The Forgotten History of Signal Detection Theory

John T. Wixted
University of California, San Diego

Signal detection theory is one of psychology’s most well-known and influential theoretical frameworks. However, the conceptual hurdles that had to be overcome before the theory could finally emerge in its modern form in the early 1950s seem to have been largely forgotten. Here, I trace the origins of signal detection theory, beginning with Fechner’s (1860/1966) Elements of Psychophysics. Over and above the Gaussian-based mathematical framework conceived by Fechner in 1860, nearly a century would pass before psychophysicists finally realized in 1953 that the distribution of sensations generated by neural noise falls above, not below, the threshold of conscious awareness. An extensive body of single-unit recording and neuroimaging research conducted since then supports the idea that sensory noise yields genuinely felt conscious sensations even in the complete absence of stimulation. That hard-to-come-by insight in 1953 led immediately to the notion of a movable decision criterion and to the methodology of receiver operating characteristic (ROC) analysis. Over the ensuing years, signal detection theory and ROC analysis have had an enormous impact on basic and applied science alike. Yet, in some quarters of our field, that fact appears to be virtually unknown. By tracing both its fascinating origins and its phenomenal impact, I hope to illustrate why no area of experimental psychology should ever be oblivious to signal detection theory.

Keywords: Gustav Fechner, high-threshold theory, Louis Thurstone, receiver operating characteristic analysis, sensory threshold

Gustav Fechner: Architect of the 2AFC Signal Detection Framework

When considering Fechner’s contributions to experimental psychology, what comes to mind is probably not signal detection theory. Instead, what is more likely to come to mind is his famous psychophysical law, according to which subjective sensation ($S$) is...
a log function of objective stimulus intensity \( I \). Fechner’s Law was based on Weber’s Law, which holds that as stimulus intensity is gradually increased, the difference is at first imperceptible, but as the stimulus intensity increases still further, the change finally becomes detectable. Let \( I \) represent the starting intensity (e.g., \( I = 100\)-g weight) and \( \Delta I \) represent the increment in intensity needed for the change to be noticeable (e.g., \( \Delta I = 5 \) g). That change is known as the just-noticeable-difference (JND), and Weber’s Law holds that \( \Delta I/I = k \), where \( k \) is a constant. Thus, the larger \( I \) is, the larger \( \Delta I \) needs to be for the change to be noticeable.

Although rarely discussed, the concept of the JND is a moving target. Imagine running different groups of blindfolded subjects and asking them to indicate when they notice a change in heaviness as water is slowly added to a 100-g cup held in one’s hand. For one group, the instructions might stipulate that they should declare a noticed change only when they are 100% certain that it has occurred. The JND for this group might be 10 g. For a second group, everything is the same except that they are asked to declare a noticed change when they are at least 50% certain that it has occurred. The average JND for this group might be lower, perhaps only 5 g. Still a third group is asked to declare a noticed change when they are only 10% certain that it has occurred. For this group, the JND might turn out to be even lower (e.g., 1 g). The point is that there is no single JND for a given intensity because it becomes ever smaller as the confidence required to detect a change decreases.

**Fechner’s Law.** Fechner argued that the right side of Weber’s Law (\( \Delta I/I = k \)) could be construed as the constant subjective change in sensation (\( \Delta S \)) that occurs when a just noticeable increment, \( \Delta I \), is added to \( I \) (Murray, 1993). Thus, \( k \) in Weber’s Law can be replaced by \( \Delta S \), such that \( \Delta S = \lambda(\Delta I/I) \), where \( \lambda \) is simply a scaling constant. Although Fechner did not point out that the JND decreases as a function of confidence, he nevertheless argued that, in the limit, Weber’s Law can be conceptualized as a differential equation according to which \( dS = \lambda(dI/I) \), where \( dS \) and \( dI \) represent infinitesimal changes in \( S \) and \( I \), respectively. When both sides of this equation are integrated, the result is as follows:

\[
S = \lambda \ln(I) + C. \tag{1}
\]

This is the second-to-last step in the derivation of Fechner’s Law. The last step is a critical one because there is some tension between it and what would later become a central tenet of signal detection theory. Thus, it is worth dwelling on that step before considering how Fechner separately conceived the 2AFC signal detection framework.

The last step in the derivation of Fechner’s Law makes the seemingly incontrovertible assumption that there is some small intensity, \( I_0 \), so small that it elicits no sensation whatsoever. That is, when stimulus intensity is \( I_0 \), \( S = 0 \). In this way of thinking, \( I_0 \) is the stimulus threshold. Because \( S = 0 \) at the threshold stimulus intensity, it follows from Equation 1 that \( 0 = \lambda \ln(I_0) + C \). Solving for \( C \) yields \( C = -\lambda \ln(I_0) \), and substituting \( -\lambda \ln(I_0) \) for \( C \) in Equation 1 yields \( S = \lambda \ln(I) - \lambda \ln(I_0) \). Simplifying this expression yields Fechner’s Law:

\[
S = \lambda \ln(I/I_0) \tag{2}
\]

According to Equation 2, sensation is a log function of stimulus intensity for any intensity greater than \( I_0 \). Note that when \( I = I_0 \) (i.e., when stimulus intensity equals but does not exceed the threshold), \( S = 0 \), as it should. Thus, for conscious sensation to occur, stimulus intensity must exceed the threshold.

Depending on how it is construed, the seemingly undeniable threshold assumption appears to be completely at odds with signal detection theory. As discussed in some detail throughout this article, signal detection theory denies the relevance of a threshold, whether the test involves a 2AFC task (where the subject must decide which of two stimuli is correct) or a yes/no task (where the subject must decide whether or not a stimulus was presented). At first glance, this might seem like a dubious idea. After all, there must be some small intensity that fails to elicit even the slightest neural response, in which case it is hard to imagine how it would give rise to a subjective sensation. For example, if I place a weight in your hand that is no heavier than a hydrogen atom, no nerve cell will fire in response. If there is no response generated by the stimulus in the nervous system, then there can be no corresponding experiential sensation upon which to base a yes/no decision. Right?

Wrong, and that is absolutely the crux of the issue. Of course there is a stimulus intensity so small in magnitude that it fails to yield any response in the nervous system. The more interesting question—the one that signal detection theory asks you to consider—is whether, on trials involving a stimulus that falls below the physiological threshold (including trials that involve no stimulus at all), one ever subjectively experiences sensation anyway. Signal detection theory holds that sensation in the absence of stimulation does, indeed, occur. As an example, most people have confidently experienced a vibrating cell phone even though no text message just arrived. I return to this pivotal issue later after first reviewing Fechner’s lesser known contribution, namely, a scaling methodology for the 2AFC task based on Gaussian error. Readers who are interested in the conceptual development of signal detection theory more so than the mathematics of it can easily skip the remainder of this section and jump to the next main section entitled “The Noise Distribution and Its Relationship to Conscious Awareness (1860–1953.)”

**Fechner’s 2AFC signal detection framework.** In a revision of the historical record, Link (1992, 1994) argued that credit for conceiving of signal detection theory belongs to Fechner (1860/1966), who, by devising a new way to scale mental experience, literally invented the field of experimental psychology itself.\(^1\) In truth, beyond Fechner’s seminal contributions, there were additional ingenious contributions still to come before the theory would emerge in its current form in 1953. Nevertheless, there is no disputing the fact that much of what we think of signal detection theory today was spelled out in detailed equations (but not in figures) by Fechner in the 19th century. The fact that Fechner did not actually draw the iconic signal detection model he had in mind—namely, two equal-variance Gaussian distributions placed on a psychological continuum with a criterion centered between them—may explain why his fundamental insight went largely overlooked until Link (1992, 1994) drew attention to it.

Earlier in the 19th century, Carl Friedrich Gauss conceptualized physical measurements in terms of a true value plus random error.

---

\(^1\) Fechner invented experimental psychology in the sense that he was the first to objectively scale mental events. Others credit Wundt as being the “father of psychology” because of his efforts to establish psychology as a discipline separate from philosophy, which, according to Diamond (2001), can be traced at least as far back to Wundt (1862).
For example, imagine using an unbiased but slightly erratic balance scale to measure Weight A many times, and the readings turn out to have a mean ($X_A$) of 50 g and a standard deviation ($s_A$) of 2 g. Doing the same for a slightly heavier Weight B yields $X_B = 53$ g and $s_B = 2$ g. Thus, for this equal-variance scenario, where $s = s_A = s_B = 2$ g, we can say that $(\overline{X}_B - \overline{X}_A)/s = (53 - 50 \text{ g})/2 \text{ g} = 1.5$. In other words, as illustrated in upper panel of Figure 1, $X_A$ and $X_B$ are 1.5 standard deviations apart.

