Ashby, F. G., & Wenger, M. J. (2023).
Statistical decision theory.
In F. G. Ashby, H. Colonius, & E. Dzhafarov
(Eds.), *The new handbook of mathematical psychology, Volume 3* (pp. 265–310).
Cambridge University Press.

Contents

7

Statistical Decision Theory

F. Gregory Ashby[a] and Michael J. Wenger[b]

7.1 Introduction

In 2002, Estes referred to signal detection theory (SDT) as "the most towering achievement of basic psychological research in the last half century" (p. 15). SDT is, by far, the most dominant model in psychophysics, and its multidimensional generalization has become the default approach for defining and studying perceptual interactions. The name "signal detection theory" refers to applications of the theory to tasks in which only one stimulus dimension is relevant, and the most common version requires participants to detect a signal embedded in noise. Tasks that require attention to more than one stimulus dimension typically require a decision more complex than simple detection – for example, the participant may be required to identify the presented stimulus uniquely, or assign it to a predetermined category. In such cases, the same statistical model is more appropriately called general recognition theory (GRT). We refer to both approaches by the term *statistical decision theory.*

This chapter reviews statistical decision theory, beginning with its origins, laying out its foundations in one dimension and its extension to two or more dimensions. We describe applications of the theory to identification and classification tasks, to the perception of configurality and holism, to the modeling of response times (RTs), and finally we consider extensions to neuroscience. An overarching theme of this chapter is that statistical decision theory provides a consistently evolving, general and powerful approach to modeling decision processes involved in sensation, perception, and cognition.

[a] University of California, Santa Barbara, USA
[b] University of Oklahoma, USA

7.2 Historical Precedents

Statistical decision theory emerged when two simple propositions were applied to a new experimental paradigm that eventually formed the foundation of psychophysics and much of experimental psychology. The first of these is the proposition that one can experience the qualia of a known stimulus (such as light) even in the absence of that stimulus. Perhaps the most famous example of this is the Helmholtz (1867) thought experiment on phosphenes: mechanical pressure on the eye causes the subjective experience of patterns of light even in a dark room.[1] Similarly, one can fail to experience the qualia of a known stimulus even when that stimulus is present (e.g., a light is on, but may be too dim to see). It appears that thinking about such possibilities was at the root of the classic two-alternative forced-choice design, and that thoughts about these possibilities are evident in work by both Fechner and Thurstone (Fechner, 1860; Link, 1994; Wixted, 2020).

The second of the two simple propositions is the idea that encoded psychological information may be a combination of a fixed value and random error. The formal notion of this possibility in human measurement can be traced at least to the work of Gauss (Dunnington, Gray, & Dohse, 2004), and the general notion of random variation in subjective human experience dates at least to the work of Cattell (Fullerton & Cattell, 1892) and Thurstone (Thurstone, 1927a, 1927b). However, the formal treatment of randomness in support of decision-making, as it has come to be expressed in statistical decision theory, emerged from the (at times contentious) debates that Fisher had with Neyman and Pearson (Fisher, 1955; Neyman & Pearson, 1933). In particular, Neyman and Pearson's distinction between Type I and Type II errors – corresponding to false alarms and misses, respectively – was offered as a refinement to Fisher's notion of a p-value, which itself had originally been proposed as an objective, though informal index of the level of trust in a null hypothesis (Lenhard, 2006).

The initial linking of these two simple propositions occurred in early work on radar and sonar and other areas of electronics and electrical engineering. It appears that the basic vocabulary of SDT – hits, misses, false alarms, and correct rejections – emerged from the World War II need, for example, to determine whether to drop a bomb or a depth charge on a possible enemy submarine (Marcum, 1947). Likewise, as noted by Wixted (2020), the idea that the probabilistic behavior of photographic film and television tubes might provide a model for the human visual system had already been

[1] Curiosity about phosphenes predates Helmholtz, as sketches of phosphenes can be found in Newton's notes (http://cudl.lib.cam.ac.uk/view/MS-ADD-0397).

considered in electrical engineering (Rose, 1942, 1948). The explicit merging of these ideas and their application to the analysis of both human behavior and the performance of engineered systems appears to have occurred at about the same time at MIT and the University of Michigan (Creelman, 2015; Peterson, Birdsall, & Fox, 1954; Peterson & Birdsall, 1953; Van Meter & Middleton, 1954).

In each of these contexts, the canonical experiment includes trials in which a stimulus or signal is or is not present and the observer or system is required to respond that the signal is present or absent. This task inspired the name signal detection theory, and almost all modern applications of SDT are either to this task or to the logically equivalent two-stimulus identification task, which we consider in detail in the next section. In fact, Link (1994) rightly noted that the use of this canonical task goes back at least to Fechner's foundational work on psychophysics. As we will see, this simple experiment provides a powerful and general framework for understanding how signals are processed – either by biological or engineered systems.

To illustrate the power and generality of this accomplishment (and to reflect on Estes' evaluation), we obtained rough estimates of the number of publications that used SDT in audition and vision, in five-year increments between 1955 and early 2020.[2] We contrasted these data with the number of Ph.D.s awarded in experimental, cognitive, and human factors psychology, along with the number of Ph.D.s awarded in electrical, electronics, and communications engineering for that same range of years.[3]

Figure 7.1a plots the cumulative number of publications in audition and vision that include SDT, along with the number of Ph.D.s awarded in psychology and engineering. This presentation is somewhat misleading, so Figure 7.1b plots the same data in terms of relative cumulative number (i.e., dividing the value of each data series at time t by the value at the starting point, 1960). It becomes apparent that the increase in the use of SDT is not simply due to an increase in the number of scientists who could potentially use SDT. This powerfully underscores Estes' estimate of SDT as a towering achievement. With this historical context in mind, we now consider the details.

[2] Searches were performed using Google Scholar. The search for publications in audition was performed using "auditory OR audition OR perception "signal detection theory" -vision -visual" and the search in vision was performed using "vision OR visual OR perception "signal detection theory" -auditory -audition."

[3] National Science Foundation, National Center for Science and Engineering Statistics, Survey of Earned Doctorates.

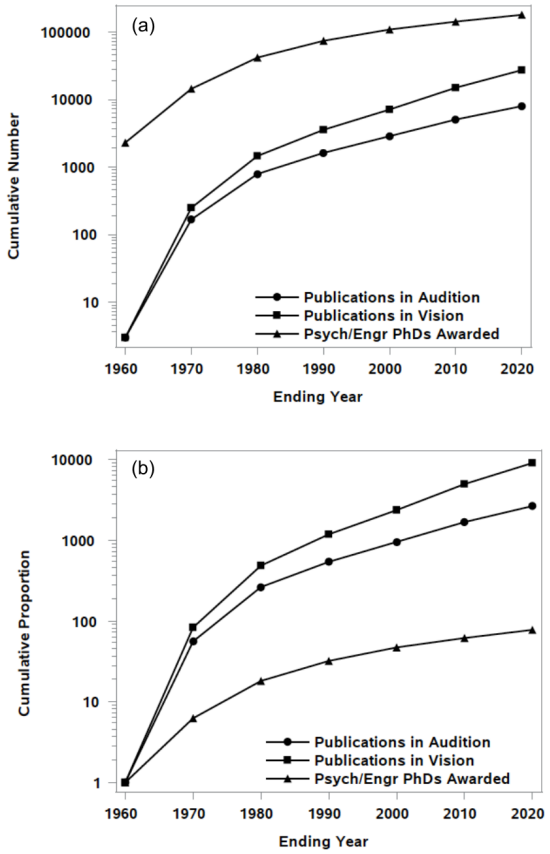

Figure 7.1 (a) Cumulative publications citing signal detection theory in audition and vision, relative to cumulative PhDs awarded in sub-disciplines of psychology and engineering, 1960-2020. (b) Relative increase in publications citing signal detection theory in audition and vision, and relative increase in PhDs awarded in sub-disciplines of psychology and engineering, 1960-2020.

7.3 One Dimension: Signal Detection Theory

The most common application of SDT is to a two-stimulus identification task – that is, a task with two stimuli and two uniquely identifying responses.[4] On each trial, the observer's task is to identify the single presented stimulus by emitting the appropriate response. In the original applications, the two stimuli were pure noise (N) and a signal of some type embedded in

[4] See Macmillan and Creelman (2005) for an excellent comprehensive treatment of the practicalities of using signal detection theory.

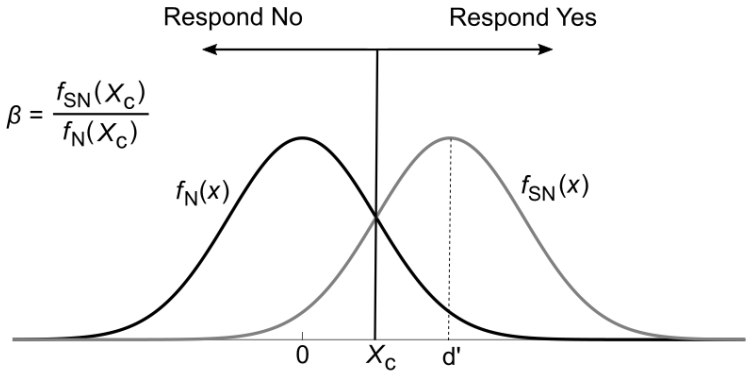

Figure 7.2 The normal, equal-variance, SDT model.

noise (SN). The observer's task was to indicate whether or not a signal was presented by responding YES or NO.

The standard SDT model for this YES-NO detection task is illustrated in Figure 7.2. The model assumes that performance in this task is based on a single sensory value, denoted by **X**. As described earlier, a fundamental assumption is that all sensations are inherently noisy, and thus **X** is a random variable. In the YES-NO detection task where the stimuli are N and SN, **X** represents sensory magnitude – for example, loudness with auditory stimuli, or brightness with visual stimuli. The probability density function (pdf) describing the distribution of sensory values on N trials is denoted by $f_{\mathrm{N}}(x)$ and $f_{\mathrm{SN}}(x)$ describes this distribution on SN trials. In Figure 7.2, both of these distributions are normal with the same variance. This normal, equal-variance model is the most commonly used model in signal detection analysis, but any distributions are possible.

Another fundamental assumption of SDT is that there is no fixed threshold on sensation that determines whether or not an observer will detect a signal. Instead the observer is assumed to set a criterion value, denoted by X_{C}, and then use the following decision rule:

$$\text{Respond YES if } \mathbf{X} > X_{\mathrm{C}};\ \text{Otherwise respond NO.} \tag{7.1}$$

Unlike the classical notion of a fixed threshold, the SDT criterion is under the observer's control. The observer is assumed to choose the value of X_{C} in a way that is typically assumed to depend on the costs of the two types of errors (i.e., misses and false alarms), the benefits of the two types of correct responses (i.e., hits and correct rejections), and on the N and SN base rates.

Thus, in SDT, control of the criterion is relegated to decision processes, whereas the classical account assumed a fixed threshold for sensation that was a feature of sensory systems.

The response accuracy data are typically reported in a confusion matrix that includes a row for every stimulus and a column for each response. The entry in row i and column j is the number of stimulus i trials for which the observer responded j. When there are only two stimuli and two responses, then the confusion matrix is 2×2. The entries in row i add to the number of stimulus i trials in the experiment, and therefore do not depend on the data. As a result, each row includes only one degree of freedom (i.e., only one independent data value), so no information is lost if only one entry in each row is reported. The standard is to report the entries in the column associated with the YES response. These are used to estimate the probability of a false alarm (i.e., responding YES on N trials) and the probability of a hit (responding YES on SN trials). From Figure 7.2 it is easily seen that

$$P(\mathrm{FA}) = 1 - F_{\mathrm{N}}(X_{\mathrm{C}}), \tag{7.2}$$

where $F_{\mathrm{N}}(X_{\mathrm{C}})$ is the cumulative distribution function of the N distribution, evaluated at X_{C}. Similarly,

$$P(\mathrm{H}) = 1 - F_{\mathrm{SN}}(X_{\mathrm{C}}). \tag{7.3}$$

In any two-stimulus identification task, the data have two degrees of freedom [e.g., $P(\mathrm{H})$ and $P(\mathrm{FA})$]. The SDT model shown in Figure 7.2 has two free parameters – the location of the response criterion, denoted by X_{C}, and the distance between the means of the N and SN distributions in standard deviation units, denoted by d'. If the normal, equal-variance model is assumed then X_{C} and d' can be estimated by inverting Eqs. 7.2 and 7.3. Specifically, X_{C} is estimated by inverting inverting Eq. 7.2 to produce

$$\hat{X}_{\mathrm{C}} = \Phi^{-1}\left[1 - \hat{P}(\mathrm{FA})\right], \tag{7.4}$$

where Φ^{-1} is the inverse-Z transformation (i.e., $\Phi^{-1}(p)$ is the Z value that has area to the left equal to p) and $\hat{P}(\mathrm{FA})$ is the observed proportion of false alarms. Note from Figure 7.2 that d' equals the standardized distance from the mean of the N distribution to X_{C} (i.e., X_{C}) plus the distance from X_{C} to the mean of the SN distribution. Therefore,

$$\hat{d}' = \hat{X}_{\mathrm{C}} - \Phi^{-1}\left[1 - \hat{P}(\mathrm{H})\right]. \tag{7.5}$$

Note also that d' is the standardized distance between the means (i.e., the mean difference divided by the common standard deviation). As a result,

the common variance is not identifiable, in the sense that any combination of mean differences and standard deviations that combine to produce the same d' will make identical predictions. As a result, we can set the common standard deviation to 1 without loss of generality.

The two degrees of freedom in the data can be used to estimate X_C and d', but then there are no data left to test the model's goodness-of-fit. Given that the model can perfectly fit any observed values of $P(H)$ and $P(FA)$, then an obvious question is why fit this model to two-stimulus identification data? The most common reason, which has been confirmed in thousands of applications, is that SDT is highly successful at separating perceptual and decisional effects. In particular, manipulations that should only affect sensory magnitude – such as increasing or decreasing signal intensity – mostly cause d' to change but not X_C, whereas manipulations that should only affect the observer's decision about how to act on their sensory experience – such as changing the costs and benefits associated with the various possible outcomes – mostly cause X_C to change but not d'. In contrast, any of these changes are likely to cause accuracy to change, so without SDT, it is generally impossible to know whether a change in accuracy is due to a change in perception or a change in decision strategy. SDT offers a highly effective method for solving this problem.

