Y_{2,k}}\left(-\frac{a_1}{a_2}y + \frac{\theta}{a_2}\right) dF_{Y_{1,j}}(y) \quad \text{(D.25)}$$

holds for all $\theta$ such that $\theta_1 < \theta < \theta_2$. The length of this continuous interval of $\theta$ values is

$$\begin{aligned} \theta_2 - \theta_1 &= a_1(\beta_2 - c_1) - a_1(\beta_1 - c_2) \\ &= a_1\left((\beta_2 - \beta_1) + (c_2 - c_1)\right) \\ &= a_1\left(|I_\beta| + \left|D_{\theta_2}^+\right|\right) \\ &> 0 \end{aligned} \quad \text{(D.26)}$$

since $\left|D_{\theta_2}^+\right| = c_2 - c_1 > 0$, $a_1 > 0$ and $|I_\beta| = \beta_2 - \beta_1 \geqslant 0$. In order to shift the general interval $D_\theta^+$ from $D_{\theta_1}^+$ to $D_{\theta_2}^+$, then $D_\theta^+$ must move across the width of $I_\beta$ plus the width of $D_{\theta_2}^+$, which is what shown in Equation D.26.

The left and right-hand sides of Equation D.25 contribute to Equations D.4 and D.5 respectively. Since the interval of integration in Equation D.25, $y \in D_\theta^+ \cap I_\beta$, is a subset of the interval of integration in Equations D.4 and D.5, then there is a *strict inequality* between Equations D.4 and D.5 for any given $\theta \in (\theta_1, \theta_2)$. This length of this interval is $\theta_2 - \theta_1 > 0$, and therefore $F_{\Theta_j}(\theta) > F_{\Theta_k}(\theta)$ holds over a non-zero interval of $\theta$ values. Since $F_{\Theta_j}(\theta) \geqslant F_{\Theta_k}(\theta)$ holds for all $\theta \in \mathbb{R}$ (Equations D.3 through D.14), then $\Theta_j \overset{st}{<} \Theta_k$, which is to say that $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{<} (a_1 Y_{1,k} + a_2 Y_{2,k})$. This completes the proof of Theorem 4. Q.E.D.

## D.2 Corollaries to Theorem 4

Theorem 4 applies to when there are only two divisions. Hence the index $\eta$ in Equation D.1 was equal to either 1 or 2. Theorem 4 can be easily extended to more than two divisions, and to cover conditions involving non-strict stochastic ordering. Seven corollaries are given. Corollaries 10, and 11 extend the results of Theorem 4 to cover more general situations where there are than two divisions and more than two $Y$-variables per division. Corollary 12 deals with the expected values of the weighted sums that result from this extension of the theorem. Corollaries 13 to 16 deal with the effect that non-strict stochastic ordering (within each division) has on the strict stochastic ordering of the weighted sums of $Y$-variables. In the corollaries, let $m$ be the total number of divisions, where $m \geqslant 2$ could be finite or infinite (trivially, $m = 1$ is also possible), and let $a_1$, $a_2$ ... $a_m$ be any positive constants. Under this extension, say that the definition in Equation D.1 still applies for $\eta = 1, 2 \ldots m$. Of course, if $a_\eta = 1$ for all $\eta$, then any weighted sum of random variables

becomes a plain sum, and if $a_\eta = \frac{1}{m}$ for all $\eta$, then the weighted sum is an arithmetic mean.

**Corollary 10** *If* $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ *for all* $\eta = 1, 2 \ldots m$, *then*

$$\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right) \quad \overset{st}{<} \quad \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right) \tag{D.27}$$

*holds for any* $m \geqslant 2$.

**Proof by induction.** Suppose that $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ for all $\eta = 1, 2 \ldots m$. For $m = 2$, Equation D.27 holds because $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{<} (a_1 Y_{1,k} + a_2 Y_{2,k})$, by Theorem 4. By assumption, $Y_{\gamma+1,j} \overset{st}{<} Y_{\gamma+1,k}$ holds for all $\gamma = 1, 2 \ldots (m-1)$, so if $\sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,j} \overset{st}{<} \sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,k}$, is true, then by Theorem 4, $\left(a_{\gamma+1} Y_{\gamma+1,j} + \sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(a_{\gamma+1} Y_{\gamma+1,k} + \sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,k}\right)$ is also true, which is to say that $\left(\sum_{\eta=1}^{\gamma+1} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(\sum_{\eta=1}^{\gamma+1} a_\eta Y_{\eta,k}\right)$ is true. If Equation D.27 holds for $m = \gamma$, then it also holds for $m = \gamma + 1$, and since Equation D.27 holds for $m = 2$, it therefore holds for all $m = 2, 3 \ldots$, which completes the proof.

**Corollary 11** *If the Y-variables in each of m divisions form a strictly stochastically ordered set, and the same ordering holds within each set, then the weighted sums across divisions follow the same ordering. Without loss of generality, if* $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \ldots$ *holds for all divisions,* $\eta = 1, 2 \ldots m$, *then*

$$\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,1}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,2}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,3}\right) \overset{st}{<} \ldots \tag{D.28}$$

**Proof.** In Equation D.1 (p. 287), the $\varepsilon^{th}$ $Y$-variable in the $\eta^{th}$ division is $Y_{\eta,\varepsilon}$, where $\eta = 1, 2 \ldots m$ denotes the division, and $\varepsilon = 1, 2, 3 \ldots$ denotes the index-value within each division. By assumption, the stochastic ordering within each division (ordered according to the same second index values, $\varepsilon$) is the same for all divisions. Corollary 11 follows then from the application of Corollary 10 to sets of $Y$-variables taken across divisions and defined by successive overlapping pairs of $\varepsilon$ values (1 and 2, then 2 and 3, ... ), where each set of $Y$-variables consists of all $m$ $Y$-variables, one per division, that share the same second index value, $\varepsilon$.

Note that Equations D.27 and D.28 are only guaranteed to hold *if* the stochastic ordering is the same across divisions. If the stochastic ordering is different across divisions, or if there is no stochastic ordering within some divisions or within parts of some divisions, then the equations may hold, but they do not necessarily hold.

**Corollary 12** *Without loss of generality, if* $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \ldots$ *holds for all divisions,* $\eta = 1, 2 \ldots m$*, then*

$$E\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,1}\right) < E\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,2}\right) < E\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,3}\right) < \ldots \tag{D.29}$$

Corollary 4 in Appendix C showed that the stochastic ordering of a set of random variables results in the numerical ordering of their expected values. Corollary 12 follows from Corollary 4 applied to the weighted sums that result from Corollary 11 (particularly Equation D.28).

The next two corollaries are concerned with *non-strict* stochastic ordering between the $j^{th}$ and the $k^{th}$ $Y$-variables in each of two divisions.

**Corollary 13** *If* $Y_{1,j} \overset{st}{\leqslant} Y_{1,k}$ *and* $Y_{2,j} \overset{st}{\leqslant} Y_{2,k}$*, then* $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{\leqslant} (a_1 Y_{1,k} + a_2 Y_{2,k})$.