Why represent the distance between the two means this way (i.e., in terms of the standard deviation) instead of simply noting that the means are 3 g apart? Because the measured distance between the two means in standard deviation units would not change even if we repeated this exercise with a different but equally erratic balance scale that measured weight in ounces instead of grams. The means would be 3 g apart using one scale and 0.11 oz apart using the other, but no matter what interval-scale

![Distribution of A and B](image)

**Figure 1.** Upper panel: On a given trial, two weights (Weight A and Weight B) are individually measured using a balance scale, and the observed values are $A$ and $B$, respectively. Assume that, over many trials, $\overline{X}_A = 50$ g, $\overline{X}_B = 53$ g, and $s_A = s_B = s = 2$ g. Lower panel: How often would a random draw from the distribution of scores for Weight A exceed a random draw from the distribution of scores for Weight B? This question is equivalent to asking how often $B - A$ would be negative. As illustrated in the lower panel, the distribution of $B - A$ difference scores would have a mean of $\overline{X}_{B-A} = \overline{X}_B - \overline{X}_A = 3$ gm, and a standard deviation of $s_{B-A} = \sqrt{s_A^2 + s_B^2} = 2\sqrt{2}$. For this difference-score distribution, the zero-point on the x-axis falls 1.06 standard deviations below the mean. That is, as shown in the equation at the bottom of the figure, $z_{A>B} = -1.06$. The proportion of the difference-score distribution that falls below zero (i.e., integrating the distribution from $-\infty$ up to 0) is equal to .144 (shaded region). That is, $\Phi(z_{A>B}) = \Phi(-1.06) = .144$. 


unit of weight we use in the physical world, for equally precise scales, we would still find that the means are 1.5 \( s \) apart. Thus, an advantage of expressing the difference between the two means in terms of standard deviation units is that it generalizes across all interval-scale measures.

When we move from the world of physical scaling to Fechner's world of psychophysical scaling, we lose the metric provided by the physical balance scale (e.g., subjective sensations are not directly measured in grams or ounces), but we can still estimate the psychological distance between the means of the sensations generated by Weight A and Weight B in terms of the standard deviation of subjective sensations across trials. To do so, it might seem that the obvious approach would be to treat subjects like balance scales by asking them to subjectively rate sensations of heaviness across trials using a numerical scale. Conceptually, this direct magnitude estimation approach makes sense, and there are times when it seems to work remarkably well (e.g., Dubé, Tong, Westfall, & Bauer, 2019; Mickes, Wixted, & Wais, 2007). However, a reasonable concern is that the ratings, unlike physical distance measured in grams or ounces, might not lie on an interval measurement scale (e.g., Krueger, 1989). If not, then it would not make sense to scale the psychological distance between the two means in terms of a standard deviation computed from those ratings.

To avoid that problem, Fechner used an indirect approach to psychological scaling by having observers rate the relative sensations generated by two weights presented on each trial. Instead of providing a subjective measure of heaviness using a rating scale, observers were asked to pick the heavier of the two weights (i.e., a 2AFC task was used). As Luce and Krumhansl (1988) observed, “This orientation reflects the belief that differences between sensations can be detected, but that their absolute magnitudes are less well apprehended” (p. 39). Using what he called the “method of right and wrong cases,” Fechner (1860/1996) presented observers with two weights lifted in close succession, a fixed standard stimulus (e.g., 50 g in the example used above) and a comparison stimulus (e.g., 53 g) on many trials. Which weight was lifted first (standard or comparison) and in which hand (left or right) was counterbalanced, and the observer’s job was to decide which of the two weights felt heavier. Because of internal Gaussian error in the sensations generated by each weight, the lighter weight (Weight A) would sometimes be incorrectly judged as the heavier weight, just as would sometimes happen if imperfect balance scales were used to measure the two weights. The frequency with which such errors occurred is what Fechner used to scale sensations.

To appreciate the logic of Fechner’s approach, consider again physical measurements obtained using a balance scale (upper panel of Figure 1). Knowing that \( \overline{X}_A \) and \( \overline{X}_B \) are 1.5 standard deviations apart allows us to predict how often a physical measurement of Weight A would erroneously exceed that of Weight B. In formal terms, the question is this: over a large number of trials, how often would a random draw from the distribution of scores for Weight A (random variable A) exceed an independent random draw from the distribution of scores for Weight B (random variable B)? This question is equivalent to asking how often the random variable \( B - A \) would be negative, which can be answered by considering the distribution of difference scores. As illustrated in the lower panel of Figure 1, if the original distributions of balance-scale measurements of Weight A and Weight B are 1.5 standard deviations apart (top panel of Figure 1), then, across trials, one would expect that Weight A would be erroneously found to be heavier than Weight B 14.4% of the time (i.e., 85.6% correct). Interestingly, although \( s \) was set to 2 g in this example, the logic applies no matter what interval-scale units are used to measure \( s \). That fact is what made it possible for Fechner to scale subjective sensations in terms of standard deviation units, working backward from behavioral errors.

Figure 2 corresponds to Figure 1 except that it represents subjective sensations rather than physical measurements. On each trial, Weight A generates sensation \( \alpha \) and Weight B generates sensation \( \beta \). To emphasize the fact that these sensations are not directly measurable on an interval scale, lower-case Greek letters are used to represent them as well as their corresponding means and standard deviations. Otherwise, everything is the same. With regard to behavioral performance, imagine that all we know is that when observers judge the weights over many trials, Weight A is erroneously found to be heavier than Weight B 14.4% of the time. Starting at the bottom of Figure 2 (the distribution of difference scores) and working backward from that behavioral error measure, we could infer that \( \mu_B \) and \( \mu_A \) are separated by 1.5 standard deviation units. Today, we would say that \( d' = 1.5 \), but Fechner called his measure \( t \), which is equal to .5\( d' \). In other words, he measured the psychological distance from the mean of Weight A to the midpoint between the means of Weight A and Weight B (Link, 1994, 2001) such that \( t = 0.75 \). However, the two measures \( (d' \text{ and } t) \) convey exactly the same scaling information because they are linearly related to each other. Thus, as long ago as 1860, Fechner had already worked out the Gaussian-based 2AFC signal detection framework. It was an undeniably monumental and enduring advance, but there was still much more work to be done.

**Louis Leon Thurstone: Psychologically Scaling Physically Unmeasurable Stimuli**

Thurstone’s (1927) key insight was to appreciate that there are qualities of stimuli that can be scaled in psychological space in precisely the same way that Fechner scaled weights despite the fact that those qualities cannot be measured in physical space. To use one of Thurstone’s examples, consider two samples of handwriting, which are judged by an observer to differ in their physical beauty (even though it would be hard to physically measure how much they differ in that regard). The fact that one regards one sample as being more beautiful than the other already means they differ in psychological space. However, we’d like to specify the psychological distance between them on a unidimensional interval scale, which would allow this scaling exercise to be extended to many more handwriting samples, placing them all on an interval scale of beauty.

Like Fechner, Thurstone assumed that each handwriting sample generates an internal sensation of beauty that varies from trial to trial. Thurstone avoided the use of the term “sensation” to avoid implying anything about whether the internal signal was physical, mental, or some combination of the two, but I will continue to use that term here for consistency. For handwriting samples A and B, denote their respective mean sensations of beauty \( \mu_A \) and \( \mu_B \) and their respective standard deviations of beauty \( \sigma_A \) and \( \sigma_B \), with \( \mu_B > \mu_A \) (i.e., sample B is more beautiful than sample A). To scale the psychological distance between \( \mu_B \) and \( \mu_A \) we simplify by assuming that \( \sigma_A = \sigma_B = \sigma \), as we did earlier for psychophysical scaling example involving weights.
Thurstone presented both samples on each trial and asked subjects to select the more beautiful of the two. Thus, like Fechner, he used a 2AFC task. For the 2AFC task, each trial is conceptualized as a random draw from distribution B (yielding sensation $/H_9252/H_9251$/) and an independent random draw from distribution A (yielding sensation $/H_9251$/). These random variables, $/H_9252$/ and $/H_9251/$, represent how beautiful handwriting samples B and A seem, respectively, on a particular trial. Across many trials, a new distribution of difference scores ($/H_9252/H_9251$/) will be created. As before, this distribution of difference scores (illustrated in the lower panel of Figure 2) has a mean equal to $/H_9262/H_9262$/ and a standard deviation equal to $/H_9268/H_9268$/.