7.3.1 The receiver operating characteristic

A standard way to summarize the results of a YES-NO detection experiment is via the receiver operating characteristic (ROC), which plots $P(H)$ (on the ordinate) against the probability of a false alarm $P(FA)$ (on the abscissa). The standard approach is to plot data from a variety of conditions that cause X_C to change, but not d'. Examples are shown in Figure 7.3. Because each point on any one curve is associated with a different value of X_C but the same value of d', these are iso-sensitivity contours. Technically, other kinds of curves could be plotted in the same space (e.g., iso-bias curves), but because iso-sensitivity contours are so common, this is almost always what is meant by an ROC curve. For any positive value of d', the iso-sensitivity curve must fall completely in the upper left half of the plot. The main diagonal, in which $P(H) = P(FA)$ (denoted by the dotted line) corresponds to $d' = 0$. Any curve (or point) below this diagonal indicates a negative d'. Since pure guessing should produce $d' = 0$, a (significantly) negative d' should only occur because of participant deception or because the observer is using a highly suboptimal decision rule.

There are several popular experimental designs that are used to estimate

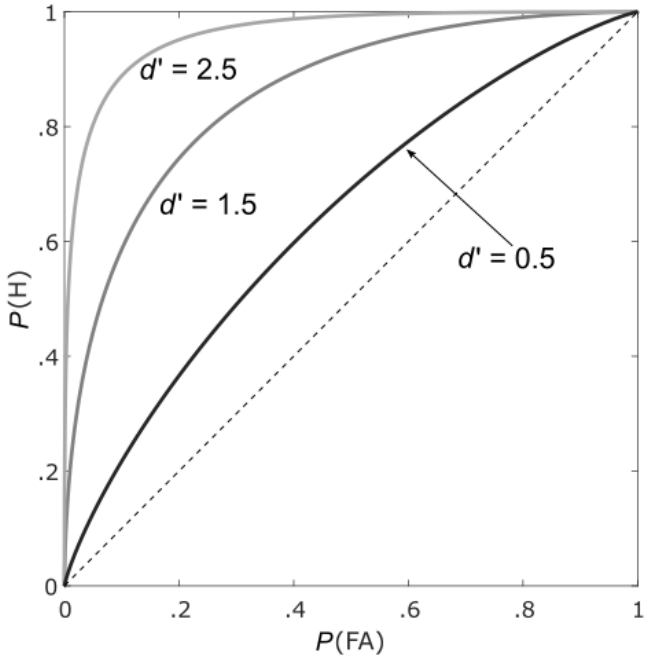

Figure 7.3 An ROC showing iso-sensitivity contours for three different values of d'.

iso-sensitivity curves. One approach is to include a variety of conditions in which the stimulus characteristics remain fixed, but different payoffs are used to encourage participants to change their criterion for responding YES. Another approach, which uses the same N and SN trials but is experimentally more efficient, is to ask observers to rate the intensity of the signal on each trial. Given an r-point rating scale, $r-1$ points on an iso-sensitivity curve can be estimated by assuming that observers construct $r-1$ criteria, denoted by $X_1, X_2, ..., X_{r-1}$, and respond with rating i if and only if $X_{i-1} < \mathbf{X} \leq X_i$, where $X_0 = -\infty$ and $X_r = \infty$. The i^{th} point on the curve is then estimated via

$$\hat{P}(\text{FA}_i) = \hat{P}(R > i|\text{N}) \tag{7.6}$$

and

$$\hat{P}(\text{H}_i) = \hat{P}(R > i|\text{SN}), \tag{7.7}$$

where R is the observer's rating.

The optimal decision strategy in any two-stimulus identification task de-

pends on the likelihood ratio

$$L(x) = \frac{f_{\mathrm{SN}}(x)}{f_{\mathrm{N}}(x)}. \tag{7.8}$$

In particular, if the goal is to maximize the probability of a correct decision, then the optimal decision rule is to

$$\text{Respond YES if } L(\mathbf{X}) > \frac{P(\mathrm{N})}{P(\mathrm{SN})};\ \text{Otherwise respond NO}, \tag{7.9}$$

where $P(\mathrm{N})$ and $P(\mathrm{SN})$ are the probabilities that N and SN, respectively are presented on each trial (i.e., the stimulus base rates). Thus, if SN and N are equally likely, then the optimal strategy is to respond YES if the current sensory magnitude is more likely to be a sample from the SN distribution than from the N distribution. If the sample is more likely from the N distribution, then the NO response should be given. This is the scenario in Figure 7.2. If there are more N trials than SN trials, then the Eq. 7.9 decision rule indicates that stronger evidence is required before responding YES.

In some applications, the different types of errors may incur different penalties and the different types of correct decisions may bring different benefits. Let $V_{I,J}$ denote the value (either positive or negative) of responding J (e.g., YES or NO) on trials when stimulus I was presented (e.g., SN or N). Then the decision rule that maximizes value is (e.g., Green and Swets 1966)

$$\text{Respond YES if } L(\mathbf{X}) > \frac{(V_{\mathrm{N,NO}} + V_{\mathrm{N,YES}})P(\mathrm{N})}{(V_{\mathrm{SN,YES}} + V_{\mathrm{SN,NO}})P(\mathrm{SN})};\ \text{Otherwise respond NO.} \tag{7.10}$$

Note that according to this rule, if the only change in the outcomes is to increase the reward for a correct rejection – that is to increase the (positive) value of $V_{\mathrm{N,NO}}$ – then the observer should increase the criterion, since this will ensure more NO responses. In contrast, if the only change is to increase the penalty for a false alarm – that is to decrease the (negative) value of $V_{\mathrm{N,YES}}$ – then the observer should decrease the criterion, since this will ensure fewer YES responses.

Because of the important role that the likelihood ratio plays in optimal responding, the Eq. 7.1 decision rule is sometimes reformulated in terms of the likelihood ratio:

$$\text{Respond YES if } L(\mathbf{X}) > \beta;\ \text{Otherwise respond NO.} \tag{7.11}$$

In this version of the theory, β can be interpreted as the value of the likeli-

hood ratio at the criterion X_C – that is,

$$\beta = L(X_C) = \frac{f_{SN}(X_C)}{f_N(X_C)}. \tag{7.12}$$

As with X_C, the criterion β is assumed to be under the observer's control. Setting $\beta = P(N)/P(SN)$ maximizes accuracy (i.e., see Eq. 7.9), but the observer is free to set β at some other value. For example, the optimal value of β must be learned, and during this learning process, suboptimal values of β are to be expected.

Note that the Eq. 7.1 and 7.11 decision rules are equivalent if the likelihood ratio increases monotonically with $\mathbf{X}$. The Eq. 7.1 decision rule responds NO to any $\mathbf{X} < X_C$ and YES to any $\mathbf{X} > X_C$, but if the likelihood ratio increases monotonically with $\mathbf{X}$, then the likelihood ratio is less than β for any $\mathbf{X} < X_C$ and greater than β for any $\mathbf{X} > X_C$, so under these conditions, the two decision rules always give the same response. This raises the obvious question of how one could tell from empirical data whether the likelihood ratio of the SN and N sensory distributions is or is not monotonically increasing with sensory magnitude. The key to answering this question is provided by the following result.

Theorem 7.1 *For any differentiable ROC curve, the likelihood ratio*

$$L(x) = \frac{f_{SN}(x)}{f_N(x)} \tag{7.13}$$

is a monotonically increasing function of x (i.e., sensory magnitude) if and only if the ROC is concave down.

Proof By definition, a differentiable function is concave down if and only if its slope is monotonically decreasing. The slope of the ROC curve is

$$\frac{dP(H)}{dP(FA)} = \frac{d[1 - F_{SN}(x)]}{d[1 - F_N(x)]} = \frac{f_{SN}(x)}{f_N(x)} = L(x). \tag{7.14}$$

Therefore, the slope of the ROC curve equals the likelihood ratio, which proves the theorem. □

If the likelihood ratio increases monotonically with sensory magnitude, then the more intense the sensation, the greater the confidence that a signal was presented (i.e., SN). This makes sense, so we would expect empirical ROCs to be concave down, and in fact, the evidence strongly supports this prediction (Green & Swets, 1966). In other words, the empirical evidence supports the assumption that the likelihood ratio of the SN and N sensory

distributions increases monotonically with sensory magnitude. These data rule out many alternative models of the N and SN distributions in which the likelihood ratio is not monotonic. Perhaps the best-known model in this class is the normal, unequal-variance model, which is illustrated in Figure 7.4. The top panel shows a N distribution with small variance and two alternative SN distributions, both with larger variances. The bottom panel shows the ROC curves predicted by this model under the assumption that the observer uses the Eq. 7.1 decision rule.

Figure 7.4 displays several features worth noting. First, the likelihood ratio is not monotonically increasing. Note that, as expected, the SN distribution has higher likelihood for large sensory magnitudes, but nonintuitively, it also has higher likelihood for small magnitudes (i.e., magnitudes below the mean of the N distribution). Therefore, as sensory magnitude increases, the likelihood ratio is initially large (i.e., greater than 1), is then small (less than 1), and finally becomes large again (greater than 1). Because of this non-monotonicity, the Eq. 7.1 decision rule is not optimal. Instead, the optimal strategy (i.e., described by Eq. 7.11) is to respond YES to small and large sensory magnitudes (when $L(\mathbf{X}) > 1$) and NO only for magnitudes of intermediate value (when $L(\mathbf{X}) < 1$).

Second, note that the ROC curves shown in Figure 7.4B are not concave down. Instead, the upper right portion of both curves displays a pronounced violation of concavity. Furthermore, note that both ROCs dip below the main diagonal, which as mentioned earlier, reflects suboptimal decision making. This is because the predicted ROC curves shown in Figure 7.4 were generated under the assumption that the observer is using the Eq. 7.1 decision rule, which is highly suboptimal for small sensory magnitudes.

Third, neither ROC curve in Figure 7.4B is symmetric around the negative diagonal. In fact, many empirical ROCs, albeit concave down, are skewed in this same manner (Green & Swets, 1966), and this is the main reason that the normal, unequal-variance model is popular. In other words, this model accounts for the many reports that empirical ROC curves are skewed, but it is inconsistent with the ubiquitous finding that empirical ROCs are concave down.

Finally, note that the standard measure of sensitivity, namely d', is not defined in this model. Traditionally, d' is defined as the distance between the N and SN means divided by the common standard deviation. In the normal, unequal-variance model, however, there is no common standard deviation, so the traditional d' is undefined. This is also a common problem with multivariate extensions of SDT.

In summary, empirical ROC curves are concave down and are either ap-

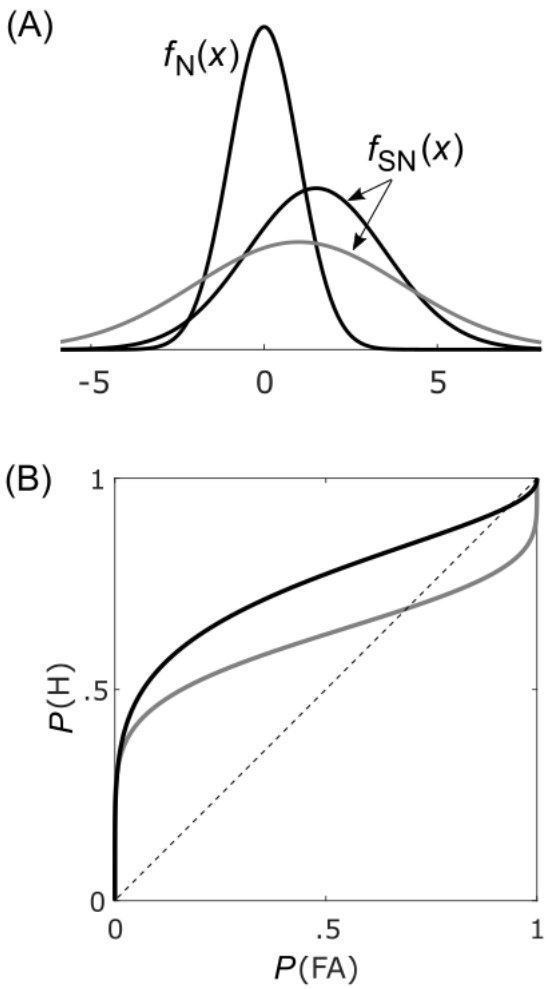

Figure 7.4 (A) The normal, unequal-variance model of SDT. The N distribution is normal with mean 0 and variance 1. Two alternative SN distributions are shown. The pdf in black is normal with mean 1.5 and standard deviation 2, whereas the pdf in gray is normal with mean 1 and standard deviation 3. (B) The ROC showing the iso-sensitivity contours predicted by the two models shown in panel A. Both curves assume the N distribution is normal with mean 0 and variance 1. The black curve assumes the SN distribution has mean 1.5 and standard deviation 2, whereas the gray curve assumes the SN distribution has mean 1 and standard deviation 3.

proximately symmetric about the negative diagonal or skewed in the direction shown in Figure 7.4. The normal, equal-variance model accounts for symmetric ROCs that are concave down, but as it turns out, so do many

other models. Killeen and Taylor (2004) describe the necessary conditions on the N and SN distributions for a SDT model to predict symmetric ROCs.[5] In addition, many SDT models account for skewed ROCs that are concave down. Included in this list, for example, are models in which the N and SN distributions are both exponential or Rayleigh distributions.

7.3.2 Application to other tasks

Although the original applications of SDT in psychology were to YES-NO detection tasks, the theory has also been applied to a variety of other tasks. First, applications to any two-stimulus, identification task are identical except for re-labeling of the stimuli and responses. For example, suppose the stimuli are "A" and "B" and their identifying responses are "a" and "b". If A and B are different stimuli then they must differ in some way. If they differ on some quantitative (i.e., prothetic) dimension, then associate the stimulus with the smaller value with N and its associated response with NO. If they differ on some qualitative (i.e., metathetic) dimension, then the association of A and B to N and SN is arbitrary. Either way, once the associations are complete, the SDT model is identical to the model for the YES-NO detection task.