**Proof.** This has already essentially be shown in the proof of Theorem 4. In particular, Equations D.3 through to D.14 showed that $P(a_1 Y_{1,j} + a_2 Y_{2,j} \leqslant \theta) \geqslant P(a_1 Y_{1,k} + a_2 Y_{2,k} \leqslant \theta)$ for all values of $\theta$, which is equivalent to showing that $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{\leqslant} (a_1 Y_{1,k} + a_2 Y_{2,k})$.

**Corollary 14** *If either* **(a)** $Y_{1,j} \overset{st}{\leqslant} Y_{1,k}$ *and* $Y_{2,j} \overset{st}{<} Y_{2,k}$*, or* **(b)** $Y_{1,j} \overset{st}{<} Y_{1,k}$ *and* $Y_{2,j} \overset{st}{\leqslant} Y_{2,k}$*, then* $(a_1 Y_{1,j} + a_2 Y_{2,j}) \overset{st}{<} (a_1 Y_{1,k} + a_2 Y_{2,k})$.

**Proof.** The difference between Corollaries 13 and 14 is whether the resulting weighted sums are strictly or non-strictly ordered. Since strict ordering implies non-strict ordering, but not vice versa (Appendix C, p. 260), then Equations D.3 through to D.14 (in the proof of Theorem 4) still hold under either of conditions (a) or (b).

Under condition (a), if $Y_{1,j} \overset{st}{\leqslant} Y_{1,k}$, then either $Y_{1,j} \overset{st}{<} Y_{1,k}$, or $Y_{1,j} = Y_{1,k}$. If $Y_{1,j} \overset{st}{<} Y_{1,k}$ holds in condition (a), then Corollary 14 reduces to Theorem 4, which proves the conclusion in Corollary 14. If $Y_{1,j} = Y_{1,k}$ holds in condition (a), then $F_{Y_{1,k}}(y) = F_{Y_{1,j}}(y)$ for all $y$, which implies that the non-strict inequality in Equation D.11 can be replaced by an equality without affecting the rest of the proof of Theorem 4. Since the rest of the proof is unaffected, then the conclusion of Theorem 4 (with a *strict* ordering of weighted sums) remains unchanged Hence, strict stochastic ordering of weighted sums occurs when $Y_{1,j} = Y_{1,k}$ as well as when $Y_{1,j} \overset{st}{<} Y_{1,k}$ (i.e. under condition (a) in Corollary 14. Furthermore, condition (b) is merely a restatement of condition (a), because the first indices $\eta = 1$ and $\eta = 2$ are arbitrary (i.e. it does not matter which division is labelled first and which is labelled second). Hence, the same result applies as in condition (a).

The results in Corollaries 13 and 14 can be extended to when there are more than two divisions and, as before, it is a question of strict versus non-strict order in the result. An

extension of Corollary 14 is presented first, so it can be used to demonstrate the extension to Corollary 13.

**Corollary 15** *Assume there are $m \geqslant 2$ divisions, and that non-strict ordering $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ holds for all $\eta = 1, 2 \ldots m$. If $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ also holds for at least one value $\eta \in \{1, 2, \ldots, m\}$ (i.e. there is strict ordering in at least one division), then $\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right)$ (Equation D.27) holds.*

**Proof by induction.** Without loss of generality, assume that *strict* stochastic ordering occurs between the $j^{th}$ and $k^{th}$ $Y$-variables in the first division ($\eta = 1$), that is $Y_{1,j} \overset{st}{<} Y_{1,k}$. For all of the other divisions ($\eta = 2 \ldots m$) $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ holds, and possibly $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$. For the second division, $\left(\sum_{\eta=1}^{2} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(\sum_{\eta=1}^{2} a_\eta Y_{\eta,k}\right)$ holds, either by Theorem 4 (if $Y_{2,j} \overset{st}{<} Y_{2,k}$) or by Corollary 14 (if $Y_{2,j} \overset{st}{\leqslant} Y_{2,k}$). This shows that Equation D.27 holds for $m = 2$ divisions. By assumption, $Y_{\gamma+1,j} \overset{st}{\leqslant} Y_{\gamma+1,k}$ holds for all $\gamma = 1, 2 \ldots (m-1)$, so if $\left(\sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(\sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,k}\right)$ is true, then by Corollary 14, $\left(a_{\gamma+1} Y_{\gamma+1,j} + \sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(a_{\gamma+1} Y_{\gamma+1,k} + \sum_{\eta=1}^{\gamma} a_\eta Y_{\eta,k}\right)$ is also true, which is to say that $\left(\sum_{\eta=1}^{\gamma+1} a_\eta Y_{\eta,j}\right) \overset{st}{<} \left(\sum_{\eta=1}^{\gamma+1} a_\eta Y_{\eta,k}\right)$ is true. This implies that if Equation D.27 holds for $m = \gamma$, then it also holds for $m = \gamma + 1$, and since Equation D.27 holds for $m = 2$, it therefore holds for all $m = 2, 3 \ldots$. This completes the proof.

Assuming non-strict stochastic ordering between the $j^{th}$ and $k^{th}$ $Y$-variables for each division, Corollary 15 shows that if there is even *one* division which also has *strict* stochastic ordering, then the stochastic ordering of the weighted sum across all divisions is also strict. This holds true regardless of whether the ordering is strict or non-strict in the rest of the divisions. It also holds if $Y_{\eta,j} = Y_{\eta,k}$ holds for some divisions since $Y_{\eta,j} = Y_{\eta,k}$ implies $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$.

**Corollary 16** *Assume there are $m \geqslant 2$ divisions, and $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ holds for all $\eta = 1, 2 \ldots m$. The weighted sums, $\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right)$ and $\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right)$, in Equation D.27 only follow a non-strict stochastic ordering if the j-versus-k pairwise stochastic ordering is non-strict in all of the contributing divisions. Since $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ holds for all $\eta$, then the weighted sums follow the non-strict ordering*

$$\left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,j}\right) \quad \overset{st}{\leqslant} \quad \left(\sum_{\eta=1}^{m} a_\eta Y_{\eta,k}\right) \tag{D.30}$$

*only if $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ <u>does not hold</u> for any of the divisions, $\eta = 1, 2 \ldots m$.*

Corollary 16 is the complement of Corollary 15. Corollary 16 follows from repeated

application of Corollary 13 to an increasing sum for each of the $j^{th}$ and $k^{th}$ $Y$-variables, along the same lines as in the proof of Corollary 15. $Y_{\eta,j} \overset{st}{\leqslant} Y_{\eta,k}$ holds for all $\eta$ and if at any point, as $Y$-variables from more divisions are combined, there is *even one division* for which $Y_{\eta,j} \overset{st}{<} Y_{\eta,k}$ holds, then Equation D.27 holds, otherwise Equation D.30 holds.

Corollaries 15 and 16 imply that the stochastic ordering of a sequence of weighted sums of random variables (Equation D.28) and the numeric ordering of their associated expected values (Equation D.29) hold for any arbitrary combination of strict and non-strict inequalities in the ordering sequence. This result holds, subject to the conditions given in Corollaries 15 and 16, *if* the ordering sequence is the same for all of the divisions involved. For any given successive pair of second index values, $\varepsilon$ and $\varepsilon + 1$ (1 and 2, 2 and 3, ... ), the division with strict ordering does not have to be the same for different $\varepsilon$ values. This implies that Equation D.28 may hold even when there is no *single* division that has strict stochastic ordering among *all* of its $Y$-variables.