The scaling metric that Thurstone used is known as the Law of Comparative Judgment, the general equation for which is usually expressed in the following form (e.g., Luce, 1994, Equation 1; Thurstone, 1927, Equation 2):

$$/H_9262/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268$/ $/H_9268}$/


Using Thurstone scaling, this is how far apart the two handwriting samples are on an interval scale in psychological space. For this example, we find that $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262$/ $/H_9262}$/.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
example, the result from Equation 3 corresponds to what is depicted in the upper panel of Figure 2, which is to say that the two samples are 1.5 standard deviations apart.

Although it is not obvious given how Fechner (1860/1966) and Thurstone (1927) presented their mathematical derivations, it should now be clear that they used the exact same scaling approach, as illustrated in Figures 1 and 2. Moreover, although both used the 2AFC task, the \( d' \) score they derived, as illustrated in the upper panel of Figure 2, is the value that theoretically corresponds to a yes/no task version of the task in which only one weight or one handwriting sample was presented on each trial. For example, on each trial, the question would be “Is this the heavier of the two weights (yes or no)?” or “Is this the more beautiful of the two handwriting samples (yes or no)?” Using this task, one could compute \( d' \) from the hit rate (proportion of trials involving stimulus B correctly judged as such) and the false alarm rate (proportion of trials involving stimulus A incorrectly judged as being stimulus B). The computational formula is \( d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR) \), where \( HR \) and \( FAR \) are the hit and false alarm rates, respectively (Macmillan & Creelman, 2005, p. 369, Equation 1.5). Theoretically, at least, we would find that \( d' = 1.5 \) using this approach as well (see Appendix A for a discussion of the fact that discriminability on the 2AFC task exceeds discriminability on the yes/no task, which is what modern-day signal detection theorists would likely emphasize).

A special version of the yes/no detection task involves making a decision on each trial about the presence or absence of a stimulus (instead of making a decision about which of two possible stimuli was presented). Although neither Fechner nor Thurstone addressed the issue, the signal detection model for this task consists of a signal distribution representing sensations on stimulus-present trials and a noise distribution representing sensations on stimulus-absent trials. For the types of stimuli that Thurstone considered, the concept of a noise distribution (i.e., the distribution of sensations on stimulus-absent trials) does not even come into play because the task cannot be broken down into stimulus-present trials and a noise distribution representing sensations on stimulus-absent trials. For example, the result from Equation 3 corresponds to what is depicted in the upper panel of Figure 2, which is to say that the two samples are 1.5 standard deviations apart.

The Noise Distribution and Its Relationship to Conscious Awareness (1860–1953)

Although never specifically referring to or illustrating a noise distribution, per se, Fechner (1860/1966) clearly conceptualized sensations arising from “noise” in terms of spontaneous neural activity. According to this idea, sensory neurons are not completely quiescent in the absence of physical stimulation. Presumably, though not specifically mentioned by Fechner, spontaneous activity varies from trial to trial, thereby giving rise to what might be regarded as a distribution of noise activity across trials. The issue under consideration now is how to conceptualize the placement of that noise distribution in relation to a threshold of conscious awareness. To illustrate the key concepts, I return to a consideration of physically measurable stimuli (unlike what Thurstone considered) and switch from using examples involving sensations of heaviness to examples involving auditory or visual sensations (e.g., loudness or brightness).

Neural Noise Below the Threshold of Conscious Awareness

Figure 3 depicts a discrimination task similar to that illustrated in Figure 2 (now for sound) except that I have depicted a hypothetical boundary—a threshold—of conscious awareness. We might think of this threshold (\( T \)) as corresponding to the amount of activity in auditory neurons required for experiential sensation to occur, which is the amount of neural activity generated by a stimulus with physical intensity \( I_s \) in Fechner’s law. In Figure 3, the noise distribution is situated below the boundary of conscious awareness, which means that although spontaneous noise is occurring in sensory pathways of the neural system, the observer has no conscious awareness of it. The distributions lie on a dimension labeled “sensation of loudness,” but a more complete label might be “degree of neural activity in auditory sensory channels that give rise to the sensation of loudness.” As depicted in Figure 3, a certain threshold level of neural activity is required before any conscious sensation of loudness will occur.

Placing the noise distribution below the boundary of conscious awareness serves to illustrate the intriguing way that Fechner thought about how neural noise interacts with the neural activity generated by a physical stimulus to give rise to the conscious experience of sensation. He referred to these below-threshold sensations as “negative sensations,” a description that corresponds to the fact that his famous psychophysical function returns a negative value whenever \( I < I_s \) (Equation 2). In some ways, specifying a noise distribution that falls below the threshold of conscious awareness seems nonsensical, and Fechner received much grief from critics for proposing it. What, exactly, is a negative sensation? As noted by Murray (1990), Fechner wanted it to be clearly understood that by “negative” he did not mean “the opposite.” Thus, for example, in an experiment on the sensation of warmth, neural noise did not theoretically give rise to subthreshold levels of coldness. Instead, negative sensations were more like imaginary numbers, but his critics would have none of it. As Fechner put it:

It is incontestably to be desired that the controversy over negative sensations should come to an end once and for all; but if my experience up to now is any guide, even my ghost will have no peace because of it. (Murray, 1990, p. 58)

He was apparently right about that. What Fechner did not realize was that the solution to his negative-sensation problem was the one that signal detection theorists would provide nearly a century later, namely, the noise distribution falls above, not below, the threshold of conscious awareness (i.e., neural noise generates genuinely felt
positive sensations, not imaginary negative sensations). Remark-
ably, Fechner seemed close to arriving at that conclusion himself.

Fechner’s Conceptualization of Neural Noise

In Fechner’s view, a distinction can be drawn between two
stimulus-intensity thresholds (Murray, 1993). One is \( I_0 \), which, as
noted earlier, represents the stimulus intensity required to achieve
conscious awareness (often referred to as the absolute threshold).
The other is \( I_0' \), which, as used here, represents the stimulus
intensity required to induce a change in neural activity in the
relevant sensory channel. Stimulation greater than \( I_0' \) but less than
\( I_0 \) adds to the baseline neural activity, moving the mean of the
unconscious neural noise distribution to the right in Figure 3
without necessarily moving it into the realm of conscious aware-
ness. Eventually, however, as stimulus intensity increases still
further, the right tail of the signal-plus-noise distribution would
begin to exceed the threshold, creating conscious sensations on
some nontrivial proportion of the trials. Viewed in that light, what
we refer to as \( I_0' \) is not precisely definable. A common but arbitrary
definition is that \( I_0' \) is equal to the intensity required to produce
conscious awareness of the stimulus 50% of the time. In other
words, as interpreted here, \( I_0' \) is the intensity required such that the
mean of the distribution of sensations falls at the threshold of
conscious awareness.

In the auditory domain, Fechner viewed neural noise as usually
falling below the threshold of conscious awareness. However, he
did not believe that the auditory noise distribution always falls
below the threshold of conscious awareness, noting, for example,
that “In abnormal conditions, the ear may be subjected to internal
excitation which exceeds its threshold” (Fechner, 1860/1966, p.
210). Moreover, in the visual domain, he argued that the noise
distribution typically falls above the threshold: “as we have noted
several times, the eye is always above threshold because of its
internal excitation, so that each external light stimulus can only
add to the excitation already present” (Fechner, 1860/1966, p.
200). Here, he is clearly describing the modern view of signal
detection theory, where the visual noise distribution falls above the
threshold of conscious awareness (he referred to it using a term
translated as “dark light”), and the signal distribution consists of
the neural signal generated by a visual stimulus added to the
above-threshold neural noise that was already there (Scheerer,
1987).