In addition, SDT has been applied to a variety of different types of experiments that include multiple stimuli. The most widely used is probably the two sample, two-alternative forced-choice task. On each trial, two stimuli are presented – one N and one SN (or one A and one B), and the observer's task is to identify which one is SN (or e.g., B). SDT assumes that exposure to the two stimuli produces two sensory magnitudes – one that is a random sample from the N distribution and one randomly sampled from the SN distribution – and that the observer identifies the larger of these as SN. A well known result, described in the following proposition is that the probability correct in this task equals the area under the ROC that results from the YES-NO detection task (Green, 1964; Green & Swets, 1966).

Theorem 7.2 *SDT predicts that the area under the ROC curve (AUC) equals the probability correct in a two-sample, two-alternative forced-choice task.*

Proof If we let $w = P(\text{FA})$ and define the function g such that $g(w) = P(\text{H})$

[5] Specifically, the ROC is symmetric if the SN cumulative distribution function is generated by applying a strictly decreasing involution to the survivor function of the N distribution (Killeen & Taylor, 2004). An involution is a transformation that is its own inverse. So for example, if T is an involution then $T\{T[1 - F(x)]\} = 1 - F(x)$.

then

$$\begin{aligned} AUC &= \int_0^1 g(w)dw \\ &= \int_0^1 [1 - F_{\mathrm{SN}}(X_{\mathrm{C}})]\, d\,[1 - F_{\mathrm{N}}(X_{\mathrm{C}})] \\ &= \int_{+\infty}^{-\infty} [1 - F_{\mathrm{SN}}(X_{\mathrm{C}})]\, \frac{d\,[1 - F_{\mathrm{N}}(X_{\mathrm{C}})]}{dX_{\mathrm{C}}} dX_{\mathrm{C}} \qquad (7.15) \\ &= \int_{+\infty}^{-\infty} [1 - F_{\mathrm{SN}}(X_{\mathrm{C}})]\, [-f_N(X_{\mathrm{C}})]\, dX_{\mathrm{C}} \\ &= \int_{-\infty}^{+\infty} f_N(X_{\mathrm{C}})\, [1 - F_{\mathrm{SN}}(X_{\mathrm{C}})]\, dX_{\mathrm{C}} \qquad (7.16) \\ &= P(X_{\mathrm{SN}} > X_{\mathrm{N}}). \end{aligned}$$

The limits in Eq. 7.15 are from $+\infty$ to $-\infty$ because $P(\mathrm{FA}) = 0$ when $X_{\mathrm{C}} = +\infty$ and $P(\mathrm{FA}) = 1$ when $X_{\mathrm{C}} = -\infty$. The last equality holds because the integrand in Eq. 7.16 gives the likelihood that the sample from the N distribution equals X_{C} and the sample from the SN distribution is greater than this value. □

AUC is a widely used measure of bias-free classifier performance. For example, compared to d', it has a number of distinct advantages. Perhaps the most important is that AUC is a nonparametric measure that makes no assumptions about the underlying N and SN distributions. In contrast, d' is unambiguously defined only when the N and SN distributions have variances that are equal.

The two-sample, two-alternative forced-choice task is closely related to multiple-look experiments, in which the observer is presented with r independent samples of either N or SN on each trial (e.g., Green and Swets 1966). As in the YES-NO detection task, the observer's task is to respond YES or NO, depending on whether the r samples were all SNs or Ns. Another well-known result relates the performance of an ideal observer in the multiple-look experiment to performance in the YES-NO detection task.

Theorem 7.3 *Suppose an ideal observer with perfect memory participates in a multiple-look experiment in which r independent samples of N or SN are presented on each trial. Denote the d' of this observer in the YES-NO detection task as d'_{YN} and the d' in the multiple-look experiment as d'_r. Then*

the normal, equal-variance model predicts that

$$d'_r = \sqrt{r}\, d'_{YN}. \tag{7.17}$$

Proof In the multiple-look experiment, each of the r N or SN samples generates its own sensory value. Denote the i^{th} of these by x_i, and the collection of all r by the vector $\underline{x}' = [x_1, x_2, ..., x_r]$. Under the assumptions of the proposition, note that on N trials, $\underline{x}$ has an r-dimensional multivariate Z distribution, and on SN trials it has an r-dimensional multivariate normal distribution with mean vector $\underline{\mu}' = [d'_{\text{YN}}, d'_{\text{YN}}, ..., d'_{\text{YN}}]$ and variance-covariance matrix equal to the identity. Since the variance equals 1 in all directions, the standardized distance between the N and SN means equals

$$\begin{aligned} d'_r &= \sqrt{(d'_{\text{YN}} - 0)^2 + (d'_{\text{YN}} - 0)^2 + ... + (d'_{\text{YN}} - 0)^2} \\ &= \sqrt{r\, d'^2_{\text{YN}}} \\ &= \sqrt{r}\, d'_{\text{YN}}. \end{aligned}$$

□

Estimation of d'_r for human observers shows that it increases with r, but more slowly than predicted by Eq. 7.17 (Green & Swets, 1966). The most likely reason is that human observers do not have perfect memory, and thus are unable to take full advantage of all r stimulus samples.

7.3.3 Extensions

Marr (1982) famously proposed the hierarchical classification of mathematical models as computational, algorithmic, or implementational. In mathematical psychology, Marr's algorithmic-level models are often referred to as process models. SDT provides a computational-level description of decision making, since it makes no attempt to describe the underlying algorithms or perceptual or cognitive processes that mediate decision making. During the 1970's, great efforts were devoted to developing process models of decision making, and currently there are several different process interpretations of SDT. Perhaps the most popular is provided by the drift-diffusion model (Link & Heath, 1975; Ratcliff, 1978), which is illustrated in Figure 7.5. The idea is that instead of representing the sensory effects of the stimulus on each trial with a single random sample from the N or SN distributions, as in classical SDT, the observer is assumed to repeatedly sample the presented

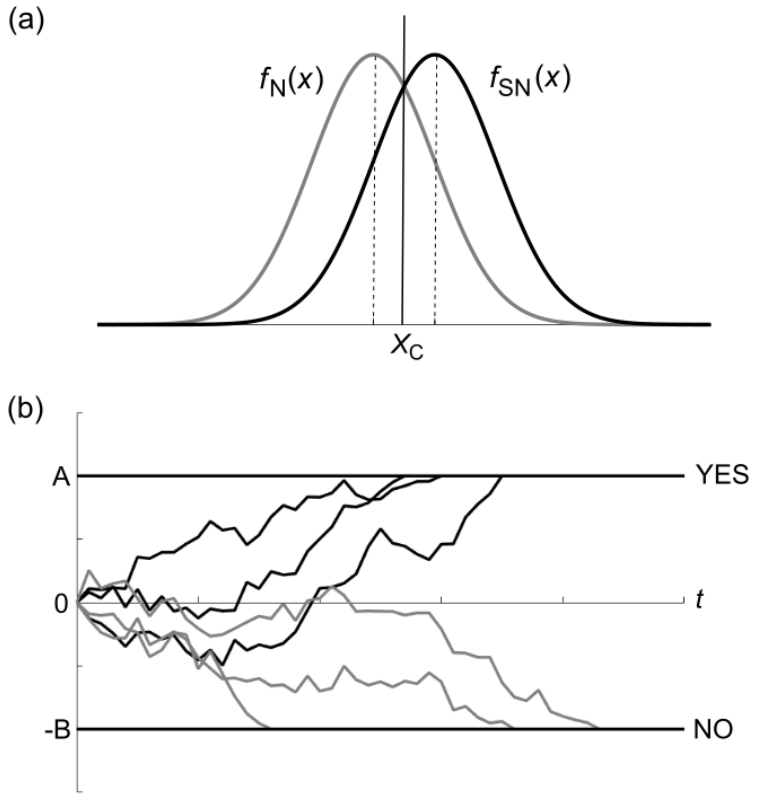

Figure 7.5 (a) The normal, equal-variance model in which $d' = 1$. (b) A drift-diffusion model in which the drift is determined by random sampling from the N or SN distribution. Samples larger than X_C push the drift up, whereas samples smaller than X_C push it down. Sample paths are shown for six hypothetical trials – three SN trials (in black) and three N trials (in gray).

stimulus as long as it is available. Each sample $\mathbf{X}$ is compared to the criterion X_C by computing the difference $\mathbf{X} - X_C$, and these differences are accumulated. The sampling and accumulating processes continue until the resulting sum (or integral) first exceeds an upper criterion A or falls below a lower criterion -B (i.e., see Figure 7.5b). Sampling terminates with a YES response in the former case, and with a NO response in the latter case.

This version of the drift-diffusion model includes the d' and X_C parameters of SDT plus the response criteria A and B. However, in addition to predicting accuracy data, the diffusion model also predicts RTs because closed-form expressions exist for first passage times (i.e., time when the process first

crosses a response threshold). As a result, there are more data to fit, and therefore more degrees of freedom available for parameter estimation. Several computer packages are available that automate this parameter estimation process (Vandekerckhove & Tuerlinckx, 2007; Wiecki, Sofer, & Frank, 2013).

Note that the drift-diffusion model can represent a response bias in two different ways. One is to place X_C at some point where the likelihood ratio is different from 1 (assuming equal base rates and payoffs), and another is to set A $\neq$ B. Of course, the classical SDT model can account for bias only by adjusting X_C. Consider a condition in which the observer adopts a conservative criterion and therefore is biased towards responding NO. Thus, according to SDT, X_C is set at some point where the likelihood ratio is greater than 1 (i.e, $\beta > 1$). Now consider trials in that condition where the sensory value falls at some point where the likelihood ratio is greater than 1 but less than β. According to SDT, the observer will respond NO on this trial, even though the evidence objectively favors a YES response (because the likelihood ratio is greater than 1). Balakrishnan (1999) presented evidence against this prediction. In particular, he described results of several experiments that suggested that observers always respond with the alternative that is most likely to be correct, even if they are biased towards one response and against the other. Unfortunately, there is no way to represent this state of affairs in classical SDT. In contrast, the drift-diffusion model offers an elegant resolution to this apparent paradox. Balakrishnan's results suggest that X_C is set at the point for which $\beta = 1$ in all applications (e.g., as in Figure 7.5). A bias towards a NO response can then be implemented by setting A > B. Thus, according to this account, the evidence is always judged objectively. Evidence that objectively favors SN always makes a YES response more likely and evidence that favors N always makes a NO response more likely. Therefore, a bias towards responding NO does not color the observer's view of the world. Instead, the observer is simply willing to stop and respond NO on the basis of less overall evidence then they are willing to stop and respond YES. This more reasonable view of response bias is among the greatest advantages that the drift-diffusion model provides over and above classical SDT.

7.4 Two or More Dimensions: General Recognition Theory

SDT is useful for understanding behavior in any task in which the observer's decision is based on a single sensory dimension. Most real-world stimuli vary on multiple dimensions, however, and many perceptual decisions require attention to more than one dimension. For example, there is no single sensory

dimension that allows accurate face identification. For this reason, there is obvious value in extending SDT to multiple stimulus dimensions.

At first glance, this seems like a straightforward exercise. An obvious place to begin is by replacing the unidimensional probability distributions that are used to represent the sensory effects of a stimulus in SDT with multivariate probability distributions. But complications quickly arise even in the case of two sensory dimensions. First, some sensory dimensions interact, and the perceptual literature includes a bewildering number of terms that have been proposed to describe these interactions, including perceptual independence, separability, integrality, holism, configurality, sampling independence, dimensional orthogonality, and performance parity. How should these different types of sensory interactions be modeled? And how are they all related to each other? Second, how should the decision process be modeled? In SDT, the sensory space is a line, and in two-alternative tasks, the observer is typically assumed to divide the line into two regions – one associated with each response alternative. Fortunately, there are only a few ways to do this. In fact, a standard lecture in courses on SDT is to show that almost any decision strategy is equivalent to the Eq. 7.1 decision rule. However, if there are two sensory dimensions, then the sensory space is a plane, and there are an infinite number of qualitatively different ways to divide a plane into two regions.

Not surprisingly, the first attempt to generalize SDT to multiple stimulus dimensions, by Tanner in 1956, ignored most of these issues. Specifically, Tanner (1956) allowed for only one simple type of perceptual interaction and he assumed that observers always use an optimal decision rule. Despite these simplifying assumptions, Tanner's contribution was significant because he was the first to consider multiple sensory dimensions. Even so, it was another 30 years before a more useful multidimensional version of SDT was developed. During the late 1980s, a flurry of articles significantly generalized Tanner's approach. The title of Tanner's (1956) article was "Theory of recognition." To pay homage to his contributions, Ashby and Townsend (1986) called their more general approach, general recognition theory (GRT). GRT quickly developed: Ashby and Townsend (1986) proposed a GRT-based theory of perceptual interactions, Ashby and Gott (1988) studied decision rules in multidimensional perceptual spaces, and Ashby and Perrin (1988) used GRT to develop a unified theory of similarity and identification.

7.4.1 Identification versus categorization

GRT has been applied to a wide variety of tasks. But two tasks – identification and categorization – have emerged as the most popular, and which one is used depends on the goals of the research. In particular, identification tasks are used if the primary goal is to study perceptual representations, whereas categorization tasks are used if the primary goal is to study decision processes.

In identification tasks, there are M stimuli and M unique identifying responses. On each trial, one of the stimuli is presented, and the observer's task is to identify the stimulus by emitting the appropriate response. The data are collected in an $M \times M$ confusion matrix, in which the entry in row i and column j is the frequency with which the observer gave response j on trials when stimulus i was presented. Because the number of stimulus presentations is known, there is one constraint on each row of the confusion matrix. As a result, every confusion matrix has $M \times (M - 1)$ degrees of freedom. Note that the YES-NO detection task is a special case of this identification task in which $M = 2$ and the two stimuli to be identified are N and SN.