## D.3 Effect of transforms on weighted sums of stochastically ordered random variables

The results presented so far in this appendix deal with properties of weighted sums of $Y$-variables, as defined for each division on p. 287. This section deals with the weighted sums of s.m.i. transforms, and monotonic increasing step function transforms, of stochastically ordered random variables. In keeping with the notation in Chapter 5 and Appendix C, $Y$-variables are transformed into $R$-variables by using s.m.i. transforms, and $R$-variables are transformed into $Q$-variables by using step function transforms.

### Strictly monotonic increasing transforms

Suppose that there are $m \geqslant 2$ divisions and that there is an s.m.i. transform for each division. Specifically, let $h_\eta$ be the s.m.i. transform for the $\eta^{th}$ division. The $h$ transforms may be the same or may be different across divisions. Let

$$\begin{aligned} R_{\eta,\varepsilon} &= h_\eta(Y_{\eta,\varepsilon}) \\ &= h_\eta(x_{\eta,\varepsilon} \oplus_\eta U_{\eta,\varepsilon}), \end{aligned} \tag{D.31}$$

which defines a set of $R$-variables for each division. For the $\eta^{th}$ division, the set is $\{R_{\eta,1}, R_{\eta,2}, R_{\eta,3} \dots\}$. By Corollary 3 in Appendix C, if the underlying set of $Y$-variables is a stochastically ordered set, then so is the set of $R$-variables, and the stochastic ordering is the same for both sets. By Corollary 4, the ordering of the expected values of the $R$-variables *within each division* follows that of the expected values of the $Y$-variables.

Stochastic ordering of $Y$-variables may or may not exist within each division and even

if it does, it may or may not be the same across divisions. If such stochastic ordering does exist within each division *and it is the same across all divisions*, then the stochastic ordering of the $R$-variables within each division is also the same across all divisions and follows that of the $Y$-variables. Furthermore, Theorem 4 and Corollaries 10 to 16 can also be applied to the $R$-variables (as well as to the $Y$-variables), which implies that any weighted sums of $R$-variables will also follow the same stochastic ordering as that of the underlying $Y$-variables. If there are $m$ divisions, and if $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \dots$ holds for all divisions, $\eta = 1, 2 \dots m$, then $R_{\eta,1} \overset{st}{<} R_{\eta,2} \overset{st}{<} \dots$ also holds for all divisions, and

$$\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,1}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta R_{\eta,2}\right) \overset{st}{<} \left(\sum_{\eta=1}^{m} a_\eta R_{\eta,3}\right) \overset{st}{<} \dots \tag{D.32}$$

and

$$E\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,1}\right) < E\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,2}\right) < E\left(\sum_{\eta=1}^{m} a_\eta R_{\eta,3}\right) < \dots \tag{D.33}$$

hold, by Corollaries 11 and 12, respectively.

The weighting constants, $a_1$, $a_2$, $a_3 \dots$, for the $R$-variables need not be the same as the weighting constants, $a_1$, $a_2$, $a_3 \dots$, for the $Y$-variables in order for this to hold, although they could be.

**Monotonic increasing step function transforms**

Assume there are $m \geqslant 2$ divisions and that there is a monotonic increasing step function transform for each division that may be applied to the $R$-variables, as described for a single division in Section C.3 in Appendix C. Specifically, let $\Lambda_\eta$ be the transform for the $\eta^{th}$ division. The $\Lambda$ transforms may be the same or may be different across divisions. Let

$$Q_{\eta,\varepsilon} \quad = \quad \Lambda_\eta(R_{\eta,\varepsilon})\,, \tag{D.34}$$

where $R_{\eta,\varepsilon}$ is as defined in Equation D.31. Equation D.34 defines a set of $Q$-variables for each division. For the $\eta^{th}$ division, the set is $\{Q_{\eta,1}, Q_{\eta,2}, Q_{\eta,3} \dots\}$, and the results presented in Section C.3 apply here to each division. If the $\eta^{th}$ division has a strictly stochastically ordered set of $Y$-variables, then the related $Q$-variables also form a stochastically ordered set, although the ordering is not necessarily always strict (Corollary 8 in Appendix C has further details). In other words, if $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \dots$ holds then $Q_{\eta,1} \overset{st}{\leqslant} Q_{\eta,2} \overset{st}{\leqslant} \dots$ holds, by Corollary 8 (via intermediate $R$-variables). Furthermore, if $Y_{\eta,1} \overset{st}{<} Y_{\eta,2} \overset{st}{<} \dots$ holds for

all divisions, $\eta = 1, 2 \ldots m$, then the $Q$-variables are such that

$$\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,1}\right) \overset{st}{\leqslant} \left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,2}\right) \overset{st}{\leqslant} \left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,3}\right) \overset{st}{\leqslant} \ldots \tag{D.35}$$

is guaranteed to hold. Equation D.35 follows from Corollary 16 being applied to the $Q$-variables based on successive overlapping pairs of the second index value, $\varepsilon$ (i.e. 1 and 2, then 2 and 3, ... ). Consequently,

$$E\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,1}\right) \leqslant E\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,2}\right) \leqslant E\left(\sum_{\eta=1}^{m} a_\eta Q_{\eta,3}\right) \leqslant \ldots \tag{D.36}$$

Strict stochastic ordering may or may not apply between each weighted sum in Equation D.35, depending on conditions within each division (Section C.3 in Appendix C) and also across divisions (from Corollaries 15 and 16 applied to sets of $Q$-variables instead of $Y$-variables).

The weighting constants $a_1$,$a_2$,$a_3$, ... for the $Q$-variables need not be the same as those for the $R$-variables or those for the $Y$-variables, although they could be.

## D.4 Summary

The primary result derived in this appendix was Theorem 4, which showed that given two stochastically ordered sets (*divisions*) of random variables, each of which has the same ordering, then weighted sums of random variables taken across divisions are also stochastically ordered. Although this may seem an intuitive result, the proof of it was quite involved. The theorem is a very general result. It applies to stochastically ordered random variables, regardless of whether they are continuous, discrete or mixed, regardless of their specific distributional forms, and regardless of assumptions about correlations or independence among the random variables. Special cases covered by weighted sums of random variables include plain sums and arithmetic means.

Section D.2 presented corollaries to Theorem 4, that extended the general results of the theorem to weighted sums taken over an arbitrary number of divisions, and when there are an arbitrary number of random variables per division. The general weighted sums follow the same stochastic ordering that holds within each division, and the expected values of the weighted sums follow a similar (numerical) ordering. Conditions under which either strict or non-strict ordering of weighted sums occurred were given. It was shown that if any strictly ordered pair of random variables contributes to a pair of weighted sums of (strictly or non-strictly) stochastically ordered random variables, then the pair of weighted sums must also be strictly ordered.

Section D.3 showed how Theorem 4 and its corollaries applied to weighted sums of s.m.i.-transformed stochastically ordered random variables. The results held, even if the s.m.i.-transform was different for each division. Section D.3 also showed that stochastic ordering held for weighted sums of step function transforms of stochastically ordered random variables.