According to Murray (1993; see also Nicolas, Murray, & Farah-
mand, 1997), at about the same time that Fechner was contem-
plating this issue, von Helmholtz (1962) similarly argued that any
sensation of brightness is added to the “natural light of the retina
(i.e., the resting state of activity). This concept seems equivalent to
Fechner’s concept of dark light. Possibly in response to this idea,
Delboeuf (1873) suggested that Weber’s Law might be properly
conceptualized as a differential equation of the form
\( dS = k(dI/(I + I_n)) \), where \( I_n \) is the intensity of the resting state neural
activity. Integrating both sides of this equation and following the
same steps as before yields the following variant of Fechner’s
Law:

\[
S = \lambda \log \frac{I + I_n}{I_n}
\]

As in Fechner’s Law, when \( I = 0, S = 0 \). Critically, however,
according to a seemingly straightforward interpretation of this
formulation, \( zero \ does \ not \ mean \ the \ absence \ of \ sensation \); instead,
zero corresponds to the average nonzero sensory experience asso-

Figure 3. An illustration of the sensations generated by two sounds that differ in their intensities (Sound B is
louder than Sound A). Each sound, when presented alone across many trials, yields a distribution of sensations
of loudness, and the mean loudness for Sound B exceeds that for Sound A. Critically, according to this model,
the distribution of sensations generated by spontaneous neural activity in auditory pathways falls below the
threshold of conscious awareness (T). In other words, noise activity generates no conscious sensation whatso-
ever.
associated with resting state neural activity. Analogously, 0°C is an arbitrarily defined point on an interval temperature scale, not the minimum possible temperature, so negative values are interpretable as positive values.

As far as I can determine, Fechner, though aware of the above formulation by Delboeuf (Murray, 1993), did not pursue this line of thinking any further (at least not in his first volume; Fechner’s second volume has not yet been translated into English, incredibly). Had he done so, he might have conceived of the rest of modern-day signal detection theory. Alas, it was not until the early 1950s that the notion of an above-threshold noise distribution was born again, during an era when (a) researchers were trying to measure the stimulus intensity required to achieve conscious awareness of a stimulus and (b) statisticians were, quite independently, working out the modern view of statistics involving a null hypothesis (“noise”) versus an alternative hypothesis (“signal”). I first consider early efforts to nail down the visual threshold.

In Search of the Absolute Visual Threshold

Although Fechner considered noise in the visual system to always fall above threshold, researchers in the 1940s instead concluded that neural noise was essentially nonexistent. In a classic experiment, Hecht, Shlaer, and Pirenne (1942) set out to measure the exact stimulus intensity required for an observer to achieve conscious awareness of a flash of light. Subjects in this experiment were dark adapted for 40 min and then asked to fixate on a dim red target. On each test trial, they pressed a button to deliver a 100 ms flash of light in the periphery of their visual field, with the intensity of the flash varying randomly across trials.

A notable feature of this experiment—one that remains oddly common even today—reflects the researchers’ implicit belief that any noise distribution that might exist falls well below the threshold of conscious awareness. The notable feature is that only stimulus-present trials were used. Never did the participant press the button to deliver a flash and then report whether or not one was detected when no flash occurred at all (i.e., there were no stimulus-absent trials). Instead, a flash always occurred but with an intensity that varied from trial to trial. Thus, although the hit rate was measured, the concept of a false alarm was nowhere to be found.

In 1942, this design flaw was certainly forgivable, but the same flaw can be found even today and seems somewhat less forgivable than it once was. As noted by Swets (1961) more than a half century ago, experimenters sometimes include a few stimulus-absent trials (with error feedback) that they construe as catch trials. These trials are designed to keep subjects from randomly guessing yes in the complete absence of sensation, but what if they genuinely experience false sensations on those trials, contrary to the experimenter’s threshold theory?

The results reported by Hecht et al. (1942) showed that the probability of detecting a flash increased from close to 0 at the lowest intensity to nearly 1.0 at the highest intensity. In other words, a standard psychophysical function was obtained. Figure 4 illustrates the signal-detection interpretation of a standard psychophysical task using the common detection threshold of 50% (the function in the lower panel is an idealized depiction of the path along which empirical data typically fall). On the surface, the fact that the flash with the lowest intensity was rarely detected in the data reported by Hecht et al. (1942) seems to support the idea that noise in the visual system rarely, if ever, creates the illusion that a flash did in fact occur. However, such reasoning provides false comfort. All it really means is that the subjects likely set a conservative decision criterion for declaring that they detected a flash, preferring that the flash appear relatively bright to them before declaring that it was detected.

In an early use of ideal observer analysis (a topic I consider in more detail later), Hecht et al. (1942) also fit a Poisson model to the psychophysical data from each participant to estimate the number of quanta absorbed by the retina in response to a flash of light (which varies trial to trial even holding stimulus intensity constant). The fits were very good and indicated that the probability of consciously detecting a flash was almost perfectly predicted by the estimated probability that approximately six quanta of light or more were absorbed. The higher that estimated probability, the higher the behavioral probability of detecting the flash. The fact that the ideal observer model fit the psychophysical data almost perfectly was interpreted to mean that, near the stimulus threshold, variability in perception arises almost exclusively from stimulus variability (i.e., variability in the number of quanta that happen to be absorbed by the retina across trials). This interpretation is in stark contrast to the modern interpretation, according to which variability in sensation reflects any stimulus variability that might exist plus variation in neural noise to which the signal is

Figure 4. Upper panel: distributions of sensations generated by flashes of light across six levels of physical intensity (a noise distribution for stimulus-absent trials, if such trials were included, would fall to the left of the low-intensity distribution). The low-intensity flash almost never generates sensations that fall above the threshold of conscious awareness (T). Lower panel: psychophysical function relating the physical intensity of a flash to the percentage of flashes that are detected (i.e., that receive a yes response). In this example, the threshold is arbitrarily defined as the intensity associated with a 50% detection rate.
null hypothesis testing and the Neyman and Pearson (1933) lemma. As most students of psychology know, Fisher’s (1925) approach to hypothesis testing consists of evaluating the probability of obtaining a particular empirical result (e.g., a t-value of 1.86) by chance even if no effect is present. The key idea is that even when no signal is present, so to speak, a distribution of empirical t scores will still be obtained across many identical experiments. This distribution, though having nothing at all to do with conscious sensations, would have a mean of zero and is exactly analogous to what would later be conceptualized as the noise distribution in signal detection theory. Traditionally, a statistical test is judged to be significant if the obtained t score exceeds a criterion such that it would have been observed by chance—that is, given that the null hypothesis is true—less than 5% of the time (with α set to .05). However, although the basic null-hypothesis testing approach involves the equivalent of the noise distribution of signal detection theory, there is nothing that corresponds to the signal distribution.

Neyman and Pearson (1933) added the equivalent of the signal distribution when they sought to optimize a binary decision about an alternative hypothesis (Hₐ) versus the null hypothesis (H₀) in light of empirical data. To do so, they proposed specifying an expected effect size prior to running the experiment. With a prespecified effect size and a prespecified alpha level, one can compute the fixed sample size (N) required to have adequate power to detect the alternative hypothesis, if it is true. If β is the probability of failing to detect a true effect after testing N subjects, then power is equal to 1 − β, which is the probability of correctly detecting an effect, if it exists. In the parlance of modern-day signal detection theory, 1 − β is the scientific hit rate and α is the scientific false alarm rate.

Neyman and Pearson (1933) argued that an optimal decision maker would base a statistical decision ("signal" vs. "noise") on the likelihood of the evidence given Hₐ divided by the likelihood of the evidence given H₀ (i.e., a likelihood ratio test). According to the Neyman-Pearson lemma, the optimal decision rule involves choosing a criterion likelihood ratio that maximizes the probability of detecting Hₐ when it is true while ensuring that the probability of false alarm (i.e., the probability of detecting H₀ when it is false) is less than or equal to α. In a way, Neyman and Pearson (1933) proposed signal detection theory from a statistical point of view.

Gigerenzer and Murray (1987) argued that this new approach to statistics facilitated productive lines of thinking in domains beyond the field of statistical decision theory. Most notably, Peterson and Birdsall (1953) applied Neyman and Pearson’s (1933) reasoning to the detection of pulsed signals in noise, giving rise to modern-day signal detection theory. Peterson and Birdsall (1953) also realized, apparently for the first time, that the performance of any “receiver” can be efficiently summarized in a concise graph—the receiver operating characteristic (ROC) – by varying the criterion for detecting the signal (i.e., by varying the equivalent of the alpha

2 A reviewer drew my attention to another technical report by Marcum (1947), which clearly describes signal detection theory on page 9 (“Detection of a signal is said to occur whenever the output of the receiver exceeds a certain predetermined value hereafter called the bias level. In the absence of any signal, this bias level will on occasion be exceeded by the noise alone”). Marcum (1947) also describes manipulating the criterion, simultaneously increasing or decreasing true and false detections of a radar signal. This is the essential concept of ROC analysis, but no ROC was actually plotted.