The most useful information in identification tasks is in the confusions that observers make, so experimental conditions are selected to guarantee errors. This is usually accomplished by using highly similar stimuli, but sometimes brief exposure durations or noise masks are used instead. Anytime one stimulus is confused for another, an error occurs. Therefore, misidentifications are most commonly made because of errors in perception, rather than because of a suboptimal decision strategy. As a result, identification tasks are a good choice if the goal is to study perceptual representations. Of course, observers can also make errors if they fail to remember which response button is associated with which stimulus. Therefore, feedback is usually provided to help observers learn these associations, and some training trials are included that are excluded from the data analysis.

In the most widely used identification tasks, the stimuli are constructed by factorially combining a small number of discrete values on two sensory dimensions. The most common choice is to factorially combine two values on two dimensions to create a total of four stimuli. Each confusion matrix collected from such a 2×2 factorial design includes 12 degrees of freedom (4×3) for parameter estimation and model testing. If we call the two stimulus dimensions A and B, then we can denote the stimulus in which dimension or component A is at level i and component B is at level j by $\mathrm{A}_i\mathrm{B}_j$, and the corresponding response by $\mathrm{a}_i\mathrm{b}_j$.

Categorization experiments are identical to identification experiments, except they include fewer response alternatives than stimuli. In a categorization experiment, one of N stimuli is presented on each trial and the observer's task is to assign it to one of M categories, where $M < N$. The confusion matrix is therefore $N \times M$, and it contains $N \times (M-1)$ degrees of freedom. The most common choice is $M = 2$. Note that in this case, the data include N degrees of freedom. In most cases the categories are novel, in the sense that they were created specifically to use in the experiment. As a result, accurate responding requires the observer to learn the structure of these categories, most commonly via trial-by-trial feedback provided by the experimenter. Errors are most likely to occur because the observer is using a suboptimal strategy to assign stimuli to categories. Misperceptions are just as likely as in identification experiments, but they tend to have little effect on accuracy. For example, confusing one stimulus with another in the same category does not change the response, and therefore has no observable effect on behavior. For these reasons, categorization experiments are a good choice if the goal is to study decision processes.

7.4.2 Modeling perceptual and decisional interactions

One of the foundational motivations for the generalization of SDT to multiple dimensions was to model perceptual interactions in a theoretically rigorous way (Ashby & Townsend, 1986). For much of the middle portion of the twentieth century, this issue was addressed almost completely in terms of operational definitions (e.g., Garner and Felfoldy 1970; Garner, Hake, and Eriksen 1956; Garner and Morton 1969).[6] Ashby and Townsend (1986) created GRT principally as a theoretical structure to define perceptual independence, perceptual separability, and decisional separability. These definitions are now standard in the field. They also showed how these theoretical primitives relate to a variety of other independence-related terms that were popular in the literature.

A GRT model of the 2×2 factorial identification experiment is shown in Figure 7.6. The ellipses denote the contours of equal likelihood for the four bivariate perceptual distributions, where $f_{ij}(x_1, x_2)$ denotes the perceptual distribution associated with stimulus $\mathrm{A}_i\mathrm{B}_j$. Note that the marginal distributions associated with this stimulus are denoted by $g_{ij}(x_1)$ and $g_{ij}(x_2)$ for dimensions x_1 and x_2, respectively. Also shown are the decision bounds that divide the perceptual plane into four response regions.

[6] Use of the term operational is not to be confused here with the logic of operationism or converging operations (Bridgman, 1945; Von Der Heide, Wenger, Bittner, & Fitousi, 2018).

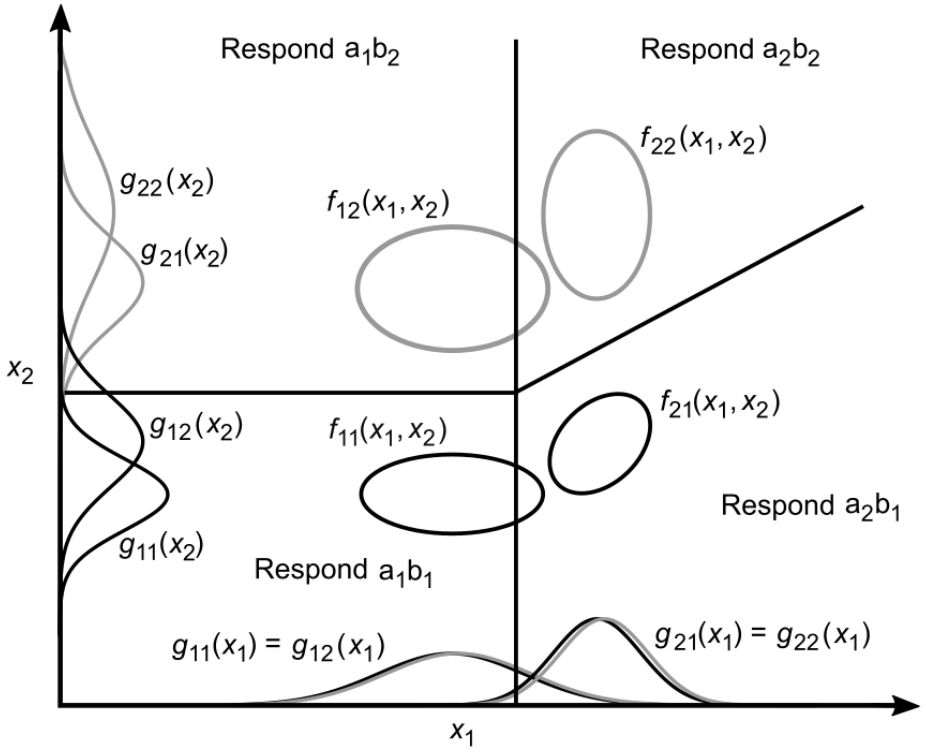

Figure 7.6 A GRT model of the 2×2 factorial identification experiment. The ellipses denote the contours of equal likelihood for the four bivariate perceptual distributions.

According to GRT, stimulus components A and B satisfy *perceptual independence* in stimulus A_iB_j if and only if the perceived value of component A is statistically independent of the perceived value of component B on trials when stimulus A_iB_j is presented. More specifically, perceptual independence of components A and B holds in stimulus A_iB_j if and only if

$$f_{ij}(x_1, x_2) = g_{ij}(x_1)g_{ij}(x_2), \tag{7.18}$$

for all values of x_1 and x_2. If perceptual independence is violated, then components A and B are perceived dependently.

Note that perceptual independence is a property of a single stimulus, in the sense, for example, that perceptual independence could hold for one stimulus and be violated for all others. In the Figure 7.6 example, the distributions are all bivariate normal, so independence is equivalent to zero correlation. Note that perceptual independence appears to be satisfied in all stimuli except

A_2B_1, which displays a positive correlation between perceived values of the A and B stimulus components.

Component A is *perceptually separable* from component B if the observer's perception of A does not change when the level of B is varied. In other words, if components A and B are perceptually separable, then it is easy to attend to one and ignore the other. If this is impossible – that is, if the perception of A changes when B changes, then component A is *perceptually integral* with component B. Classic separable dimensions are color and shape, whereas classic integral dimensions are the saturation and brightness of a color patch. In GRT, all information about the perception of component A on trials when stimulus A_iB_j is presented is contained in the marginal distribution $g_{ij}(x_1)$. Therefore, component A is perceptually separable from B if and only if

$$g_{11}(x_1) = g_{12}(x_1), \text{ and } g_{21}(x_1) = g_{22}(x_1), \text{ for all values of } x_1. \tag{7.19}$$

Equation 7.19 guarantees that the perception of component A_1 is the same regardless of whether it appears with B_1 or B_2, and that the same invariance holds for component A_2. In the Figure 7.6 example, note that component A is perceptually separable from component B, but component B is not perceptually separable from A. In particular, changing the level of B does not change the perception of A, but increasing the level of A from A_1 to A_2 increases the perceived value of component B. Note that unlike perceptual independence, perceptual separability is a property of multiple stimuli (i.e., all that share a common value on one stimulus dimension).

Finally, *decisional seprability* holds on dimension x_1 if the decision about whether component A is at level 1 or level 2 does not depend on the perceived value of component B. Mathematically, this condition holds if and only if the observer uses the following decision rule to determine the level of component A:

$$\text{The level of component A is 1 if } \mathbf{X}_1 \leq X_1; \text{ Otherwise the level is 2,} \tag{7.20}$$

for some constant criterion X_1. This decision rule is equivalent to using a decision bound on dimension x_1 that is parallel to the x_2 axis (and therefore orthogonal to the x_1 axis). In the Figure 7.6 example, note that decisional separability holds on dimension x_1, but not on dimension x_2.

GRT has also been used successfully to formalize and study the notion of holistic or configural perception or processing (e.g., see discussions in Piepers and Robbins 2012; Richler and Gauthier 2014). GRT was first used to model the potential perceptual and decisional interactions that constitute holistic or configural perception by O'Toole, Wenger, and Townsend (2001),

and it was first applied to the holistic or configural perception of faces by Wenger and Ingvalson (2002, 2003). More recently, Townsend and Wenger (2015) used GRT to propose a set of working axioms for holistic or configural perception.

As an example of how GRT has been used to study holistic processing, consider face perception, and more specifically, the composite face effect (Young, Hellawell, & Hay, 1987), which is frequently cited as a hallmark of holistic perception (Murphy, Gray, & Cook, 2017). The composite face illusion occurs in tasks where observers are presented with an image of a face, divided into top and bottom portions roughly at the nose. Observers are asked to identify either the top or bottom half while ignoring the other half. The top and bottom portions can be drawn from either the same or different faces, the faces can be either familiar (e.g., famous) or unfamiliar, and the two halves can be either aligned or misaligned. The composite face effect is that identification of one half is impaired when the top and bottom halves are from different faces, and this impairment is greatest when the two halves are from different familiar identities.

The first step in modeling the composite face effect with GRT is to represent the space of perceptual evidence supporting identification of the two halves. For simplicity, consider the simplest case in which the top and bottom halves are always aligned. Let component A denote the top half face and component B denote the bottom half, with the subscript denoting the identity of the face. So in stimuli A_1B_1 and A_2B_2, the top and bottom halves are from the same face, whereas in stimuli A_1B_2 and A_2B_1, the two halves are from different faces.

The next step is to construct a null model that does not display any type of holism or configurality. We do this by assuming perceptual independence for all stimuli, perceptual and decisional separability on both dimensions, and that all variances are equal. In this model, which is illustrated in Figure 7.7a, the identity of the top half of the stimulus does not affect the perceptual representation or the decision made about the bottom half.

The final step is to build a model that assumes holistic perception when the top and bottom halves are from the same face, but not when they mismatch. There are several ways to do this. One is to assume a positive perceptual dependence when the two halves match and a negative dependence when they mismatch. This model, which is illustrated in Figure 7.7b, corresponds to the type of within-stimulus relationships that are implied in the vernacular use of holism, configurality, or Gestalt (O'Toole et al., 2001; Townsend & Wenger, 2015). A second way is to change the marginal means of the distributions such that confusability increases when the bottom and

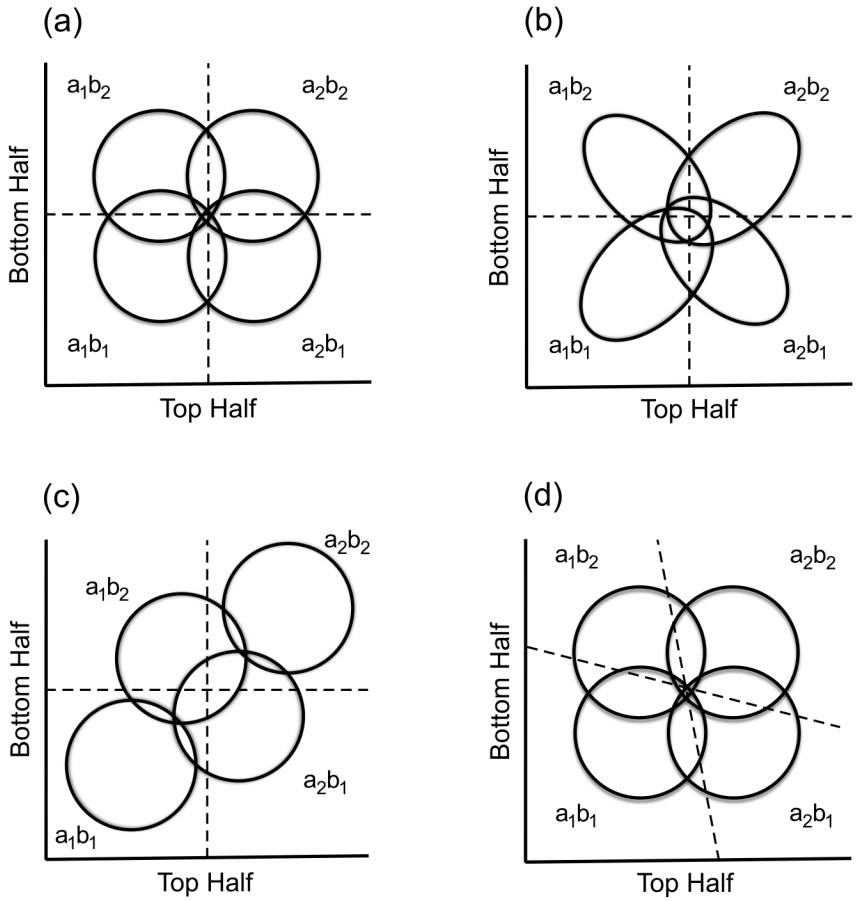

Figure 7.7 Alternative GRT models of the holism or configurality thought to underlie the composite face effect: (a) lack of holism or configurality, (b) positive perceptual dependencies when the halves match and negative perceptual dependencies when they mismatch, (c) shifting the perceptual means to model increased accuracy when the halves match and decreased accuracy when they mismatch, (d) accounting for increased accuracy when the halves match by shifting the decision bounds.

top are mismatched and decreases when they are matched (Figure 7.7c). The same effect could be obtained by the third possible way of modeling holism: by shifting the decision bounds (Figure 7.7d). Of course, these possibilities could also be combined in a variety of ways.