# Appendix E

# Non-linear least-squares regression of the FORA

The following is a development of the least-squares regression of FORA equations of the form of Equation 6.5 (p. 150). Below, $y_i$ denotes the $i^{th}$ point on an empirical FORA (i.e. the average empirical measure value for a combination-size of $i$), and $A_i$ denotes the regression estimate of $y_i$. Although the equations are developed with $\mathcal{A}$ as the measure in mind (hence the use of $A_1$ as a mnemonic), they could be derived from any measure of sensitivity. Anywhere that $\mathcal{A}$ may appear below, it can be replaced by the measure of interest.

Given a set of data from $m$ replications, let

$$\begin{aligned} y_1 &= \text{the first mean-}\mathcal{A}\text{ value} \\ y_2 &= y_1 + \delta_1 \\ y_3 &= y_1 + \delta_1 + \delta_2 \\ &\vdots \\ y_m &= y_1 + \sum_{j=2}^{m} \delta_j, \end{aligned}$$

where $\delta_1$, $\delta_2$ ... denote successive *increments* in the empirical FORA. Suppose that $\delta_j \simeq \kappa j^{\mu}$ for some empirical constants $\kappa$ and $\mu$. The constraint $\kappa > 0$ is desirable in order for the predicted values of $\mathcal{A}$ to *increase* as the number of replications increases. Furthermore, $\mu < -1$ is also desirable so that the regression series is constrained to converge as $m$ tends to $\infty$, particularly for measures of performance such as $\mathcal{A}$ that are constrained to take on finite values.

According to Equation 6.5, the predicted values are

$$\begin{aligned}
\widehat{y}_1 &= A_1 \\
\widehat{y}_2 &= A_1 + \kappa 2^{\mu} \\
\widehat{y}_3 &= A_1 + \kappa 2^{\mu} + \kappa 3^{\mu} \\
&\vdots \\
\widehat{y}_m &= A_1 + \kappa 2^{\mu} + \kappa 3^{\mu} + \ldots + \kappa m^{\mu}.
\end{aligned}$$

In general, the predicted value of the FORA after $i$ replications is

$$\widehat{y}_i = A_1 + \kappa \sum_{j=2}^{i} j^{\mu}, \qquad i \geqslant 2.$$

Fitting a regression model of this form means that

$$\begin{aligned}
y_i &= \widehat{y}_i + \widehat{e}_i \\
&= \begin{cases} A_1 + \widehat{e}_1 & i = 1 \\ A_1 + \kappa \sum_{j=2}^{i} j^{\mu} + \widehat{e}_i & i \geqslant 2, \end{cases}
\end{aligned}$$

where

$$\widehat{e}_i = \begin{cases} y_1 - A_1 & i = 1 \\ y_i - \left( A_1 + \kappa \sum_{j=2}^{i} j^{\mu} \right) & i \geqslant 2. \end{cases}$$

Applying a least-squares criterion to fit such a regression function to an empirical FORA involves finding values $A_1$, $\kappa$ and $\mu$ ($\kappa > 0$, $\mu < -1$) that minimise the total squared error, $\sum_{i=1}^{m} \widehat{e}_i^2$. The total squared error is

$$\begin{aligned}
\sum_{i=1}^{m} \widehat{e}_i^2 &= (y_1 - A_1)^2 + \sum_{i=2}^{m} \left( y_i - \left( A_1 + \kappa \sum_{j=2}^{i} j^{\mu} \right) \right)^2 \\
&= y_1^2 - 2A_1 y_1 + A_1^2 + \sum_{i=2}^{m} y_i^2 - 2 \sum_{i=2}^{m} y_i \left( A_1 + \kappa \sum_{j=2}^{i} j^{\mu} \right) + \ldots \\
&\quad \cdots + \sum_{i=2}^{m} \left( A_1^2 + 2A_1 \kappa \sum_{j=2}^{i} j^{\mu} + \kappa^2 \left( \sum_{j=2}^{i} j^{\mu} \right)^2 \right) \\
&= \sum_{i=1}^{m} y_i^2 - 2A_1 \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} A_1^2 - 2\kappa \sum_{i=2}^{m} y_i \sum_{j=2}^{i} j^{\mu} + \ldots \\
&\quad \cdots + 2A_1 \kappa \sum_{i=2}^{m} \sum_{j=2}^{i} j^{\mu} + \kappa^2 \sum_{i=2}^{m} \left( \sum_{j=2}^{i} j^{\mu} \right)^2 . \qquad \text{(E.1)}
\end{aligned}$$

Note that equal weighting is given to all of the data points, $y_i$. If unequal weightings were appropriate instead,[1] then they should be applied to Equation E.1 at this point.

Equation E.1 is a function defined in a 3-dimensional parameter space in which the data values, $y_i$, are fixed, and $A_1$, $\kappa$ and $\mu$ vary. The global minimum of the value of Equation E.1 is found somewhere in this parameter space, but where it lies is not clear.

The next step is to find equations for the mimimum taken with respect of each of the three variables. This means working out the partial derivatives of $\sum_{i=1}^{m} \widehat{e}_i^2$ with respect to $\kappa$, $\mu$ and $A_1$ in turn, setting each to zero, simplifying if possible and solving for $\kappa$, $\mu$ and $A_1$ respectively. First with respect to $\kappa$, from Equation E.1 it can be seen that

$$\frac{\partial}{\partial \kappa} \sum_{i=1}^{m} \widehat{e}_i^2 = -2 \sum_{i=2}^{m} y_i \sum_{j=2}^{i} j^{\mu} + 2A_1 \sum_{i=2}^{m} \sum_{j=2}^{i} j^{\mu} + 2\kappa \sum_{i=2}^{m} \left( \sum_{j=2}^{i} j^{\mu} \right)^2 . \qquad \text{(E.2)}$$

Setting this to zero, dividing by $-2$ and collecting terms gives

$$\sum_{i=2}^{m} (y_i - A_1) \sum_{j=2}^{i} j^{\mu} = \kappa \sum_{i=2}^{m} \left( \sum_{j=2}^{i} j^{\mu} \right)^2 . \qquad \text{(E.3)}$$

[1] Possibly to compensate for statistical factors that may affect FORA regression, as was discussed in Section 6.2.1.