3 Note that, from here on, I use α and β as they are commonly used in statistics (unlike in Figure 2, where they represent psychological random variables).
level).\(^4\) If the alpha level is set to a liberal value (e.g., $\alpha = .20$), many signals will be significant, so the hit rate and the false alarm rate will both be high. By contrast, if the alpha level is set to a low (conservative) value (e.g., if $\alpha$ is reduced from .05 to .005), few signals will be significant, so the hit rate and the false alarm rate will both be low. Thus, holding the quality of the detection device constant, one can achieve a whole range of hit and false alarm rates that can be plotted to reveal the ROC (an analytical approach that I consider in more detail in a later section).

**Sequential sampling.** Interestingly, whereas Neyman and Pearson’s (1933) statistical decision theory gave rise to signal detection theory, a different statistical decision theory advocated by Wald (1945; Wald & Wolfowitz, 1948) gave rise to modern-day sequential sampling models (Brown & Heathcote, 2008; Busemeyer & Townsend, 1993; Link, 1975, 1992; Ratcliff, 1978; Ratcliff & Murdock, 1976; Ratcliff & Smith, 2004; Usher & McClelland, 2001; Vickers, 1970). In Wald’s approach, instead of performing a statistical test following a predetermined (fixed) set of trials, a likelihood ratio test is performed after each trial, and testing terminates when a criterion value is achieved. Using that approach, one can most efficiently distinguish between two competing hypotheses (i.e., using the minimum number of observations) for a fixed error rate.

In psychology, both modeling approaches—signal detection models and sequential sampling models—are thriving today. As a general rule, they exist side by side, with signal detection models often used to interpret binary decisions made with a certain level of confidence, and sequential sampling models often used to interpret binary decisions made with a certain reaction time (RT). Indeed, a fascinating paradox is that signal detection theory is well-suited to conceptualizing confidence but not RTs, whereas the reverse is true of most sequential sampling models (Pleskac & Busemeyer, 2010).\(^5\) Later, I briefly consider sequential sampling accounts of detection and confidence, but, in what follows, I mostly concentrate on the evolution of signal detection theory—its interpretation of binary (e.g., yes/no) decision making as well as its interpretation of confidence—in the years following the publication of Peterson and Birdsall’s seminal technical report in 1953.

**Sensory Noise Above the Threshold of Conscious Awareness**

Later in 1953, another University of Michigan technical report (No. 18) entitled “A New Theory of Visual Detection” by Tanner and Swets (1953) applied signal detection theory to experiential sensation in the visual domain. At about the same time, Smith and Wilson (1953), from the University of Washington and the Massachusetts Institute of Technology, respectively, presented a (now standard) signal detection model for an auditory tone-detection task involving multiple observers. Their Figure 11 shows a Gaussian distribution representing subjective sensations that arise on blank (i.e., stimulus-absent) trials. The reference list in this paper makes no mention of any of the psychophysical work I have considered this far (instead, it, too, mainly cites the statistical literature), and it appears to be a separate development of the theory in the auditory domain (Swets, 1973).

The extent to which any of these developments were directly influenced by Fechner is not entirely clear. None of the relevant 1953 publications cited Fechner, suggesting that the Gaussian-based signal detection framework applied to psychophysics arose from a different line of thought rather than as an addition to Fechner’s ideas. Then again, in questioning the notion of a fixed threshold, Swets (1961) observed that “although Fechner started the study of sensory functions along lines we are now questioning, he also anticipated the present line of attack in both of its major aspects” (p. 169). The two major respects anticipated by Fechner involved the relevance of statistical decision theory and the notion that there is a grading of sensory excitation below the threshold (i.e., negative sensations).

In any case, the next year saw additional papers on the same topic (e.g., Munson & Karlin, 1954; Tanner & Swets, 1954). Though the paper by Munson and Karlin (1954) had little impact (according to ISI, it has been cited only 18 times, with the most recent one occurring in 1986), Tanner and Swets (1954) was quite influential and is still often cited today. In that paper, they followed up on the technical report they had published the year before and made the following profound arguments: (a) consciously accessible neural signals are inherently noisy in the absence of stimulation (the key insight), (b) the presentation of a stimulus yields a neural response consisting of signal plus noise (echoing, but not citing, Fechner), and (c) observers base a decision on the strength of neural activity by setting an adjustable decision criterion. If the neural activity exceeds that criterion, the decision is yes (a stimulus was subjectively detected on this trial); otherwise, the decision is no (a stimulus was not subjectively detected on this trial). Figure 5 illustrates this theory. A notable feature of this figure—one that is far more important than is generally appreciated—is that virtually the entire noise distribution falls above the threshold of conscious awareness. From this perspective, the threshold, though it exists, is a largely irrelevant consideration.

According to this new way of thinking, the critical boundary is not the threshold of conscious awareness; instead, the critical boundary is the placement of the decision criterion (c). Unlike the threshold of conscious awareness, which is a fixed boundary, the placement of the decision criterion is under the control of the observer. Encouraging the observer to adopt a conservative criterion (shifting it to the right) would result in a lower hit rate and a lower false alarm rate. This is essentially what catch trials do, ultimately creating the false impression that the entire noise distribution falls below the threshold of conscious awareness. The same is true of psychophysical studies involving stimulus-present trials in which stimulus magnitude is varied over a wide range. If many trials involve a strong stimulus, the subject can comfortably set a conservative criterion and still detect many signals. Under such conditions, trials involving no signal at all (i.e., stimulus absent trials) would never yield a detect decision, creating the impression that sensations associated with neural noise fall below the threshold of conscious awareness. However, instead of falling

---

\(^4\) In a footnote, Swets (1986) says “Theodore G. Birdsall, in the Electrical Engineering Department of The University of Michigan, first taught me about ROCs when he invented them” (p. 100).

\(^5\) Signal detection theory does provide a rough guide for conceptualizing RTs in that it is typically assumed that the farther a sensation falls from the decision criterion, the faster and more confident the decision will be. However, it makes no predictions about (for example) RT distributions and thus cannot account for them.
Barlow (1956) replicated Hecht et al. (1942) with one small search for the visual threshold discussed earlier. For example, reconsidering the absolute visual threshold, the noise distribution falls below the conservative criterion induced by the experimental procedure.

**Reconsidering the Absolute Visual Threshold**

By 1956, these ideas were having a profound influence on the search for the visual threshold discussed earlier. For example, Barlow (1956) replicated Hecht et al. (1942) with one small change: observers were encouraged to not only report when a flash was definitely seen (high confidence) but also when it was possible that a flash was seen (low confidence). False positives were not observed for high-confidence “seen” responses (not surprisingly given that high confidence corresponds to a conservative placement of the decision criterion), but they were observed for low-confidence “possible” responses. Moreover, the estimated threshold—arbitrarily defined as the physical stimulus intensity required for a flash to be detected 50% of the time—changed depending on whether or not possible responses were counted. A threshold of conscious awareness is not something that should change as a function of how certain the observer is that a flash occurred, yet it did change. Barlow interpreted these results to mean that internal noise can give rise to the (false) experiential sensation of a flash. Sakitt (1972) later reported similar results using a 7-point confidence scale. In that study, the false-positive rate (a measure that is completely ignored when only stimulus-present trials are used) clearly varied as a function of the confidence used to detect a flash.

As Swets (1961) pointed out, and as illustrated in Figure 5, signal detection theory does not deny the existence of a threshold. Signal detection theory simply asserts that when an observer is highly attuned to detecting an extremely weak sensory signal, spontaneous neural activity in the closely monitored sensory channel will occasionally produce genuinely felt (i.e., above-threshold experiential) sensations. Thus, the threshold is usually low enough that it is not a particularly important consideration. But to say so is not to say that there is no stimulus intensity so weak that it could never be detected. Of course a signal can be that weak. As indicated earlier, a hydrogen atom placed in one’s hand will generate no response in the nervous system and so cannot possibly be truly detected. The key point, however, is that an observer who is trying hard to detect an extremely weak signal in a sensory channel will not subjectively experience absolute nothingness on such trials. Instead, the observer will come into contact with neural noise and sometimes report that, yes, on this trial, a hydrogen atom was placed in my hand. This will happen because the observer experiences a genuinely felt (albeit false) sensation.