Two significant points have been made by applying GRT to the issue of holism. The first is that, just as there are varieties of independence in per-

ception (Ashby & Townsend, 1986), there are a variety of ways to obtain patterns of data from which one can infer holism or configurality. The second is that analysis of a task by way of GRT can provide important insights into the extent to which the task is capable of testing a hypothesis. For example, GRT simulations reported by Richler, Gauthier, Wenger, and Palmeri (2008) demonstrated that the standard method of testing the composite face effect (see discussions in Richler and Gauthier 2013; Rossion 2013) does not provide data that would allow for testing the strong hypothesis that holism is a within-stimulus effect.

7.4.3 Applications to categorization tasks

In principle, the application of GRT to categorization tasks is the same as its application to identification. In both cases, the data are summarized in a confusion matrix, and the primary focus is on the pattern of errors made by observers. One important statistical difference however, is that, for the same number of stimuli, categorization data have fewer degrees of freedom – often far fewer. For example, the most common categorization experiments include two categories. Therefore, with M stimuli, an identification confusion matrix includes $M \times (M - 1)$ degrees of freedom and the corresponding categorization confusion matrix includes only M degrees of freedom (i.e., since the confusion matrix has order $M \times 2$). Because the data include fewer degrees of freedom, GRT applications to categorization tasks include simplifying assumptions that reduce the number of free parameters, relative to GRT applications to identification data.

In fact, when applied to categorization data, the most common assumption is that all perceptual representations are multivariate normally distributed with known means and with variance-covariance matrix equal to σ^2I, where σ^2 is the common noise variance on each dimension and I is the identity matrix. Thus, only one free parameter is typically assigned to model all perceptual representations (i.e., σ^2), and all other parameters are used to model decision bounds. This choice reflects the assumption that in categorization experiments, errors are more likely caused by suboptimal decision strategies than by faulty perception. Allocating the lion's share of parameters to the decision bounds provides the best opportunity to characterize these suboptimalities.

The mean of each perceptual distribution describes the mean perceived value of each stimulus. In some cases, these could come from previous multidimensional scaling or psychophysical modeling of the stimuli. For example, in the case of sine-wave gratings (such as Gabor patches) that vary in spa-

tial frequency and orientation, a psychophysical model that describes the transformation from stimulus space to perceptual space was provided by Treutwein, Rentschler, and Caelli (1989). Another possibility, especially for dimensions that are perceptually separable, is to use Stevens' exponent. For example, the Stevens exponent for brightness is 0.33, so the mean brightness of each stimulus could be computed from $kI^{0.33}$, where I is the physical intensity of the stimulus and k is an arbitrary constant that can be set for convenience. When GRT models are fit to categorization data under these assumptions about the perceptual representations, they are often referred to as decision bound models. One advantage they have over GRT models with more complex perceptual representations, which is illustrated in the next result (due to Ashby and Maddox 1993), is that no numerical integration is needed to fit any of the most common models.

Theorem 7.4 *Consider a categorization task with two categories, A and B, and a decision bound model with one linear boundary. Let the random vector* $\underline{\boldsymbol{X}}_i$ *denote the perceived value of stimulus* S_i*. Assume that* $\underline{\boldsymbol{X}}_i$ *has a multivariate normal distribution with known mean* $\underline{\mu}_i$ *and variance-covariance matrix* $\sigma^2 I$*. Then the decision bound is the set of all points that satisfy*

$$\underline{b}'\underline{\mu}_i + c = 0, \tag{7.21}$$

for some vector of constants $\underline{b}$ *and constant* c*. This model, called the general linear classifier, predicts that*

$$P(A|S_i) = \Phi\left(\frac{\underline{b}'\underline{\mu}_i + c}{\sigma\sqrt{\underline{b}'\underline{b}}}\right), \tag{7.22}$$

where Φ *is the cumulative distribution function of a standard normal (i.e., Z) distribution.*

Proof Under the conditions of the proposition, the decision rule of the general linear classifier is "Respond A if $h(\underline{\mathbf{X}}_i) > 0$; otherwise respond B." Therefore,

$$P(A|\mathrm{S}_i) = P\left[h(\underline{\mathbf{X}}_i) > 0|\mathrm{S}_i\right]. \tag{7.23}$$

Now $\underline{\mathbf{X}}_i$ has a multivariate normal distribution with mean vector $\underline{\mu}_i$ and variance-covariance matrix $\sigma^2\mathrm{I}$. As a result, $h(\underline{\mathbf{X}}_i)$ has a univariate normal distribution with mean $\underline{\mathrm{b}}'\underline{\mu}_i + c$ and variance $\sigma^2\underline{\mathrm{b}}'\underline{\mathrm{b}}$. The result follows immediately from these observations. □

Since the $\underline{\mu}_i$ are assumed to be known, the parameters of the model are

the noise variance σ^2 and the decision bound parameters $\underline{b}$ and c. If there are r perceptual dimensions, then $\underline{b}$ has order $r \times 1$. However, without loss of generality, one entry in $\underline{b}$ can be set arbitrarily, so $\underline{b}$ has only $r - 1$ free parameters.[7] Therefore, if the perceptual space is two dimensional, this model has three free parameters (i.e., one slope parameter, the decision bound intercept c, and the noise variance σ^2).

Predictions for the decision bound model that assumes a quadratic decision bound, called the general quadratic classifier, were derived by Ashby and Maddox (1993). Predictions for models that assume some form of decisional separability can be found in Ashby and Valentin (2018). For these models, the decision bound is compatible with an explicit rule that is easily verbalized. For example, the rule: "Respond A if $\mathbf{X}_1 > c_1$ and $\mathbf{X}_2 > c_2$; otherwise respond B" is equivalent to the conjunction rule "Respond A if the stimulus is large on both dimensions; otherwise respond B." Ashby and Valentin (2018) also described predictions of models that assume the participant guesses randomly on every trial.

Criterial noise can be added to decision bound models by assuming that the decision rule is "Respond A if $h(\underline{\mathbf{X}}) > \boldsymbol{\epsilon}_c$; otherwise respond B," where $\boldsymbol{\epsilon}_c$ is normally distributed with mean 0 and variances σ_c^2. If the decision bound is linear, then it is straightforward to show that perceptual and criterial noise are not separately identifiable (Ashby & Maddox, 1993). Instead, only the sum of the perceptual and criterial noise variances is estimable. For this reason, it makes no difference whether we assume that the noise is perceptual or decisional (or some combination of the two). Once predicted probabilities are computed, the parameters can be estimated by finding the numerical values that maximize the likilihood-related statistic L^* in Eq. 7.33 below.

7.4.4 Applications to identification tasks

GRT has been used to analyze data from identification confusion matrices in two different ways. One approach is to compute certain summary statistics from the empirical confusion matrix and then to check whether these satisfy conditions that are characteristic of perceptual independence, perceptual separability, or decisional separability. The other approach is to fit GRT models to the entire confusion matrix. To test various assumptions

[7] For example, assume $r = 2$. Then note that at least one of b_1 and b_2 (i.e., the entries in $\underline{b}$) must be nonzero. Note that the decision rule "Respond A if $h(\underline{x}) > 0$" is unchanged if we divide both sides by a positive constant. Therefore, without loss of generality, we can divide both sides by $\sqrt{b_1^2 + b_2^2}$. Note that the sum of the squared entries in the revised $\underline{b}$ vector now equals 1. As a result, we can always replace b_2 with $\sqrt{1 - b_1^2}$.

about perceptual and decisional processing – for example, to test whether perceptual independence holds – a version of the model that assumes perceptual independence is fit to the data as well as a version that makes no assumptions about independence. This latter version contains the former version as a special case (i.e., in which all covariance parameters are set to zero), so it can never fit worse. After fitting these two models, we conclude that perceptual independence is violated if the more general model fits significantly better than the more restricted model that assumes perceptual independence (Ashby & Perrin, 1988; Thomas, 2001).[8] Because these approaches are so different, we discuss each in turn.

It is important to note however, that regardless of which method is used, there are certain nonidentifiabilities in GRT models that could limit the conclusions that are possible to draw from any such analyses (e.g., Menneer, Wenger, and Blaha 2010; Silbert and Thomas 2013). The problems are most severe when GRT is applied to identification data from 2×2 factorial designs (i.e., with stimuli A_1B_1, A_1B_2, A_2B_1, and A_2B_2). For example, Silbert and Thomas (2013) showed that in 2×2 applications where there are two intersecting linear decision bounds that do not satisfy decisional separability, there always exists an alternative model that makes the exact same empirical predictions and satisfies decisional separability (and these two models are related by a linear transformation). Thus, in standard applications of GRT to identification experiments that use a 2×2 factorial design, decisional separability is not testable, nor are the slopes of the decision bounds uniquely estimable. It turns out however, that for a variety of reasons, these nonidentifiabilities are not catastrophic.

First, there are no identifiability problems if the perceptual dimensions are known. Obviously, the linear transformation that rotates intersecting linear bounds so that one is vertical and one is horizontal also rotates the perceptual dimensions. So although decisional separability holds in the new model, the separability is with respect to novel dimensions. In other words, one interpretation of the identifiability problem is that if the best-fitting GRT model to some single confusion matrix collected in an experiment that used a 2×2 factorial design assumes intersecting linear bounds that violate decisional separability, then there is always an alternative GRT model that fits equally well and assumes that the observer made decisions by selectively attending to some different perceptual dimensions. With complex stimuli, such as faces, this will often be difficult to rule out. However, with many

[8] Note that many of the statistical packages written for estimating GRT models provide estimates of parameter variability and/or confidence intervals, allowing one to determine whether (for example) a parameter estimate can be inferred to be reliably different from 0.

simple stimuli, this possibility is straightforward to reject. For example, consider sine-wave gratings (e.g., such as Gabor patches) that are created by factorially combining two spatial frequencies (bar widths) and two (bar) orientations. An enormous visual perception literature tells us that humans treat these two dimensions as primary (e.g., DeValois and De Valois 1990). So any conclusions about decisional separability drawn from a GRT analysis should be immune to identifiability problems because the mathematically equivalent model that makes different assumptions about decisional separability must assume that the observer perceived the stimuli in a way that is incompatible with the visual perception literature.

Second, the problems do not generally exist with 3×3 or larger factorial designs (as used for example, by Ashby, Waldron, Lee, and Berkman 2001). In the 3×3 case, the GRT model with linear bounds requires at least 4 decision bounds to divide the perceptual space into nine response regions (e.g., in a tic-tac-toe configuration). Typically, two will have a generally vertical orientation in the two-dimensional perceptual space and two will have a generally horizontal orientation. Linear transformations will rotate the vertical-tending bounds by the same amount, and the horizontal-tending bounds by the same amount. Therefore, unless the two vertical-tending bounds are parallel and the two horizontal-tending bounds are parallel, there is no linear transformation that guarantees decisional separability for all 4 bounds. For example, if the two vertical-tending bounds are not parallel, then the linear transformation that makes one perfectly vertical (guaranteeing decisional separability) will leave the other oblique to the abscissa (causing a violation of decisional separability). Thus, in 3×3 (or higher) designs, decisional separability is typically identifiable and testable.

Third, there are simple experimental manipulations that can be added to the basic 2×2 identification experiment to test for decisional separability. Currently, more than 30 different qualitative differences have been identified in the learning and performance of tasks in which observers use strategies that satisfy versus violate decisional separability (for a review of most of these, see Ashby and Valentin 2017). For example, switching the locations of the response buttons interferes with performance if decisional separability fails more than if decisional separability holds (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004), and delaying feedback by a few seconds has a similar effect, but on learning, rather than performance (Crossley & Ashby, 2015; Dunn, Newell, & Kalish, 2012; Maddox, Ashby, & Bohil, 2003; Maddox & David, 2005).

Fourth, one could analyze the 2×2 data using GRT-wIND (GRT with INDividual differences; Soto, Vucovich, Musgrave, and Ashby 2015), which

was inspired by the INDSCAL model of multidimensional scaling (Carroll & Chang, 1970). Like INDSCAL, GRT-wIND is fit to the data from all individuals simultaneously. All observers are assumed to share the same group perceptual distributions (see Silbert and Thomas 2017 for discussion of this assumption), but different observers are allowed different linear bounds and they are assumed to allocate different amounts of attention to each perceptual dimension. The model does not suffer from the identifiability problems identified by Silbert and Thomas (2013), even in the 2×2 case, because with different linear bounds for each observer, there is no linear transformation that simultaneously makes all these bounds satisfy decisional separability.

Summary statistics approach

The first approach that used GRT to test perceptual and decisional assumptions was based on parametric and non-parametric summary statistics that were derived from the identification-confusion matrix (see, e.g., Figure 11, p. 172, Ashby and Townsend 1986). This later evolved to an approach known as multidimensional signal detection analysis (MSDA, Kadlec 1995; Kadlec and Townsend 1992a, 1992b), which extended the concepts originally presented by Ashby and Townsend (1986) and combined those equalities with tests of equalities on Gaussian SDT parameters. This was later both simplified and refined, as summarized by Silbert and Hawkins (2016), under the strong assumption that decisional separability always holds (see also Silbert and Thomas 2013).

The most popular summary statistics tests use measures called *marginal response invariance* and *report independence* to draw inferences about perceptual separability and perceptual independence. Marginal response invariance holds at the ith level of the first dimension if the following equality holds:

$$\begin{aligned} P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_1) &= P(\mathrm{a}_i\mathrm{b}_1|\mathrm{A}_i\mathrm{B}_1) + P(\mathrm{a}_i\mathrm{b}_2|\mathrm{A}_i\mathrm{B}_1) \\ &= P(\mathrm{a}_i\mathrm{b}_1|\mathrm{A}_i\mathrm{B}_2) + P(\mathrm{a}_i\mathrm{b}_2|\mathrm{A}_i\mathrm{B}_2) \\ &= P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_2), \end{aligned} \tag{7.24}$$

where, as before, $P(\mathrm{a}_k\mathrm{b}_m|\mathrm{A}_i\mathrm{B}_j)$ is the probability that the participant responded $\mathrm{a}_k\mathrm{b}_m$ on trials when stimulus $\mathrm{A}_i\mathrm{B}_j$ was presented. Marginal response invariance provides information about perceptual separability so long as decisional separability holds. If decisional separability does hold, then a failure of marginal response invariance at any level of a given dimension implies that perceptual separability fails on that dimension (Ashby &

Townsend, 1986). If the marginal d's are also unequal on that dimension, then our conclusion that perceptual separability fails is further bolstered.