Similarly from Equation E.1, differentiating the total squared-error with respect to $\mu$ gives

$$\begin{aligned}\frac{\partial}{\partial\mu}\sum_{i=1}^{m}\widehat{e}_i^2 &= -2\kappa\sum_{i=2}^{m}y_i\sum_{j=2}^{i}j^\mu\ln(j)+2A_1\kappa\sum_{i=2}^{m}\sum_{j=2}^{i}j^\mu\ln(j)+\ldots \\ &\quad \cdots+2\kappa^2\sum_{i=2}^{m}\left(\sum_{j=2}^{i}j^\mu\right)\left(\sum_{j=2}^{i}j^\mu\ln(j)\right). \qquad \text{(E.4)}\end{aligned}$$

Setting this to zero, dividing by $-2\kappa$ and collecting terms gives

$$\sum_{i=2}^{m}(y_i-A_1)\sum_{j=2}^{i}j^\mu\ln(j) = \kappa\sum_{i=2}^{m}\left(\sum_{j=2}^{i}j^\mu\right)\left(\sum_{j=2}^{i}j^\mu\ln(j)\right). \qquad \text{(E.5)}$$

Differentiating Equation E.1 with respect to $A_1$ gives

$$\frac{\partial}{\partial A_1}\sum_{i=1}^{m}\widehat{e}_i^2 = -2\sum_{i=1}^{m}y_i+2\sum_{i=1}^{m}A_1+2\kappa\sum_{i=2}^{m}\sum_{j=2}^{i}j^\mu. \qquad \text{(E.6)}$$

Setting this to zero and dividing by $-2$ gives

$$\sum_{i=1}^{m}(y_i-A_1) = \kappa\sum_{i=2}^{m}\sum_{j=2}^{i}j^\mu. \qquad \text{(E.7)}$$

In order to find a minimum, Equations E.3, E.5 and E.7 need to be solved simultaneously for $A_1$, $\kappa$ and $\mu$. Finding such a point does not guarantee that the point found *is* in fact a minimum. Whether it is a minimum or not could be found by deriving second order partial-derivatives, which are not presented here. It is preferable that all of the second order partial-derivatives of Equation E.1 are positive, because that would guarantee that the triplet $(A_1,\kappa,\mu)$ was a minimum (Courant, 1937). However, this is not necessarily the case. A quick check shows that all of the second order partial-derivatives are guaranteed to be positive *except for* $\frac{\partial^2}{\partial\kappa\partial\mu}\sum_{i=1}^{m}\widehat{e}_i^2$ (where $\frac{\partial^2}{\partial\kappa\partial\mu}=\frac{\partial^2}{\partial\mu\partial\kappa}$), and $\frac{\partial^2}{\partial\mu^2}\sum_{i=1}^{m}\widehat{e}_i^2$. Whether these latter derivatives are positive or otherwise depends on the data points, $y_i$, and the parameter triplet. Pragmatically, Equations E.3, E.5 and E.7 can be solved simultaneously, and the result used nevertheless. From the data sets analysed in Chapters 6, 7 and 8, this approach seems sufficient to achieve good fits to empirical FORAs, indicating that a minimum is found rather than a maximum.

Note the expression $(y_i - A_1)$ in the left-hand sides of Equations E.3, E.5 and E.7 is what $\kappa \sum_{j=2}^{i} j^{\mu}$ (for $2 \leqslant i \leqslant m$) approximates, namely the increments in performance. If the data points are exactly of the form in Equation 6.2, then the left-hand sides of Equations E.3, E.5 and E.7 are exactly equal to the right-hand sides of the same equations and, the sum of the squared residuals in Equation E.1 is zero.

Assuming that a particular parameter triplet $(A_1, \kappa, \mu)$ results in a minimum, then Equation E.6 will equal zero and Equation E.7 will hold. If so, then this gives a simple expression for $A_1$ in terms of only $\kappa$, $\mu$, and the data points, $y_i$. A rearrangement of Equation E.7 gives[2]

$$\begin{aligned} \sum_{i=1}^{m} A_1 &= \sum_{i=1}^{m} y_i - \kappa \sum_{i=2}^{m} \sum_{j=2}^{i} j^{\mu}, \\ &= y_1 + \sum_{i=2}^{m} \left( y_i - \kappa \sum_{j=2}^{i} j^{\mu} \right), \end{aligned}$$

which implies that

$$A_1 = \frac{1}{m} \left( y_1 + \sum_{i=2}^{m} \left( y_i - \kappa \sum_{j=2}^{i} j^{\mu} \right) \right). \tag{E.8}$$

It can be seen that if the log-log-plot is exactly linear and Equation 6.2 holds exactly, then Equation E.8 reduces to

$$\begin{aligned} A_1 &= \frac{1}{m} \left( y_1 + \sum_{i=2}^{m} \left( y_i - \sum_{j=2}^{i} \delta_j \right) \right) \\ &= \frac{1}{m} \left( y_1 + \sum_{i=2}^{m} y_1 \right) \\ &= y_1. \end{aligned} \tag{E.9}$$

Similarly, expressions for $\kappa$ in terms of only $A_1$, $\mu$, and the data points, $y_i$, can be derived from Equations E.3, E.5 and E.7. The three expressions (which are not given here) are all different and not as simple as Equation E.8. There does not seem to be an obvious simplification that gives an expression for $\mu$.

[2]Note the lower limits in the sums across $i$.

# Appendix F

# Computational aspects of all combinations analysis

Some of the computational details of all combinations analysis and of FORA regression are described here. Much of this material relates to a computer program called PCTRIAL, which was written to perform various analyses described in this thesis. It was developed primarily for ACA, for calculating FORAs and estimating FORA asymptotes. Code for calculating mean ROC curves, transform-average GOC curves and transfer functions was also integrated into the program.

**The PCTRIAL computer program.** PCTRIAL performs ACA, computes an empirical FORA for each of the measures $\mathcal{A}$, $d'$, $\mathcal{D}_2$ and $P(C)$, and estimates asymptotic performance for each measure. It can also run multiple random resamplings of sets of replications from a larger data set, calculate FORAs and asymptotes and collate results across resamplings, as reported in Section 7.4.

PCTRIAL was written in Borland Pascal 7.0 using the 16-bit Borland IDE (the program development environment), and operates as a DOS protected mode program on an IBM-compatible personal computer. It made use of two code units developed by Linton Miller called BIGARRAYS and PPMDUMP. BIGARRAYS allowed circumvention of the Borland Compiler's 64 kilobyte limitations. PPMDUMP saved graphics screens to file. PCTRIAL is not general-purpose. A moderate amount of redevelopment and dedicated code is needed for each new data set and the program is dependent on the Borland libraries.

Miller (1998; cited in Lapsley Miller, 1999) independently wrote his own ACA program. It was written in C and implemented on Unix systems. Results from Miller's program cross-checked successfully against results from PCTRIAL. Most of the experimental FORAs reported in detail in Chapters 6, 7 and 8 were calculated using PCTRIAL, except those in Section 8.4 which came from Miller's program.

**GOC analysis.** The transform-average GOC curves presented in Chapter 3 were calculated using the generalised GOC algorithm described in Section 2.4.2. This was implemented in PCTRIAL using an insert-sort algorithm (Aho, Hopcroft, & Ullman, 1983). Apart from the transform-average GOC curves, and the use of arithmetic mean ratings in Chapter 4, *all other GOC analyses* in this thesis, including the FORAs in Chapters 6 to 8, were computed using *sums of integer ratings* under the conventional GOC algorithm outlined in Section 2.4.2. This was primarily because of speed. PCTRIAL calculated GOC curves using an index-sort algorithm (or bin-sort; Aho et al., 1983) using sums of integer-ratings as the sorting-key. Index-sorting is the fastest type of sorting algorithm, requiring only $O(n)$ steps to sort a set of size $n$. The restriction on using an index-sort is that the sorting-key must be integer-valued, since array indices are integer-valued. The computation time required to obtain any single GOC curve, such as the GOC curve in Figure 2.5, is negligible under either GOC algorithm (Section 2.4.2). If millions of GOC curves are calculated, as was the case in Chapters 6, 7 and 8, then the computation time per GOC curve is important.