**Threshold Theory Post-1953**

The signal detection perspective introduced in the early 1950s was undeniably revolutionary, but not everyone was convinced. Standing against signal detection theory, even to this day, are models that maintain the assumption of a fixed threshold that divides conscious experience into discrete categories (as opposed to the continuum envisioned by signal detection theory). Indeed, several variants of threshold theory emerged post-1953 (see Kellen & Klauer, 2018, and Rotello, 2017, for reviews). It is important to consider these theories in some detail because, although the term threshold theory might sound somewhat technical, it could also be called “how you think right now (probably).” In other words, at least in my experience, most people are intuitive threshold theorists, usually subscribing instinctively to the simplest version of it known as high-threshold theory. Thus, in experimental psychology, the competition between signal detection theory and threshold theory may always be part of the landscape.

In this section, after briefly describing threshold models, I consider what they have to say about the experiential status of false alarms and what they imply about the dependent measure that should be used to gauge performance on yes/no signal detection tasks. In subsequent sections on ROC analysis and the confidence-accuracy relationship, I also consider what they predict (and what signal detection theory predicts) in those domains.

**Modern Variants of Threshold Theory**

**High-threshold theory.** The simplest and most intuitive version of threshold theory is known as high-threshold theory, which many (e.g., Link, 1992) attribute to an unpublished report by Blackwell (1953). The first published report mentioning this theory that I have been able to find is the paper by Smith and Wilson (1953) cited earlier. According to this theory, and in agreement with almost everyone’s intuition, stimulus-absent trials result in a below-threshold signal precisely because no stimulus was presented. After all, how can a stimulus be detected if it was not presented? On stimulus-present trials, by contrast, a signal will exceed the threshold with probability $p$ and fail to exceed the threshold with probability $1 - p$. The higher $p$ is, the better detection performance is. A critical assumption of threshold theory is that, on below-threshold trials, no matter how carefully an observer monitors a sensory channel for evidence that a stimulus was presented, no signal whatsoever in that channel is detected. Even so, under such conditions, observers sometimes randomly guess that a stimulus was presented, which is why false alarms sometimes occur. Critically, therefore, according to this theory, all false alarms (as well as some hits) are random guesses.

**Two-high-threshold theory.** Another theory known as two-high-threshold theory (or double high-threshold theory) adds to
High-threshold theory by assuming a second threshold that can only be exceeded on stimulus-absent trials (Green & Swets, 1966). On stimulus-present trials in which a sensation generated by a signal exceeds the signal threshold, the subject correctly responds yes; on stimulus-absent trials in which a sensation generated by noise exceeds the noise threshold, the subject correctly responds no. On stimulus-present and stimulus-absent trials in which the relevant threshold is not exceeded, there is no diagnostic sensory information upon which to base a decision, so the subject simply guesses yes or no. Like high-threshold theory, this theory therefore assumes that all false alarms (as well as some hits) are random guesses. Two-high-threshold theory has been advocated in recent years, mainly for memory tasks (e.g., Bröder & Schütz, 2009; Bröder, Kellen, Schütz, & Rohrmeier, 2013; Kellen & Klauer, 2011; Malmbarg, 2002; Province & Rouder, 2012).

Low-threshold theory. An interesting variant of threshold theory was proposed by Luce (1963). This theory is known as low-threshold theory (Luce, 1963), and it has enjoyed a recent resurgence (Kellen, Erdfelder, Malmbarg, Düb, & Criss, 2016; Starns, Dubé, & Freling, 2018). Low-threshold theory assumes that the threshold for conscious awareness is not placed entirely below the noise distribution (as in signal detection theory) or entirely above it (as in high-threshold theory). Thus, because some of the noise distribution falls above the threshold, low-threshold theory can handle a genuinely experienced false sensation. Like high-threshold theory, this theory assumes that subjects respond yes if the threshold is exceeded and guess yes with a certain probability if not.

The Subjective Experience of False Alarms

As noted earlier, many people have experienced the strong subjective sensation of their cell phones vibrating in response to an apparent text message that did not actually just arrive. Despite the absence of physical vibration, one may be absolutely certain that a text was just received because the experiential sensation of a vibrating phone was unmistakable. Yet, when the phone is checked, it becomes clear that it was a high-confidence false alarm. This phenomenon is so common that it has a name: phantom vibration syndrome (e.g., Rosenberger, 2015). High-threshold theory and two-high-threshold theory have no easy way to make sense of this almost universal experiential phenomenon, but low-threshold theory can explain it because that theory allows for the possibility that neural noise can exceed the detection threshold (Starns & Ma, 2018).

The same consideration applies to recognition memory, where instead of indicating whether or not a flash or a tone was detected, subjects indicate whether or not a test item (e.g., a word) appeared on an earlier list. On a recognition test, false alarms (i.e., yes responses to lures) appear to reflect a genuinely experienced false sense of prior occurrence, even when subjects claim to recollect the lure’s prior occurrence on the list. Dual-process models hold that recognition decisions are based either on a nonspecific sense of familiarity signal (as standard single-process models assume) or on the specific recollection of details. Such models typically assume that recollection is a high-threshold process (e.g., Diana, Reder, Arndt, & Park, 2006; Yonelinas, 1994, 2002). By using Tulving’s (1985) Remember-Know procedure, subjects can theoretically indicate which process supported their yes decision by saying Remember if it was based on a threshold recollection signal and saying Know if it was based on a continuous familiarity signal. Interestingly, in every study that has used the Remember–Know procedure, Remember false alarms occur. That is, subjects report the subjective impression of recollection (not the subjective impression of random guessing). Indeed, using the Deese-Roediger-McDermott procedure, Remember false alarms occur as often to critical lures as they do to targets (Roediger & McDermott, 1995, Experiment 2). Consistent with the idea that such reports reflect genuinely experienced false recollection, Remember false alarms are made more quickly and with higher confidence than Know false alarms (e.g., Duarte, Henson, & Graham, 2008; Wheeler & Buckner, 2004; Wiesmann & Ishai, 2008; Wixted & Stretch, 2004). Whereas signal detection theory and low-threshold theory naturally accommodate results like these, they are hard to reconcile with either high-threshold or two-high-threshold theory.

Measuring Discriminative Performance

Signal detection theory holds that the ability to discriminate two states of the world (e.g., Signal vs. Noise) is represented by the standardized distance between the Gaussian sensory distributions they generate. Today, that ubiquitous measure is known as $d'$, the computational formula for which is given by

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

This formula applies to the simplest case, namely, a yes/no detection task in which the variances of the signal and noise distributions are assumed to be equal.6

By contrast, as detailed in Appendix B, high-threshold theory gives rise to a measure of discriminative performance often referred to as the standard correction for guessing:

$$p = \frac{HR - FAR}{1 - FAR}$$

As also detailed in Appendix B, the simplest version of two-high-threshold theory instead holds that:

$$p = HR - FAR.$$  

Low-threshold theory does not yield a singular measure of discriminative performance (Macmillan & Creelman, 2005, p. 88). Equations 4–6 underscore a key point that is too often overlooked: whatever measure of performance an experimenter chooses to use on a yes/no detection task or a 2AFC discrimination task, that measure necessarily embraces specific theoretical assumptions. That is, no matter how the $HR$ and $FAR$ are combined to yield a dependent measure, and whether you know it or not, that measure embraces a specific theory. Although every experimenter should know this, in my experience, many do not, so they end up unwittingly using a measure that embraces a theory that makes a lot less sense than signal detection theory.

If the objective is to measure underlying psychological processes, which was Fechner’s objective at the dawn of experimental psychology, then there is no escape from this reality, no matter how much one would like to use a theory-free measure of detection.

---

5 For the yes/no task, the formula can be written as: $d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR)$. For a 2AFC task, $d' = [\Phi^{-1}(HR) - \Phi^{-1}(FAR)]\sqrt{2}$ (see Appendix A).
performance (Swets, 1986). Every measure of the mind embraces specific theoretical assumptions, including the supposedly atheoretical, nonparametric $A'$ (Pollack & Norman, 1964; Snodgrass & Corwin, 1988). $A'$ is widely used, even today, because it is supposedly nonparametric and theory-free, but the truth is that it embraces unreasonable theoretical assumptions that no researcher would consciously embrace (Macmillan & Creelman, 1996; Pastore, Crawley, Berens, & Skelly, 2003). When measuring the mind, there is no choice but to embrace a theory, so the only rational course of action is to choose your measure—and, therefore, your theory—wisely.