Before GRT, the most popular method for assessing separability was via a categorization task called the *filtering task*, which uses the same stimuli as the 2×2 identification task, but asks observers to report the level of component A or the level of component B, rather than identify the stimulus uniquely. Ashby and Maddox (1994) proposed an RT version of marginal response invariance for this task that they called *marginal RT invariance*. Specifically, for $i = 1$ or 2, marginal RT invariance holds for component A if

$$P(\mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1, \mathrm{a}_i) = P(\mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2, \mathrm{a}_i), \text{ for all } t > 0, \tag{7.25}$$

where **RT** is the RT and a_i indicates that the observer responded that the level of component A was i. Ashby and Maddox (1994) showed that if decisional separability holds and if RT decreases with the distance from the percept to the decision bound – an assumption called the RT-distance hypothesis – then perceptual separability holds if and only if marginal RT invariance is satisfied for both correct and incorrect responses.

Ashby and Maddox (1994) only investigated tasks with two response alternatives (i.e., the filtering and redundancy tasks popularized by Garner 1974). Townsend, Houpt, and Silbert (2012) applied a similar approach to the 2×2 identification task. They defined an RT invariance condition similar to marginal RT invariance that they called timed marginal response invariance. This condition holds in the 2×2 identification task for level i of component A if, for all $t > 0$

$$\begin{aligned} P(\mathrm{a}_i\mathrm{b}_1, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1) + P(\mathrm{a}_i\mathrm{b}_2, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1) \\ = P(\mathrm{a}_i\mathrm{b}_1, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2) + P(\mathrm{a}_i\mathrm{b}_2, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2). \end{aligned} \tag{7.26}$$

Rather than assume the RT-distance hypothesis, Townsend et al. (2012) investigated predictions of a general class of models that assumed processing of the two stimulus components occurs in parallel. Within this class of models, they showed that if perceptual and decisional separability hold then timed marginal response invariance must also hold.[9]

The parallel models considered by Townsend et al. (2012) are grounded on the assumptions of stochastic linear systems, in which the activation in a channel is proportional to the magnitude of its input (Townsend & Wenger, 2004; Wenger & Townsend, 2006). There is a channel for each level of every stimulus component, and each channel accumulates evidence that

[9] Note that this result is weaker than the if and only if result that is possible if the RT-distance hypothesis is assumed to hold in the filtering task.

the relevant stimulus component is at the level to which the channel is tuned. In GRT and signal detection theory, if the likelihood ratio is monotonic, then evidence increases with distance from boundary (or criterion). For this reason, the parallel models considered by Townsend et al. (2012) are closely related to the models that Ashby and Maddox (1994) considered, which satisfy the RT-distance hypothesis.

Given this similarity, it is not surprising that marginal RT invariance and timed marginal response invariance are closely related. First, note that

$$P(\mathrm{a}_i\mathrm{b}_1, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_j) + P(\mathrm{a}_i\mathrm{b}_2, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_j) = P(\mathrm{a}_i, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_j). \tag{7.27}$$

Next note that marginal RT invariance is equivalent to assuming that for all $t > 0$:

$$\frac{P(\mathrm{a}_i, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_1)}{P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_1)} = \frac{P(\mathrm{a}_i, \mathbf{RT} \leq t|\mathrm{A}_i\mathrm{B}_2)}{P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_2)}. \tag{7.28}$$

Now Townsend et al. (2012) showed that if timed marginal response invariance holds then so does marginal response invariance (i.e., Eq. 7.24). Therefore, if the joint probabilities in the numerators of Eq. 7.28 are equal for all t, then the probabilties in the denominators are also equal. Therefore, when applied to the filtering task, marginal RT invariance and timed marginal response invariance are equivalent.

The summary statistics described so far are targeted at testing for perceptual separability. Another set of statistics targets perceptual independence. Report independence (called sampling independence in the early literature) is assessed for each individual stimulus and provides information about perceptual independence, again assuming that decisional separability holds. Report independence holds in the 2×2 identification task for stimulus A_iB_j if:

$$\begin{aligned} P(\mathrm{a}_i\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j) &= P(\mathrm{a}_i|\mathrm{A}_i\mathrm{B}_j) \times P(\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j) \\ &= [P(\mathrm{a}_i\mathrm{b}_1|\mathrm{A}_i\mathrm{B}_j) + P(\mathrm{a}_i\mathrm{b}_2|\mathrm{A}_i\mathrm{B}_j)] \\ &\quad \times [P(\mathrm{a}_1\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j) + P(\mathrm{a}_2\mathrm{b}_j|\mathrm{A}_i\mathrm{B}_j)]\,. \end{aligned} \tag{7.29}$$

Ashby and Townsend (1986) showed that if decisional separability holds, then a failure of report independence implies a violation of perceptual independence.

Townsend et al. (2012) also proposed an RT invariance condition that is similar to report independence. Specifically, timed report independence

holds for the response $a_i b_j$ with stimulus $A_k B_m$ if for all times $t > 0$

$$P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m, \mathrm{a}_i\mathrm{b}_j) \times P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m) = P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m, \mathrm{a}_i) \times P(\mathbf{RT} \leq t|\mathrm{A}_k\mathrm{B}_m, \mathrm{b}_j). \quad (7.30)$$

Townsend et al. (2012) showed that, within the class of parallel models they were considering, if decisional separability and perceptual independence both hold then timed report independence must hold.

Summary statistics approaches are complemented by the model-fitting approach described next. Indeed, since at least the work of Thomas (2001), there has been a focus on combining summary statistics and Gaussian model-fitting as complementary sources of converging evidence in supporting inferences (Cornes, Donnelly, Godwin, & Wenger, 2011; Von Der Heide et al., 2018; Wenger & Rhoten, 2020).

Fitting the Gaussian model to identification data

As mentioned earlier, a second approach for analyzing data from identification experiments is to fit a variety of different GRT models to the confusion matrices. Assumptions about perceptual interactions and decision processes can be tested by comparing model fits of nested models in which the restricted model makes some specific assumption, such as perceptual separability, and the more general model does not (Ashby & Perrin, 1988; Thomas, 2001). The primary advantage of this approach over the summary statistics approach is that, although it is parametric, it makes fewer assumptions about perceptual and decisional processes, and therefore should be less prone to false conclusions. The trade-off though is that it is more computationally intensive, since it often requires numerical integration.

This model-fitting approach is necessarily parametric since numerical predictions are possible only if a specific functional form is specified for the perceptual distributions. All previous applications of this approach have assumed that the perceptual distributions are multivariate normal. Furthermore, all applications have assumed that there are only two relevant sensory dimensions. In principle, the fitting algorithms (described below) are straightforward to extend to more than two dimensions, but such models could include many more free parameters and therefore significantly increased computation time. Thus, to date, applications that have fit GRT models to identification confusion matrices have assumed that the sensory distributions are bivariate normal pdfs, and the response regions are defined on a plane. A variety of different assumptions about decision processes are possible. Figure 7.8 illustrates six of these.

In an identification experiment with M stimuli, the resulting confusion

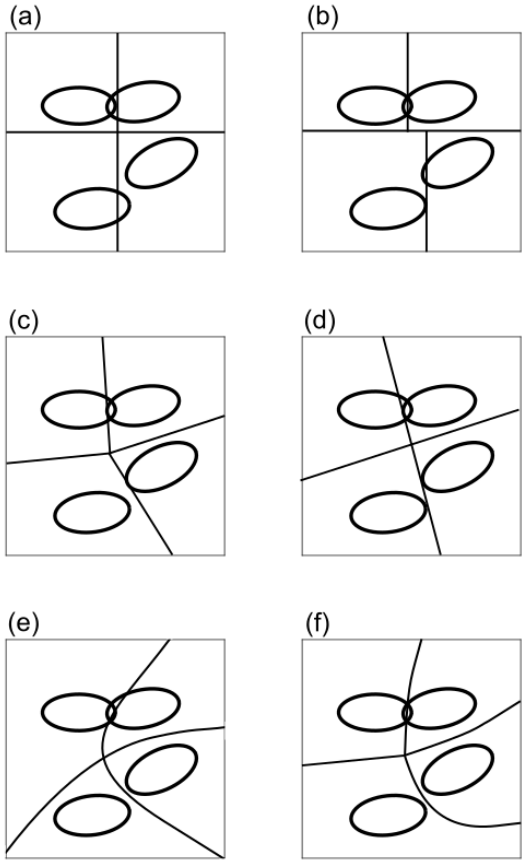

Figure 7.8 Different types of decision bounds used in GRT modeling. (a) Decisional separability is satisfied on both dimensions. (b) Decisional separability is satisfied on dimension 2, but not on dimension 1. (c) Decision bounds of the minimum distance classifier. (d) Decision bounds of the general linear classifier. (e) Decision bounds of the general quadratic classifier. (f) Decision bounds of the optimal classifier.

matrix includes $M \times (M-1)$ degrees freedom. As a result, this value fixes the maximum number of parameters that can be estimated. A bivariate normal distribution has a maximum of 5 free parameters – a mean and variance on both dimensions, and a covariance. Therefore, the smallest value of M for which the most general possible GRT model can be estimated is $M = 6$. In this case, the 6×6 confusion matrix has 30 degrees of freedom, and the 6 perceptual distributions needed to model the perceptual effects of the 6 stimuli have 30 parameters. However, the origin and unit of measurement on each perceptual dimension are arbitrary. Therefore, without loss of gen-

erality, the means on both dimensions can be set to 0 in any one perceptual distribution (to set the origin) and the variances in that distribution can be set to 1 (to set the unit of measurement). This reduces the number of free parameters to 26 (i.e., to $5M - 4$), which leaves a maximum of four parameters to model the decision bounds and assess the validity of the model. The fact that only four degrees of freedom remain rules out some decision models (e.g., the general quadratic classifier), but not all. For example, in Figure 7.8, the minimum distance and optimal classifiers have no free decision parameters, and the model that assumes decisional separability has only two free parameters (i.e., the two intercepts).

As the order of the confusion matrix increases above six, the degrees of freedom increases faster than the number of free parameters in the full GRT model. As a result, the larger the matrix, the more extra degrees of freedom there are to estimate decision bound parameters and to test the validity of the model. For example, Ashby et al. (2001) fit the full model to a variety of different 9×9 confusion matrices, which each have 72 degrees of freedom, and with nine stimuli the full model has only 41 free perceptual parameters.

On the other hand, note that the 2×2 factorial design, which as previously mentioned is the most popular identification experiment, includes only four stimuli. Therefore, the full model includes 16 free perceptual parameters (i.e, $5 \times 4 - 4$) and each confusion matrix includes only 12 degrees of freedom (i.e, 4×3). As a result, the full GRT model is not estimable in these experiments. So when GRT models are fit to single confusion matrices from 2×2 factorial designs, some assumptions must be made to reduce the number of free parameters.

When fitting any GRT model to identification data, parameter estimation is accomplished via the method of maximum likelihood. Denote the M stimuli by $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_M$ and the corresponding M responses by $\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_M$. Let n_{ij} denote the entry in row i and column j of the confusion matrix – that is, the frequency with which the observer responded R_j on trials when stimulus S_i was presented. Note that the n_{ij} are random variables, and the entries in each row of the confusion matrix have a multinomial distribution. In particular, if $P(\mathrm{R}_j|\mathrm{S}_i)$ is the true probability that response R_j is given on trials when stimulus S_i is presented, then the probability of observing the response frequencies $n_{i1}, n_{i2}, \ldots, n_{iM}$ in row i equals

$$P(n_{i1},n_{i2},\ldots,n_{iM}|\mathrm{S}_i) = \frac{N_i!}{n_{i1}!\, n_{i2}! \cdots n_{iM}!} P(\mathrm{R}_1|\mathrm{S}_i)^{n_{i1}} P(\mathrm{R}_2|\mathrm{S}_i)^{n_{i2}} \cdots P(\mathrm{R}_M|\mathrm{S}_i)^{n_{iM}}, \quad (7.31)$$

where N_i is the total number of stimulus S_i presentations (i.e., so $N_i =$

$\sum_j n_{ij}$). The probability or likelihood of observing the entire confusion matrix is the product of the probabilities of observing each row:

$$L = \prod_{i=1}^{M} P(n_{i1}, n_{i2}, \ldots, n_{iM}|\mathrm{S}_i). \tag{7.32}$$

In all Gaussian GRT models, $P(\mathrm{R}_j|\mathrm{S}_i)$ is computed by integrating a multivariate normal pdf over some response region, but different models make different assumptions about the pdf and about the shape of the region. The maximum likelihood parameter estimates are the numerical values of all model parameters that maximize the likelihood L of Eq. 7.32.

Two simplifications are common. First, some of the $P(\mathrm{R}_j|\mathrm{S}_i)^{n_{ij}}$ could be very small numbers, so it is common to find parameter values that maximize $\log L$ rather than L. Since log is a monotonic transformation, the parameter values that maximize L will also maximize $\log L$. Second, note that the factorial terms in Eq. 7.31 do not depend on the values of any model parameters, and therefore they are typically excluded from the parameter estimation process. Therefore, the common practice is to find the maximum likelihood estimates of all parameters by maximizing the monotonically related term

$$L^* = \sum_{i=1}^{M} \sum_{j=1}^{M} n_{ij} \log P(\mathrm{R}_j|\mathrm{S}_i), \tag{7.33}$$

where as already mentioned, the predicted probabilities $P(\mathrm{R}_j|\mathrm{S}_i)$ are computed by integrating under the multivariate normal pdf that models the sensory representation of stimulus S_i over the R_j response region.

The difficulty of computing the integrals required to maximize L^* depends on the nature of the decision bounds assumed by the model. Decisional separability simplifies things considerably because then the integral under a bivariate normal pdf reduces to the integral under a univariate normal marginal pdf. Under these conditions, Wickens (1992) derived the first and second derivatives necessary to estimate parameters of the model quickly using the Newton-Raphson method. In models that do not assume decisional separability, the integrals are under the bivariate normal pdf over irregularly shaped regions of the plane. As a result, numerical integration is required.