**Calculation of $\mathcal{A}$.** Some measures of sensitivity, including $\mathcal{A}$, can be calculated for a GOC or ROC curve without calculating the curve itself. As noted in Sections 1.3 and 2.4.2, empirical ROC and GOC curves can be calculated by cumulating tallies of the number of times each rating or sum-of-ratings occurred. The cumulative tallies are divided by the number of stimuli per event in order to derive empirical hit and false alarm rates. If the aim of data analysis is to work out the area under the curve, but not the curve itself, much computing time can be saved because hit and false alarm rates do not need to be calculated explicitly. In Borland Pascal at least, it is several times faster to perform arithmetic on integer-typed variables than on floating-point variables. For speed, PCTRIAL used integer additions and multiplications to arrive at an integer-valued, scaled version of $\mathcal{A}$, $I_c$. The area under the curve was then given by $\mathcal{A} = I_c/(2n_{\mathrm{SN}}n_{\mathrm{N}})$, where $n_{\mathrm{SN}}$ and $n_{\mathrm{N}}$ are the number of stimuli per event for the $SN$ and $N$ events respectively.[1] The efficiency in calculating $\mathcal{A}$ also has benefits for calculating $d'$ and $\mathcal{D}_2$ when the latter measures are derived as transforms of $\mathcal{A}$.

**All combinations analysis.** There is a major redundancy in ACA that can be used to substantially reduce computation time. The redundancy has to do with calculating GOC measures by using complementary combinations. It was first mentioned with respect to partial-ACA in Section 7.3.2, and is illustrated here by means of an example.

Assume $m = 6$ replications were run in an experiment and these are numbered from 1 to 6. GOC analysis requires calculating the sum-of-ratings[2] for each stimulus for each

[1] $I_c$ is similar to, but not necessarily identical to, twice the Mann-Whitney U statistic (Bamber, 1975). The potential discrepancy between the two values has to do with tied rating values.

[2] The description here is in terms of sums of raw ratings, but could apply equally well to sums or averages of transformed ratings in transform-averaged GOC analysis.

| combinations of size 2 | complementary combinations of size 4 |
|---|---|
| 1 2 | 3 4 5 6 |
| 1 3 | 2 4 5 6 |
| 1 4 | 2 3 5 6 |
| 1 5 | 2 3 4 6 |
| 1 6 | 2 3 4 5 |
| 2 3 | 1 4 5 6 |
| . | . |
| . | . |
| . | . |
| 4 6 | 1 2 3 5 |
| 5 6 | 1 2 3 4 |

TABLE F.1: A list showing combinations of size 2 taken from the set of integers from 1 to 6, along with their complementary combinations of size 4. The combinations of size 2 are in lexicographic order, while their complementary combinations are in reverse lexicographic order.

combination of replications taken from the data set. Assume that GOC curves for all combinations of size $\xi = 2$ are calculated. Table F.1 lists combinations of size 2 in lexicographic order. Each subset of size 2 has a complementary set composed of the 4 remaining replications. Note that the list of complementary combinations is an exhaustive list of combinations of size 4, which is given in reverse-lexicographic order. The general rule is that each possible combination of size $\xi$ taken from a set of size $m$ is complementary to a combination of size $m - \xi$.

The only exception to the general rule is when $\xi$ equals the total number of replications, $m$, in which case there is only 1 combination (the set of all integers from 1 to $m$) and it has no complement. There is only one GOC curve of combination-size $m$ and it requires that the total $m$-replication sum of ratings per stimulus be calculated for all stimuli. In PCTRIAL, the sums for this particular GOC curve are calculated first and stored in an array before the rest of the ACA is done. The sum of ratings per stimulus for a complementary combination of size $m - \xi$ is equal to the total sum of ratings per stimulus (stored in the array) *minus* the sum of ratings for the combination of size $\xi$. For the example in Table F.1, given the sum for all 6 replications, if the sum for a combination of size 2 is calculated, then the sum for a combination of size 4 can be calculated by performing one subtraction rather than 3 additions. This is a saving of two arithmetic operations at the expense of extra storage of the six-replication sums. The saving is small when $m$ is small, but for larger values (such as $m = 24$ for Taylor et al.'s (1991) experiment in Chapter 6), the savings are considerable. The end result is that sensitivity measures for two GOC

| first half of the combinations of size 3 | complementary combinations of size 3 |
|---|---|
| 1 2 3 | 4 5 6 |
| 1 2 4 | 3 5 6 |
| 1 2 5 | 3 4 6 |
| 1 2 6 | 3 4 5 |
| 1 3 4 | 2 5 6 |
| 1 3 5 | 2 4 6 |
| 1 3 6 | 2 4 5 |
| 1 4 5 | 2 3 6 |
| 1 4 6 | 2 3 5 |
| 1 5 6 | 2 3 4 |

TABLE F.2: All combinations of size 3 taken from a set of integers from 1 to 6. The first 10 combinations are shown in the left-hand column, while their complementary combinations are shown in the right-hand column.

curves can be calculated at a time rather than just one. It is important to remember that the summation and subtraction is done on a *per stimulus basis*.

There is a special case when $m$ is an even number and $\xi = \frac{m}{2}$. Table F.2 shows a listing of all combinations of size 3 taken from a set of size 6. The first column gives the first 10 out of ${}^6C_3 = 20$ combinations in lexicographic order. The second column lists complementary combinations, which are the second 10 out of 20 combinations given in reverse-lexicographic order. As before, sums-of-ratings from combinations in the first column are subtracted from rating total-sums to give the sums for the second column. This results in a saving when $m$ is even (there is no special saving to be had if $m$ is odd).

**Other efficiencies in ACA.** Complementary combinations aside, other efficiencies are possible when computing ACA. In ACA, all combinations of replications must be calculated, usually one at a time in lexicographic order. Much running time can be saved when a sum-of-ratings from a previous combination can be reused to calculate the sum-of-ratings for a subsequent combination. There are many ways to do this, but Miller (1999, personal communication) recommends dynamic programming methods for reasons of efficiency (and hence speed).

**Computing FORA regression parameters.** As described in Sections 6.2.3 and 6.2, the non-linear least-squares FORA regression of the form of Equation 6.5 requires finding a parameter triplet $(A_1, \kappa, \mu)$ that minimises the sum of the squared residuals, which is achieved by simultaneously solving Equations E.3, E.5 and E.7 given in Appendix E. The simultaneous solution was achieved in PCTRIAL by implementing a minimization search routine called the *Variable Metric Method*, which is described in Nash (1979, Chpt. 15). The algorithm uses a mixture of gradient descent and linear search methods, and generally executes very quickly.

The Variable Metric Method provides only one particular method for solving for extrema. Other algorithms may converge on slightly different solutions for the same data set, which would result in somewhat different estimated asymptote values. Results may also depend on the programming language and compiler, and on the floating-point precision that is used.