Ironically, in my experience, some researchers are reluctant to embrace the assumptions of signal detection theory because it assumes Gaussian distributions, so they choose a measure like $A'$ or proportion correct, oblivious to the much more implausible theory they just embraced instead (Macmillan & Creelman, 1996). Even more ironically, as noted by Pastore et al. (2003), some of these same researchers embrace the Gaussian assumption with alacrity when performing a statistical analysis, such as ANOVA. Ask yourself: is there a distribution of internal sensations that is more plausible than the one Fechner chose to use in 1860? If so, use it. Signal detection theory does not require the Gaussian assumption, but it does require that you choose (and defend) the distributional form that you believe to be more plausible than that of the Gaussian. Green and Swets (1966) argued that it is generally assumed that sensory events are composed of a multitude of largely independent neural events. If so, the central limit theorem justifies the Gaussian assumption with alacrity when performing a statistical analysis, such as ANOVA. Ask yourself: are there a distribution of internal sensations that is more plausible than the one Fechner chose to use in 1860? If so, use it. Signal detection theory does not require the Gaussian assumption, but it does require that you choose (and defend) the distributional form that you believe to be more plausible than that of the Gaussian. Green and Swets (1966) argued that it is generally assumed that sensory events are composed of a multitude of largely independent neural events. If so, the central limit theorem justifies the assumption of a Gaussian distribution of individual neural events. One can never be sure about the form of the underlying distribution, of course, but the rationale offered by Green and Swets (1966) seems infinitely preferable to the presence of any rationale for embracing the peculiar theoretical assumptions implicitly endorsed by a measure like $A'$.

Receiver Operating Characteristic Analysis

Although unknown to the world prior to 1930, an ROC is nothing more than a plot of the hit rate versus the false alarm rate across different levels of response bias, holding discriminability constant. Its utter simplicity belies its profound importance. Indeed, since its inception in 1953, the impact of ROC analysis is so biostate. Initially, and continuing to this day, it was used to test signal detection versus threshold theories of yes/no detection performance. Later, beginning in the 1970s, it emerged as the state-of-the-art technique in a large number of applied domains, including (a) medicine, where it has become the primary method of evaluating competing diagnostic tests (e.g., Metz, 1978, 1986), (b) machine learning, where it is used to evaluate competing pattern-recognition algorithms (e.g., Fawcett, 2006), (c) weather forecasting, where it is used to test competing weather prediction models (Marzban, 2004), and (d) most recently, eyewitness identification, where it is used to test competing lineup formats (Wixted & Mickes, 2010). Next, I consider (a) methods for generating ROC data, (b) how ROC analysis is used to test theoretical predictions, and (c) how it is used to address applied issues, with an emphasis on eyewitness identification.

Generating ROC Data

Several methods are used to induce subjects to shift the decision criterion across conditions while (hopefully) holding discriminability constant (Swets, Tanner, & Birdsall, 1955; Tanner, Swets, & Green, 1956). For example, prior to presenting the test trials, instructions can be used to encourage subjects to adopt either a conservative criterion or a liberal criterion (instruction method). In the conservative condition, the instructions might be: “Please do not indicate that you detect the signal unless you are sure that it was presented,” whereas in the liberal condition, the instructions might instead be: “Please indicate that you detect the signal even if you are not sure it was actually presented.” Alternatively, the probability of presenting a stimulus-present trial versus a stimulus-absent trial can be varied across conditions (signal presentation probability method). In the conservative condition, subjects would be told that the majority of test trials will be stimulus-absent trials, whereas in the liberal condition, they would be told that the majority of test trials will involve stimulus-present trials. Still a third approach would be to manipulate decision payoffs, differentially rewarding correct rejections to induce a conservative criterion or differentially rewarding hits to induce a liberal criterion (payoff method). Regardless of the method used, each biasing condition would have a different $HR$ and $FAR$, but they would (ideally) all reflect the same ability to detect the stimulus because stimulus magnitude is held constant.

What different biasing conditions do at a psychological level differs according to the theory used to interpret the data. According to high-threshold theory, for example, $p$ would theoretically remain constant across biasing conditions, but the rate of guessing yes in the below-threshold state would vary across conditions. If $p$ is assumed to remain constant across biasing conditions, then we can use Equation 5 to predict the relationship between $HR$ and $FAR$ when the various points are plotted on the ROC. Specifically, solving for $HR$ in Equation 5 yields $HR = FAR + (1 - FAR)p$. Thus, according to high-threshold theory, holding $p$ constant (i.e., for a fixed stimulus magnitude), $HR$ should be a linear function of $FAR$ across different levels of response bias. By contrast, and as illustrated in Figure 6, signal detection theory assumes that different biasing conditions yield different placement of the decision criterion (not different probabilities of pure guessing). It also illustrates the fact that the predicted ROC is curvilinear, not linear.

The methods discussed above yield binary ROC data because each condition yields a hit rate and a false alarm rate based on binary yes/no decisions. One concern about these methods is that $d'$ (or $p$) might not remain constant across conditions, thereby distorting the shape of the ROC. That is, the biasing manipulation might unintentionally affect $d'$ or $p$ as well as response bias (see, e.g., Mickes et al., 2017). In light of that possibility, a simpler and arguably better way to generate ROC data is to collect confidence ratings in a single condition involving a neutral response bias (the rating method). As illustrated in Figure 7, decisions made with varying levels of confidence can be conceptualized in exactly the

---

7 The simplest version of double-high-threshold theory similarly predicts a linear ROC of the form $HR = p + FAR$ (Appendix B). Low-threshold theory predicts a bitonic ROC (two linear segments with different slopes) that can approximate a curvilinear ROC (see Macmillan & Creelman, 2005, p. 86).
same was as decisions based on criteria ranging from liberal to conservative. The question of whether the ROC generated by one or more of the above methods is linear (in accordance with several threshold theories) or curvilinear (in accordance with signal detection theory) has been investigated in several areas of experimental psychology, including perception, memory, and judgment and decision-making.

ROC Analysis as a Test of Theory

Perception. It has been known for many years that, in the perception literature, ROCs are almost always curvilinear (e.g., Dubé & Rotello, 2012; Tanner & Swets, 1954). This is true for both binary ROCs (created by manipulating response bias across conditions) and confidence-based ROCs generated using the rating method. Thus, the data pose a clear problem for high-threshold theory and two-high-threshold theory, thereby discouraging the use of any measure of performance derived from them (such as Equation 5 or Equation 6).

Recognition memory (including visual working memory). In studies of recognition memory and visual working memory, confidence-based ROCs are also invariably curvilinear (e.g., Banks, 1970; Egan, 1958; Lockhart & Murdock, 1970; Wilken & Ma, 2004; Xie & Zhang, 2017a, 2017b). However, binary ROCs obtained using the signal presentation probability method are sometimes linear (e.g., Bröder & Schütz, 2009; Donkin, Tran, & Nosofsky, 2014; Rhodes, Cowan, Hardman, & Logie, 2018; Rouder et al., 2008), in accordance with the threshold view, and sometimes curvilinear, in accordance with signal detection theory (Donkin, Kary, Tahir, & Taylor, 2016; Dubé & Rotello, 2012; Dubé, Starns, Rotello, & Ratcliff, 2012).

Theoretically, confidence-based ROCs and binary ROCs should agree. However, when the signal presentation probability method is used to create a binary ROC, the points tend to be noisy and not very spread out, making it difficult to convincingly argue for one shape or the other (e.g., Dubé et al., 2011). Moreover, contrary to what is supposed to happen, when this method is used, subjects sometimes become paradoxically biased to choose the stimulus presented less often (e.g., Johnstone & Alsop, 1996; Tanner, Haller, & Atkinson, 1967; Tanner, Rauk, & Atkinson, 1970; see also Levari et al., 2018). Sequential effects (e.g., a tendency to repeat the last response) can be surprisingly strong using this method (Tanner et al., 1967, 1970), though such effects can be a problem no matter how ROC data are collected. Thus, the effect of this method on behavior—and, by extension, its possible distorting effect on the shape of the ROC—is not very well understood.