Ennis and Ashby (2003) proposed an efficient algorithm for evaluating these integrals that can be used to estimate the parameters of virtually any GRT model via standard minimization software. This algorithm was described in detail by Ashby and Soto (2015). Briefly however, the algorithm, which is described in Figure 7.9, includes the following 5 steps.

1) A set of D Z-values are preloaded into an array. Each Z-value is chosen

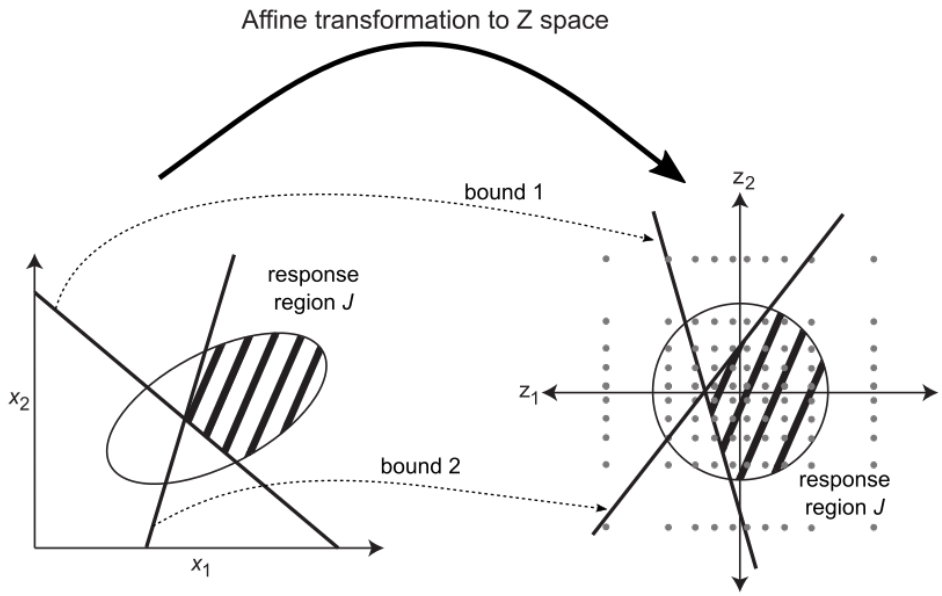

Figure 7.9 Schematic illustration of how numerical integration is performed in the multivariate normal GRT model via Cholesky factorization.

to be the center of an interval that has equal area under the Z distribution (i.e., under the pdf of a normal distribution with mean 0 and variance 1). The Cartesian product of this array with itself creates a grid of points in multidimensional space that are each the center of a rectangle (or hyperrectangle) that all have equal volumes under the multidimensional Z pdf (i.e., the gray points on the right side of Figure 7.9). If the GRT model assumes r perceptual dimensions then after this step there will be D^r grid points. For example, to fit two-dimensional GRT models, Ashby et al. (2001) set $D = 100$, which creates a grid of 10,000 points in bivariate Z-space, each of which is the center of a rectangle with volume .0001 (i.e., $.01^2$).

2) Note that GRT assumes that all entries in each row of a confusion matrix are computed by integrating under the same perceptual distribution. Different columns are associated with different response regions. The algorithm works row-by-row. The idea is to transform the perceptual distribution associated with the current row to a multivariate Z-distribution. This can always be accomplished via an affine transformation in which the linear transformation is based on the Cholesky factorization of the distribution's variance-covariance matrix. The second step is to compute this affine transformation.

3) Apply this affine transformation to the decision bounds. Since affine transformations preserve linearity, this step will convert linear bounds in perceptual space to linear bounds in Z-space.

4) Step through all D^r grid points and for each one, identify its associated response region. Each bound defines a discriminant function that assigns positive values to all points on one side and negative values to all points on the other side. With multiple bounds, each response region is characterized by a unique set of positive and negative discriminant values. So the response region of a point can be identified by examining its pattern of positive and negative discriminant values of all decision bounds after they have been transformed to Z-space.

5) Suppose the current grid point is identified as belonging to response region J. The final step is to increment the integral associated with response J by $1/D^r$.

The problems caused by insufficient degrees of freedom in 2×2 factorial designs disappear if GRT-wIND (Soto et al., 2015) is used instead of the traditional GRT model. GRT-wIND is fit simultaneously to the individual confusion matrices of all observers. Soto, Zheng, Fonseca, and Ashby (2017) developed an R package that fits this model using only a few commands. GRT-wIND assumes that all observers share the same group perceptual representation, which is described by the full GRT model, even in 2×2 factorial designs. Thus, GRT-wIND assumes that basic perceptual properties, such as perceptual separability and perceptual independence, or their violations, are shared by all observers. The model assumes that different observers produce different confusion matrices for two reasons – they allocate different amounts of attention to the two perceptual dimensions, and they use different decision bounds. Thus, fitting the model returns estimates of (1) the group perceptual representation (i.e., the full GRT perceptual model), (2) the proportion of attention allocated to the two perceptual dimensions by each observer, and (3) unique decision bounds for each observer. Soto et al. (2015) fit GRT-wIND to the confusion matrices of 24 different observers in a face identification experiment that used a 2×2 factorial design in which the 4 stimulus faces were created by crossing two facial identities with two emotional expressions. The 24 matrices included a total of 288 degrees of freedom (i.e., 24×12). The GRT-wIND model included an average of 6.67 free parameters for each individual confusion matrix, which is less than typical applications of traditional GRT models to 2×2 designs. GRT-wIND accounted for 99.52% of the variance in the 24 confusion matrices. Even more impressively, GRT-wIND provided a better fit than the best-fitting

traditional GRT model to the data of 18 of the 24 participants.[10] Furthermore, GRT-wIND suggested that in this group of 24 observers, emotional expression was perceptually separable from facial identity, but identity was not separable from expression. In contrast, a traditional GRT analysis could only report how many of the individual participants showed this pattern.

GRT accounts of identification data have been spectacularly successful. For most of the last four decades of the 20th century, the most successful model of identification, by far, was the Luce-Shepard choice model (Luce, 1963; Shepard, 1957), which assumes that

$$P(\mathrm{R}_j|\mathrm{S}_i) = \frac{\eta_{ij}\beta_j}{\sum_{k=1}^{M}\eta_{ik}\beta_k}, \tag{7.34}$$

where η_{ij} is the similarity between stimuli S_i and S_j and β_j is the bias toward response R_j (without loss of generality, one can set $\eta_{ii} = 1$ for all values of i and $\sum \beta_j = 1$). To ensure that the model is testable, similarity is assumed to be symmetric (i.e., so that $\eta_{ij} = \eta_{ji}$ for all values of i and j). The Luce-Shepard choice model was so successful that for many years, it was the standard against which competing models were compared. For example, in 1992, J. E. K. Smith summarized its performance by concluding that it "has never had a serious competitor as a model of identification data. Even when it has provided a poor model of such data, other models have done even less well" (J. K. Smith 1992, p. 199). Even so, the model was never considered completely satisfactory – primarily because a good fit provides little insight into the psychological processes of the observer producing the data. The model merely says that the probability of confusing stimulus S_j for S_i is proportional to the product of the similarity between the two stimuli and the bias toward response R_j (the denominator in Eq. 7.34 is just a normalizing constant). Also, note that the model makes no predictions about how a decision is reached. It simply predicts the proportion of R_j responses to expect over the course of a large number of S_i trials.

GRT provided the first models that ended the dominance of the Luce-Shepard choice model, at least for identification data collected from experiments with stimuli that differed on only a couple of stimulus dimensions. In virtually every such comparison, the GRT model provided a substantially better fit than the Luce-Shepard choice model, in many cases with fewer free parameters (Ashby et al., 2001). Even so, it is important to note that the Luce-Shepard choice model is still valuable, especially in the case of identi-

[10] This is because the full traditional-GRT model is not estimable in 2×2 designs, but the full GRT-wIND model is estimable.

fication experiments in which the stimuli vary on many unknown stimulus dimensions.

7.4.5 Extensions to response time

Like SDT, GRT was originally developed to account exclusively for accuracy data. Even so, there have been a number of extensions of the theory that attempt also to account for RTs. These are generally of two types. One approach is to add assumptions to GRT that allow the theory to make RT predictions but are not detailed enough to account for psychological process. Thus, like the original version of GRT (and SDT), the resulting models are descriptive, or in the language of Marr (1982), computational. The other approach is to add enough structure to GRT to model psychological process – thereby producing models that Marr identified as algorithmic. We briefly review both types in turn.

Computational-level accounts of RT

The principle example of this approach was to add an assumption called the RT-distance hypothesis to GRT, which simply assumes that RT decreases with the distance between the percept and the decision bound. This assumption was first investigated in SDT (e.g., Murdock 1985). The idea is that if decisions are made by comparing a percept to a decision bound or criterion, then the greater the distance between the two, the easier, and hence the faster the decision. This simple assumption has received considerable empirical support (Ashby, Boynton, & Lee, 1994; Murdock, 1985). As noted earlier, Ashby and Maddox (1994) showed that if the RT-distance hypothesis holds then strong nonparametric RT tests of perceptual separability are possible.

Process models of RT

This has been the more popular approach. Ashby (2000) generalized the drift-diffusion model described earlier to multiple perceptual dimensions. In this version, the perceptual representations are the same as in classical GRT. Like the drift diffusion model, application was restricted to tasks with two response alternatives. On each trial, the observer's experience with the stimulus was assumed to produce repeated samples from the relevant perceptual distribution. Each sample is compared to the decision bound and a signed distance is computed, which equals distance-to-bound if the percept is in the A region and minus distance-to-bound if it is in the B region. At this point,

the model is identical to the drift diffusion model – that is, the signed distances are cumulated, and sampling continues until the sum crosses an upper or lower barrier (exactly as in Figure 7.5, except with an "A" response replacing "YES" and a "B" response replacing "NO"). Ashby (2000) showed that this model includes the static version of GRT as a special case, and showed that the variance-covariance matrices estimated in classical applications of GRT are corrupted by decisional influences. For example, consider two conditions in which the task is identical but participants are pressed to respond more quickly in one than the other. In general, we expect more errors in the condition with speed stress. Fitting the static GRT model to these data would suggest that perceptual noise increases with speed stress. In contrast, the stochastic version of GRT accounts for these data by reducing the distance to the response barriers in the speeded condition (i.e., the numerical values of A and B in Figure 7.5), but not changing perceptual noise.

More recently, P. L. Smith (2019) proposed a similar model, except based on a circular diffusion process. The model can be applied to a variety of different tasks, but consider its application to the 2×2 factorial identification experiment with stimuli A_1B_1, A_1B_2, A_2B_1, and A_2B_2. As mentioned earlier, in static GRT models of this task, the origin of the perceptual space is arbitrary. Suppose we define the origin as the center point of the 4 perceptual means (i.e., the mean of the means), and the drift is determined by cumulating random samples from the perceptual distribution associated with the presented stimulus (e.g., scaled by some multiplicative constant). Then the drift will generally be outwards and in the direction of the perceptual mean of the stimulus. P. L. Smith (2019) assumed a single circular absorbing barrier that is divided into 4 quadrants – one associated with each response alternative. The accumulation process continues until absorption occurs, at which point the associated response is given. A response bias toward or against a particular response can be implemented by setting the angle of the response quadrant associated with that response to be greater or less than $90°$, respectively. Because this task includes more than two response alternatives, the stochastic GRT model proposed by Ashby (2000) is not even defined in this case. So in this sense, Smith's model has a considerable advantage over the model proposed by Ashby. On the other hand, the circular-diffusion model does not include decision bounds, so it is unclear how the model would account for performance differences that arise, for example, when the participant switches to or away from bounds that satisfy decisional separability.

As noted earlier, Townsend and colleagues (Townsend et al., 2012; Townsend

& Wenger, 2004; Wenger & Townsend, 2006) interpreted GRT within the framework of stochastic linear dynamical systems. These models assume the stimulus dimensions or components are processed by parallel channels that are potentially interactive (Townsend, Liu, Zhang, & Wenger, 2020). Activation in each channel is accumulated until it reaches a criterion level, and the outputs of the different channels are then passed to decisional operators (e.g., Boolean AND or OR gates). Like the drift-diffusion interpretations of GRT, these models make simultaneous accuracy and RT predictions. They have also been used to model configurality (Wenger & Townsend, 2006) and to derive new RT summary statistics that can be used to test for perceptual separability and perceptual independence.

7.4.6 Extensions to neuroscience

GRT was developed before the cognitive neuroscience revolution that began in the 1990s. As a result, for its first several decades of existence, GRT was a purely perceptual and cognitive theory. But during the past several decades there has been progress on two fronts. First, much has been learned about the architecture and functioning of the neural circuits that implement the perceptual and decision processes hypothesized by GRT. And second, GRT analyses have recently been extended to neuroscience data, in particular to data from neuroimaging experiments. This section briefly reviews these trends. For more details, see Ashby and Soto (2016) and Soto, Vucovich, and Ashby (2018).

There is now overwhelming evidence that humans have multiple learning systems that for the most part are neuroanatomically and functionally distinct (e.g., Ashby and Maddox 2005; Eichenbaum and Cohen 2001; Squire 2004). The most complete description of two of the most important learning systems is arguably provided by the COVIS theory (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Valentin, 2017). COVIS assumes separate explicit-reasoning and procedural-learning systems that compete for access to response production. The explicit-reasoning system uses executive attention and working memory to learn explicit rules, and is mediated by a broad neural network that includes the prefrontal cortex, anterior cingulate, head of the caudate nucleus, and the hippocampus. In contrast, the procedural system uses dopamine-mediated reinforcement learning when the optimal strategy is difficult or impossible to describe verbally, and key structures include the striatum and premotor cortex.

Knowing which learning system participants are using can facilitate a subsequent GRT analysis because the explicit system is constrained to use

bounds that satisfy decisional separability (at least locally), whereas the procedural system is not. The explicit system learns and applies explicit rules that can be described using Boolean algebra. More specifically, it makes independent decisions about the level of the stimulus (e.g., high versus low) on one or more dimensions and then combines the outcomes of these separate decisions using simple logical operators, such as "and" to produce conjunction rules and "or" to produce disjunctions. When translated into decision bounds, the resulting response regions can always be separated by piecewise linear bounds, in which each piece is a vertical or horizontal line segment. Thus, each piece satisfies decisional separability. In contrast, the procedural system implements less constrained decision strategies that are compatible with any of the decision bounds that are used when fitting GRT models. For these reasons, if decisional separability is assumed, then it is vital to select experimental conditions that favor explicit reasoning over procedural learning.