**Computing the Riemann zeta function.** Once a parameter triplet, $(A_1, \kappa, \mu)$, has been found via the Variable Metric Method, it remains to calculate the asymptote. The asymptote is given in Equation 6.7 as $A_\infty = A_1 + \kappa\,(\zeta(-\mu) - 1)$, where $\zeta(x)$ is the Riemann zeta function. This function has two series expansions which are

$$\zeta(x) = \sum_{j=1}^{\infty} j^{-x}, \quad x > 1 \tag{F.1}$$

$$= \frac{1}{1 - 2^{1-x}} \sum_{j=1}^{\infty} (-1)^{j+1} j^{-x} \tag{F.2}$$

(Gradshteyn & Ryzhik, 1965, Equations 9.522 (1) and (2)). Equation 6.6 (which states that $A_\infty = A_1 + \kappa \sum_{j=2}^{\infty} j^{\mu}$) uses the series in Equation F.1 to calculate $A_\infty$ by substituting $-\mu$ for $x$ in the argument, where the condition $x > 1$ equates to $\mu < -1$. The first term in the series in Equation F.1 is always equal to 1, which is not part of the series in Equation 6.6, and therefore 1 is subtracted from $\zeta(-\mu)$ in Equation 6.7.

The series terms in Equation F.1 are positive, hence the partial sums of the series approach the asymptote from below. This makes it hard to tell how close any finite partial sum is to the asymptote. The terms in Equation F.2 are of alternating sign, and the partial sums oscillate around the asymptote. The absolute difference between the asymptote and the partial sums of Equation F.2 decreases as more terms are added. Together with the alternating sign of the terms, this means that the asymptote can be approached to any desired tolerance (which was set to $5 \times 10^{-7}$ in PCTRIAL). Furthermore, Equation F.2 converges more quickly than Equation F.1 and using Equation F.2 to calculate $A_\infty$ makes it possible to work out in advance the number of terms required to achieve a given accuracy. For a given number of terms, it is even more accurate to take the average of the last two partial sums in the series, because one of these terms is always above the asymptote while the other is always below, so that their average is better than either term individually.

# Appendix G

# FORA values and regression parameters

Tables of FORA values and regression parameters for the various data sets presented in Chapters 6, 7 and 8 are given here. The contents of each table depends on what is covered in the main text. Each table gives values for the first and last points on each FORA, regression parameter values $A_1$, $\kappa$ and $\mu$ and the asymptote based on these values.[1] Theoretical performance values are given where known. The square of the correlation coefficient, $r^2$, is also given. All values of $\mathcal{D}_2$ are expressed in bits (units of information).

[1] $A_1$ represents the value of the first point on a regression-FORA for any measure of sensitivity, either $\mathcal{A}$, $d'$, $\mathcal{D}_2$ or $P(C)$, as appropriate.

| Measure | GOC $\xi = 1$ | GOC $\xi = 24$ | Asymp. | Theory | Regression parameters $A_1$ | $\kappa$ | $\mu$ | log-log $r^2$ |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{A}$ | 0.7394 | 0.8483 | 0.8585 | 0.8550 | 0.738663 | 0.166007 | -1.924461 | 0.9977 |
| $d'$ | 0.9132 | 1.4553 | 1.5310 | 1.4966 | 0.907713 | 0.699176 | -1.793913 | 0.9959 |
| $\mathcal{D}_2$ | 0.1778 | 0.3858 | 0.4222 | 0.4029 | 0.174596 | 0.246004 | -1.725072 | 0.9937 |
| $P(C)$ | 0.6793 | 0.7308 | 0.7358 | 0.7308 | 0.677812 | 0.142212 | -2.348421 | 0.7321 |

TABLE G.1: FORA values and regression parameters for four measures of sensitivity for Taylor et al.'s (1991) continuous rating scale experiment presented in Chapter 6. Combination-sizes $\xi = 1$ and 24 are the first and last points on the FORA, respectively. Estimated asymptotes are based on the given values of the regression parameters, $A_1$, $\kappa$ and $\mu$, and $r^2$ is the square of the correlation coefficient for each measure. The theoretical value for each measure is also given.

| Observer | Replication set | Figure | No. of reps ($m$) | Ave. $\mathcal{A}$ $\xi = 1$ | Ave. $\mathcal{A}$ $\xi = m$ | Asymp. | Regression parameters $A_1$ | $\kappa$ | $\mu$ | log-log $r^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | All | 7.4 | 25 | 0.8017 | 0.8567 | 0.8596 | 0.802134 | 0.093229 | -2.030904 | 0.9960 |
| 2 | 1st 25 | 7.5 | 25 | 0.7995 | 0.9121 | 0.9270 | 0.799490 | 0.138040 | -1.773156 | 0.9990 |
| 2 | 2nd 25 | 7.5 | 25 | 0.7882 | 0.8919 | 0.9047 | 0.788178 | 0.130740 | -1.793912 | 0.9993 |
| 2 | 3rd 25 | 7.5 | 25 | 0.7891 | 0.9018 | 0.9168 | 0.789276 | 0.134606 | -1.758871 | 0.9996 |
| 1 & 2 | 25 each[1] | 7.6 | 50 | 0.8006 | 0.8851 | 0.8884 | 0.801093 | 0.124602 | -1.944347 | 0.9975 |
| 2 | All[2] | 7.7 | 75 | 0.7922 | 0.9099 | 0.9160 | 0.792340 | 0.134380 | -1.775623 | — |
| 2 | All[3] | 7.8 | 75 | 0.7922 | 0.9099 | 0.9160 | 0.791970 | 0.136774 | -1.783845 | 0.9992 |

TABLE G.2: FORA values and regression parameters for Whitmore et al.'s (1993) SIFC amplitude discrimination experiment presented in Chapter 7. The figure in which a FORA appears and the number of replications involved, $m$, are given. Combination-sizes $\xi = 1$ and $m$ are the first and last points on the FORA, respectively. Estimated asymptotes are based on the given values of the regression parameters, $A_1$, $\kappa$ and $\mu$, and $r^2$ is the square of the correlation of the log-log data points for each measure. No theoretical values are given because the theory is not known.

[1]All 25 replications from Observer 1 and the first 25 replications from Observer 2 (outer points and estimated inner points).

[2]Outer points only, hence $r^2$ for Figure 7.7(b)is omitted, since any correlation is spurious.

[3]Outer points and estimated inner points.

| Measure | data set | GOC $\xi = 1$ | GOC $\xi = 64$ | Asymp. | Theory | Regression parameters $A_1$ | $\kappa$ | $\mu$ |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{A}$ | pbb | 0.7108 | 0.9390 | 0.9522 | 0.9588 | 0.709699 | 0.268060 | -1.785424 |
| $\mathcal{A}$ | db | 0.7201 | 0.9435 | 0.9537 | 0.9588 | 0.719217 | 0.285479 | -1.842825 |
| $\mathcal{A}$ | 64r | 0.8004 | 0.9483 | 0.9546 | 0.9588 | 0.799892 | 0.198505 | -1.875270 |
| $d'$ | pbb | 0.7903 | 2.1867 | 2.4817 | 2.4556 | 0.782386 | 0.981062 | -1.471322 |
| $d'$ | db | 0.8287 | 2.2407 | 2.4934 | 2.4556 | 0.820910 | 1.058591 | -1.508397 |
| $d'$ | 64r | 1.2007 | 2.3037 | 2.4619 | 2.4556 | 1.194593 | 0.917445 | -1.566729 |
| $\mathcal{D}_2$ | pbb | 0.1372 | 0.6685 | 0.7917 | 0.7520 | 0.131215 | 0.363130 | -1.452533 |
| $\mathcal{D}_2$ | db | 0.1484 | 0.6864 | 0.7824 | 0.7520 | 0.142657 | 0.407515 | -1.511043 |
| $\mathcal{D}_2$ | 64r | 0.2841 | 0.7066 | 0.7527 | 0.7520 | 0.280978 | 0.393380 | -1.633757 |
| $P(C)$ | pbb | 0.6988 | 0.8531 | 0.8825 | 0.8750 | 0.691071 | 0.116340 | -1.491838 |
| $P(C)$ | db | 0.7019 | 0.8630 | — | 0.8750 | 0.726064 | 0.035371 | -0.978434 |
| $P(C)$ | 64r | 0.7235 | 0.8750 | 0.8869 | 0.8750 | 0.722766 | 0.150813 | -1.683052 |

TABLE G.3: FORA values and regression parameters for Lapsley Miller et al.'s (1998) discrete case 2IFC experiments presented in Section 8.1. Values are given for all four measures of sensitivity. The data sets are denoted "pbb" for push-button binary-decision, "db" for derived binary-decision and "64r" for 64-point rating scale. Combination-sizes $\xi = 1$ and 64 are the first and last points on the FORA, respectively. Estimated asymptotes are based on the given values of the regression parameters, $A_1$, $\kappa$ and $\mu$, and the theoretical value for each measure is also given. FORAs based on $\mathcal{A}$ were presented in Figure 8.3. Log-log plot $r^2$ values are distorted by the fact that the the data points fall into two distant regions. For this reason, $r^2$ is not given here. There is no $P(C)$ asymptote for the derived binary-decision FORA, because $\mu = -0.9784$ was greater than $-1$, and the regression-FORA could not converge.

| Observer | GOC $\xi = 1$ | GOC $\xi = 16$ | $\mathcal{A}$ Asymp. | Regression parameters $A_1$ | $\kappa$ | $\mu$ | $r^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.6667 | 0.7755 | 0.8058 | 0.666216 | 0.120711 | -1.651875 | 0.9987 |
| 2 | 0.7652 | 0.8174 | 0.8226 | 0.765157 | 0.093935 | -2.036969 | 0.9825 |
| 3 | 0.6640 | 0.7962 | 0.8417 | 0.663162 | 0.135062 | -1.586918 | 0.9939 |
| 4 | 0.8077 | 0.8866 | 0.8957 | 0.807491 | 0.130283 | -1.967150 | 0.9999 |

TABLE G.4: FORA values and regression parameters for Lapsley Miller et al.'s (1998) continuous case 2IFC experiments presented in Section 8.2. Combination-sizes $\xi = 1$ and 16 are the first and last points on the FORA, respectively. Estimated asymptotes are based on the given values of the regression parameters, $A_1$, $\kappa$ and $\mu$, and $r^2$ is the square of the correlation coefficient for each observer. These FORAs were presented in Figure 8.6. Values are given for $\mathcal{A}$ only, and no theoretical value is given because the theory is not known.

| Observer | Measure | SNR (dB) | $\xi = 1$ | $\xi = 8$ | Asymp. | Regression parameters $A_1$ | $\kappa$ | $\mu$ | log-log $r^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | $\mathcal{A}$ | −5 | 0.5492 | 0.5732 | 0.5829 | 0.549152 | 0.034621 | -1.742952 | 0.9827 |
| 1 | $\mathcal{A}$ | 0 | 0.6108 | 0.6561 | 0.6737 | 0.610705 | 0.065509 | -1.750562 | 0.9987 |
| 1 | $\mathcal{A}$ | 4 | 0.7548 | 0.8362 | 0.8581 | 0.754786 | 0.139896 | -1.909579 | 0.9993 |
| 1 | $\mathcal{A}$ | 8 | 0.8897 | 0.9472 | 0.9546 | 0.889801 | 0.143660 | -2.265585 | 0.9996 |
| 1 | $\mathcal{A}$ | 12 | 0.9651 | 0.9917 | 0.9929 | 0.965183 | 0.108520 | -2.766443 | 0.9949 |
| 2 | $\mathcal{A}$ | −5 | 0.5424 | 0.5679 | 0.5910 | 0.542258 | 0.026269 | -1.445443 | 0.9857 |
| 2 | $\mathcal{A}$ | 0 | 0.6259 | 0.6859 | 0.7061 | 0.625936 | 0.091680 | -1.805284 | 0.9978 |
| 2 | $\mathcal{A}$ | 4 | 0.7906 | 0.8695 | 0.8877 | 0.790582 | 0.146641 | -1.981912 | 0.9999 |
| 2 | $\mathcal{A}$ | 8 | 0.9224 | 0.9618 | 0.9654 | 0.922482 | 0.118298 | -2.447972 | 0.9994 |
| 2 | $\mathcal{A}$ | 12 | 0.9900 | 0.9984 | 0.9985 | 0.990016 | 0.046553 | -3.102597 | 0.9842 |
| 1 | $d'$ | −5 | 0.1749 | 0.2609 | 0.2962 | 0.174786 | 0.123256 | -1.737135 | 0.9825 |
| 1 | $d'$ | 0 | 0.3982 | 0.5684 | 0.6381 | 0.397987 | 0.239319 | -1.726866 | 0.9986 |
| 1 | $d'$ | 4 | 0.9768 | 1.3842 | 1.5389 | 0.976331 | 0.593844 | -1.758856 | 0.9991 |
| 1 | $d'$ | 8 | 1.7362 | 2.2887 | 2.4427 | 1.736160 | 0.936085 | -1.895679 | 0.9996 |
| 1 | $d'$ | 12 | 2.5771 | 3.3892 | 3.5988 | 2.577311 | 1.412892 | -1.923481 | 0.9980 |
| 2 | $d'$ | −5 | 0.1508 | 0.2420 | 0.3247 | 0.150342 | 0.093557 | -1.443501 | 0.9856 |
| 2 | $d'$ | 0 | 0.4544 | 0.6847 | 0.7701 | 0.454318 | 0.336998 | -1.765009 | 0.9977 |
| 2 | $d'$ | 4 | 1.1448 | 1.5894 | 1.7468 | 1.144506 | 0.669772 | -1.788830 | 0.9998 |
| 2 | $d'$ | 8 | 2.0123 | 2.5059 | 2.6013 | 2.012287 | 1.010426 | -2.071087 | 0.9999 |
| 2 | $d'$ | 12 | 3.2981 | 4.1624 | 4.3555 | 3.299201 | 1.585293 | -1.977719 | 0.9854 |

TABLE G.5: FORA values and regression parameters for the 2IFC amplitude discrimination experiment in Section 8.3. Combination-sizes $\xi = 1$ and 8 are the performance values for the first and last FORA points, respectively. Estimated asymptotes are based on the given values of the regression parameters, $A_1$, $\kappa$ and $\mu$, and $r^2$ is the square of the correlation coefficient for each observer. Values are given for $\mathcal{A}$ and $d'$ only. No theoretical value is given because the theory is not known.