Soto et al. (2018) extended GRT analysis to neuroimaging data in the context of a study examining the relationship between facial identity and perceived emotion. When a visual stimulus is presented to an observer, it causes activation in many areas within the visual system. The perceptual representation modeled in GRT likely depends on activation in some higher-level visual area. If this representation violates perceptual separability (or perceptual independence), then an obvious and important question is when and where separability (or independence) was first violated within the processing stream? To address this question, Soto et al. (2018) first defined the concepts of *encoding separability* and *encoding independence*. If a stimulus dimension is encoded in some brain region of interest in exactly the same way when an irrelevant dimension is varied, then the former shows encoding separability from the latter. Similarly, if the neural representations of two stimulus dimensions are statistically independent in some region of interest, then they satisfy encoding independence. Next, Soto et al. (2018) proposed empirical tests of these constructs that are based on summary statistics derived from applying pattern classifiers to fMRI data. For example, the first step might be to construct a support vector machine that classifies the level of stimulus dimension A in some brain region of interest as 1 or 2 (following methods described e.g., by Ashby 2019). *Decoding separability* holds if the distributions of decoded values of dimension A are invariant across changes in a second, irrelevant dimension B.[11]

[11] Operationally, this can be tested in the following way. Consider an identification experiment with stimuli A_1B_1, A_1B_2, A_2B_1, and A_2B_2. First, compute the distance of each activity vector to the classifier hyperplane. Second, estimate the distributions of the A_1 and A_2

Similarly, Wenger and Rhoten (2020) demonstrated that it was possible to use the timing of a feature in EEG data to draw inferences regarding independence and separability in a study of visual perceptual learning. Specifically, they used the onset time of the lateralized readiness potential (LRP). The LRP is a negative-going waveform, measured in central electrodes contralateral to the motor response that it precedes, and is interpreted as indicating that sufficient processing has been completed in order to program the motor response. The onset time of the LRP was shown to be strongly correlated with observable RT. Consequently, when those onset times were analyzed with respect to timed marginal response invariance and timed report independence (see the subsection entitled "Summary statistics approach"), they were found to support inferences that were consistent with the inferences drawn from the response frequencies.

7.5 Concluding Remarks

The power and generality of statistical decision theory – SDT in one dimension and GRT in multiple dimensions – should confirm Estes' evaluation that SDT is "... the most towering achievement of basic psychological research in the last half century" (Estes 2002, p. 15). One would be hard-pressed to name a sub-discipline of the behavioral sciences (cognitive neuroscience included) that do not concern themselves with aspects of identification and categorization (classification). This fact, along with the fact that SDT "scales" to dealing with neurophysiological data, perhaps reinforces Wixted's opinion that "... it should not be possible to earn a Ph.D. in experimental psychology without having some degree of proficiency in signal detection theory" (Wixted 2020, p. 225). Along with these kinds of advances, we should note that a critical strength of the community of researchers associated with SDT and GRT is the unflinching willingness to tackle difficult problems, such as the identifiability issues discussed here. Investigators have added and continue to develop novel and improved methods for framing hypotheses and connecting theory and data.

7.6 Related Literature

Link (1994) and Wixted (2020) provide excellent historical overviews of the antecedents to SDT and to its early years. The original classic text on SDT

distances separately when B is at level 1 and at level 2. Finally, compare the A_1 distributions when B is at level 1 and at level 2, and also compare the A_2 distributions when B is at level 1 and level 2.

was by Green and Swets (1966). It remains relevant today, especially for its treatment of ideal observer theory. For more recent texts, see Macmillan and Creelman (2005) or Wickens (2002).

There is no text on GRT, although this topic is briefly covered by Macmillan and Creelman (2005). Even so, there are a few recent GRT tutorials, including by Ashby and Soto (2015) and Silbert and Hawkins (2016). For a review of the mathematical foundations of GRT, see Fukunaga (2013).

7.7 Acknowledgments

We thank Matthew Crossley and Jeffrey Inglis for their helpful comments on this manuscript.

References

Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology*, *44*(2), 310–329.

Ashby, F. G. (2019). *Statistical analysis of fMRI data, Second edition*. Cambridge MA: MIT press.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, *55*(1), 11–27.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*(4), 423–466.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149-178.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124–150.

Ashby, F. G., & Soto, F. A. (2015). Multidimensional signal detection theory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 13–34). New York: Oxford University Press.

Ashby, F. G., & Soto, F. A. (2016). The neural basis of general recognition theory. In *Mathematical models of perception and cognition: Volume II, A Festschrift for James T. Townsend* (pp. 1–31). New York: Routledge.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and cognitive tests. In H. Cohen & C. Lefebvre (Eds.),

Handbook of categorization in cognitive science, Second Edition (pp. 157–188). New York: Elsevier.

Ashby, F. G., & Valentin, V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience, Fourth Edition, Volume 5: Methodology* (pp. 1–41). New York: Wiley.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, *130*(1), 77–96.

Balakrishnan, J. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1189–1206.

Bridgman, P. W. (1945). Some general principles of operational analysis. *Psychological Review*, *52*(5), 246–249.

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283–319.

Cornes, K., Donnelly, N., Godwin, H., & Wenger, M. J. (2011). Perceptual and decisional factors affecting the detection of the Thatcher illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 645-668.

Creelman, C. D. (2015). Signal detection theory, history of. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences (second edition)* (pp. 952–956). New York: Elsevier.

Crossley, M. J., & Ashby, F. G. (2015). Procedural learning during declarative control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1388–1403.

DeValois, R. L., & De Valois, K. K. (1990). *Spatial vision.* New York: Oxford University Press.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859.

Dunnington, G. W., Gray, J., & Dohse, F. E. (2004). *Carl Friedrich Gauss: Titan of science.* MAA.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain.* New York: Oxford University Press.

Ennis, D. M., & Ashby, F. G. (2003). Fitting decision bound models to identification or categorization data. *Unpublished manuscript. Available at https://labs.psych.ucsb.edu/ashby/gregory/sites/labs.psych.ucsb.edu.ashby.gregory/files/pubs/ennisashby2003.pdf*.

Estes, W. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*(1), 3–25.

Fechner, G. T. (1860). *Elements of psychophysics (translated by H. E. Adler, 1966).* Leipzig: Breitkopf and Hartel (Holt, Rinehart, and Winston).

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(1), 69–78.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition, Second edition.* San Diego: Academic Press.

Fullerton, G., & Cattell, J. (1892). On the perception of small differences. University of Pennsylvania Philosophy Series, No. 2. *Philadelphia: University of*

Pennsylvania.

Garner, W. R. (1974). *The processing of information and structure*. New York: Wiley.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*(3), 225–241.

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*(3), 149–159.

Garner, W. R., & Morton, J. (1969). Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin*, *72*, 233-259.

Green, D. M. (1964). General prediction relating yes-no and forced-choice results. *The Journal of the Acoustical Society of America*, *36*(5), 1042–1042.

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

Helmholtz, H. v. (1867). *Handbuch der physiologischen Optik, vol. 9*. Leipzig: Voss.

Kadlec, H. (1995). Multidimensional signal detection analyses (MSDA) for testing separability and independence: A PASCAL program. *Behavior Research Methods, Instruments, & Computers*, *4*, 442-458.

Kadlec, H., & Townsend, J. T. (1992a). Implications of marginal and conditional detection parameters for the separabilities and independence of perceptual dimensions. *Journal of Mathematical Psychology*, *36*, 325-374.

Kadlec, H., & Townsend, J. T. (1992b). Signal detection analysis of dimensional interactions. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (p. 181-228). Hillsdale, NJ: Erlbaum.

Killeen, P. R., & Taylor, T. J. (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology*, *48*(6), 432–434.

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman–Pearson. *The British Journal for the Philosophy of Science*, *57*(1), 69–91.

Link, S. W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science*, *5*(6), 335–340.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*(1), 77–105.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Volume 1* (p. 103-190). New York: Wiley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650–662.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review*, *11*(5), 945–952.

Maddox, W. T., & David, A. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 100–107.

Marcum, J. I. (1947). *A statistical theory of target detection by pulsed radar* (Tech. Rep.). Santa Monica, CA: Rand Corporation.

Marr, D. (1982). *Vision: A computational investigation into the human represen-*

tation and processing of visual information. New York: Freeman.

Menneer, T., Wenger, M., & Blaha, L. (2010). Inferential challenges for general recognition theory: Mean-shift integrality and perceptual configurality. *Journal of Vision*, *10*(7), 1211–1211.

Murdock, B. B. (1985). An analysis of the strength-latency relationship. *Memory & Cognition*, *13*(6), 511–521.

Murphy, J., Gray, K. L. H., & Cook, R. (2017). The composite face illusion. *Psychonomic Bulletin & Review*, *24*(2), 245–261.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*(694-706), 289–337.

O'Toole, A. J., Wenger, M. J., & Townsend, J. T. (2001). Quantitative models of perceiving and remembering faces: Precedents and possibilities. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (p. 1-38). Mahwah NJ: Erlbaum.

Peterson, W. W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE professional group on information theory*, *4*(4), 171–212.

Peterson, W. W., & Birdsall, T. G. (1953). *The theory of signal detectability: Part I. The general theory. Electronic Defense Group* (Tech. Rep.). Technical Report 13, June 1953. University of Michigan.

Piepers, D., & Robbins, R. (2012). A review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology*, *3*, 559.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Richler, J. J., & Gauthier, I. (2013). When intuition fails to align with data: A reply to Rossion (2013). *Visual Cognition*, *21*(2), 254–276.

Richler, J. J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. *Psychological Bulletin*, *140*(5), 1281–1302.

Richler, J. J., Gauthier, I., Wenger, M. J., & Palmeri, T. J. (2008). Holistic processing of faces: Perceptual and decisional components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 328–342.

Rose, A. (1942). The relative sensitivities of television pickup tubes, photographic film, and the human eye. *Proceedings of the IRE*, *30*(6), 293–300.

Rose, A. (1948). The sensitivity performance of the human eye on an absolute scale. *Journal of the Optical Society of America*, *38*(2), 196–208.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139-253.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Silbert, N. H., & Hawkins, R. X. D. (2016). A tutorial on general recognition theory. *Journal of Mathematical Psychology*, *73*, 94–109.

Silbert, N. H., & Thomas, R. D. (2013). Decisional separability, model identification, and statistical inference in the general recognition theory framework. *Psychonomic Bulletin & Review*, *20*(1), 1–20.

Silbert, N. H., & Thomas, R. D. (2017). Identifiability and testability in GRT with

individual differences. *Journal of Mathematical Psychology*, *77*, 187–196.
Smith, J. K. (1992). Alternative biased choice models. *Mathematical Social Sciences*, *23*(2), 199–219.
Smith, P. L. (2019). Linking the diffusion model and general recognition theory: Circular diffusion with bivariate-normally distributed drift rates. *Journal of Mathematical Psychology*, *91*, 145–158.
Soto, F. A., Vucovich, L., Musgrave, R., & Ashby, F. G. (2015). General recognition theory with individual differences: A new method for examining perceptual and decisional interactions with an application to face perception. *Psychonomic Bulletin & Review*, *22*(1), 88–111.
Soto, F. A., Vucovich, L. E., & Ashby, F. G. (2018). Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data. *PLoS Computational Biology*, *14*(10), e1006470.
Soto, F. A., Zheng, E., Fonseca, J., & Ashby, F. G. (2017). Testing separability and independence of perceptual dimensions with general recognition theory: A tutorial and new R package (grtools). *Frontiers in Psychology*, *8*, 696.
Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171-177.
Tanner, W. (1956). Theory of recognition. *The Journal of the Acoustical Society of America*, *28*(5), 882–888.
Thomas, R. D. (2001). Characterizing perceptual interactions in face identification using multidimensional signal detection theory. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (p. 193-228). Mahwah, NJ: Erlbaum.
Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.
Thurstone, L. L. (1927b). Psychophysical analysis. *The American Journal of Psychology*, *38*(3), 368–389.
Townsend, J. T., Houpt, J. W., & Silbert, N. H. (2012). General recognition theory extended to include response times: Predictions for a class of parallel systems. *Journal of Mathematical Psychology*, *56*(6), 476–494.
Townsend, J. T., Liu, Y., Zhang, R., & Wenger, M. J. (2020). Interactive parallel models: No Virginia, violation of Miller's race inequality does not imply coactivation and yes Virginia, context invariance is testable. *The Quantitative Methods for Psychology*, *16*(2), 192–212.
Townsend, J. T., & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003–1035.
Townsend, J. T., & Wenger, M. J. (2015). On the dynamic perceptual characteristics of Gestalten: Theory-based methods. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (pp. 948–968). Oxford.
Treutwein, B., Rentschler, I., & Caelli, T. (1989). Perceptual spatial frequency-orientation surface: Psychophysics and line element theory. *Biological Cybernetics*, *60*(4), 285–295.
Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.
Van Meter, D., & Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 119–145.
Von Der Heide, R. J., Wenger, M. J., Bittner, J. L., & Fitousi, D. (2018). Con-

verging operations and the role of perceptual and decisional influences on the perception of faces: Neural and behavioral evidence. *Brain and Cognition*, *122*, 59–75.

Wenger, M. J., & Ingvalson, E. M. (2002). A decisional component of holistic encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 872-892.

Wenger, M. J., & Ingvalson, E. M. (2003). Preserving informational separability and violating decisional separability in facial perception and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1106–1118.

Wenger, M. J., & Rhoten, S. E. (2020). Perceptual learning produces perceptual objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 455—475.

Wenger, M. J., & Townsend, J. T. (2006). On the costs and benefits of faces and words. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 755-779.

Wickens, T. D. (1992). Maximum-likelihood estimation of a multivariate Gaussian rating model with excluded data. *Journal of Mathematical Psychology*, *36*(2), 213–234.

Wickens, T. D. (2002). *Elementary signal detection theory.* Oxford University Press, USA.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 14.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*, 747-759.