Against this view, two-high-threshold theorists have argued that binary ROCs are preferable because both signal detection theory

---

8 An exception is eyewitness identification ROCs because only one response is collected from each subject, in which case sequential dependencies are an irrelevant consideration.
they are and why subjects often report the experiential sensation of a true memory when they make a false alarm. To accommodate curvilinear ROCs, threshold models explain away confidence ratings instead of offering any theoretical understanding of them. Yet understanding confidence—for example, in the context of eyewitness memory—is of paramount importance. In sidestepping this key issue to accommodate curvilinear ROC data, my own view is that two-high-threshold theory is reduced to being little more than a math-modeling exercise. In that sense, curvilinear confidence-based ROC data are theoretically diagnostic in that they separate psychologically viable models that are useful in helping understand real-world decision-making from models that are psychologically less viable and much less useful in that regard (because of the arbitrary assumptions they need to fit curvilinear ROCs). I return to this important issue in a later section concerned with the confidence–accuracy relationship.

Recollection. ROC analysis has also been used to investigate the nature of the recollection process in dual-process models of recognition memory, according to which yes decisions are based on recollection or familiarity. As noted earlier, using the Remember–Know procedure, it is theoretically possible to isolate recollection-based hits (namely, hits associated with Remember judgments). If recollection is a high-threshold process (Yonelinas, 1994), then a confidence-based ROC using only Remember hits and false alarms should be linear, but if recollection is a continuous signal detection process (Wixted, 2007; Wixted & Mickes, 2010), it should be curvilinear instead. Rotello, Macmillan, Reeder, and Wong (2005) and Slotnick (2005) both reported that recollection-based ROCs—that is, ROCs computed using only Remember responses made with different levels of confidence—are curvilinear, not linear, as a signal detection theory of recollection predicts. This finding fits with the evidence reviewed earlier suggesting that Remember false alarms reflect genuinely experienced false recollections, not random guesses (see related work by Didi-Barnea, Peremen, & Goshen-Gottstein, 2016; Dunn, 2004; Johnson, McDuff, Rugg, & Norman, 2009; Mickes, Wais, & Wixted, 2009; Moran & Goshen-Gottstein, 2015; Slotnick & Dodson, 2005; Stans, Rotello, & Hautus, 2014; White & Poldrack, 2013).

Judgment and decision making. In syllogistic reasoning, the “belief bias effect” refers to the tendency to accept or reject a conclusion based on its consistency with common beliefs, regardless of its logical status. A longstanding view is that, because of this bias, people have a hard time telling the difference between logically true and logically false claims that both happen to be intuitively believable. The analyses of data supporting this perspective implicitly adopted a two-high-threshold view, effectively relying on HR – FAR as the dependent measure in analyses of variance. However, Dubé, Rotello, and Heit (2010, 2011) reported that both confidence-based and binary ROC are curvilinear, not linear. That being the case, the relevant data should not be analyzed using the intuitively appealing HR – FAR measure. After performing ROC analysis and finding that the curves from the believable and unbelievable conditions were essentially indistinguishable, Dubé et al. (2010, 2011) argued that the belief bias effect is “aptly named” (i.e., believability affects response bias, not the ability to discriminate logically true and logically false conclusions). The issue continues to be debated, with ROC analysis remaining central to the efforts to distinguish between competing...
theories (e.g., Klauser & Kellen, 2011; Stephens, Dunn, & Hayes, 2019; Trippas, Handley, & Verde, 2013).

**ROC Analysis in the Applied Domain**

In contrast to the theory-based research described above, when ROC analysis is brought to bear on an applied question, the precise shape of the ROC and which theory-based measure is the most appropriate (e.g., Equation 4 or Equation 5) are not of particular interest (Wixted & Mickes, 2018). Consider, for example, evaluating which of two diagnostic tests more effectively discriminates those who have a disease and those who do not. The relevant applied question is not which theory distinguishes the two groups in terms of an underlying latent variable measured, for example, by $d'$ (cf. Goodenough, Metz, & Lusted, 1973); instead, the question is which test can *empirically* achieve the highest HR while at the same time achieving a lower FAR compared with the other test across the full range of response bias. Except when the two ROCs being compared cross over (which is rare), the answer to that question is provided by the area under the curve (AUC). AUC is a purely geometric measure that is tied to no theory of an underlying (latent) variable. If the ROC data fall along the diagonal line of chance performance, then AUC = .5. If the data instead fall on the y axis and upper x axis (ideally at the upper left corner where those axes intersect) indicating perfect discriminability, then AUC = 1. Generally speaking, AUC will fall between .5 and 1.0, and the procedure that yields the higher AUC is the one that can achieve both a higher HR and a lower FAR than the competing procedure. Thus, AUC answers the applied question. Note that $d'$ and AUC will almost always agree, but they can disagree (Wixted & Mickes, 2018), and when they do, AUC is the relevant measure for applied purposes.

A recent example from eyewitness identification illustrates the use of ROC analysis to address an applied issue. A problem that the field has worked on for many years is the fact that a high-proportion of wrongful convictions have been attributed to eyewitness misidentifications. In an effort to reduce that problem, researchers attempted to improve upon the ubiquitous six-pack photo lineup that the police often use when they investigate crimes that someone witnessed. A photo lineup consists of one picture of the suspect’s face (the person the police believe may have committed the crime) plus five pictures of physically similar fillers who are known to be innocent. The witness can pick the suspect, pick a filler, or reject the lineup. If the witness picks the suspect, investigators become more confident that they have identified the perpetrator of the crime. All too often, the identified suspect is innocent, and that is the problem applied researchers set out to address.

An alternative lineup procedure introduced by researchers in 1985 involves presenting the photos sequentially (one at a time), with each face receiving a yes or no decision (Lindsay & Wells, 1985). Typically, the procedure terminates following the first ID (i.e., following the first yes decision). Compared with the simultaneous procedure, the sequential procedure was found to result in lower hit rate and a lower false alarm rate, where the hit rate is the proportion of target-present lineups in which the guilty suspect is correctly identified, and the false alarm rate is the proportion of target-absent lineups in which the innocent suspect is incorrectly identified (filler IDs from both lineup types are another kind of false alarm, but they are relatively inconsequential because fillers are known to be innocent).

On the surface, the fact that both the hit rate and the false alarm rate are lower for the sequential procedure seems to indicate that it merely induces a more conservative response bias. However, a measure known as the diagnosticity ratio (DR), where $\text{DR} = \frac{\text{HR}}{\text{FAR}}$, suggested that the sequential lineup is also the more accurate eyewitness identification procedure. Often, the DR (based on a single $\text{HR}–\text{FAR}$ pair) was higher for the sequential lineup compared with the simultaneous lineup. In light of such findings, approximately 30% of the more than 17,000 law enforcement agencies in the United States adopted the sequential procedure (Police Executive Research Forum, 2013). In terms of real-world impact, this work would probably appear on any short list of the most influential research in the history of experimental psychology.

As noted by Rotello and Chen (2016) and Rotello, Heit, and Dubé (2015), for the DR measure to be viable, across changes in response bias, it would have to assume that $\text{HR} = k \times \text{FAR}$ (a straight line passing through the origin) such that the diagnosticity ratio, $\frac{\text{HR}}{\text{FAR}}$, would be a constant equal to $k$. That is, if it were a valid measure of discriminability, the diagnosticity ratio would not change as a function of response bias. However, the empirical data weigh against this prediction because lineup ROCs are invariably curvilinear, even when binary ROCs are plotted (e.g., Mickes et al., 2017; Rotello et al., 2015).

The point is that the DR computed from a single $\text{HR}–\text{FAR}$ pair cannot identify the diagnostically superior lineup procedure. Instead, the question of interest is which procedure yields the higher AUC computed from a range of $\text{HR}–\text{FAR}$ pairs. When simultaneous and sequential lineups were finally compared using ROC analysis (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes, Flowe, & Wixted, 2012), it was immediately apparent that the truth may be the opposite of what the field believed to be true for nearly 25 years. As it turns out, the simultaneous lineup procedure that the police were already using is, if anything, diagnostically superior to the sequential procedure. Figure 8 presents the relevant data from Experiment 1a of Mickes et al. (2012). It may be true that the sequential procedure often induces a more conservative response bias, and many who believe that the false identification rate in the real world needs to be reduced find comfort in that fact. However, such reasoning denies signal-detection logic. If a lower false ID rate is the goal, the solution would be to induce more conservative responding using the diagnostically superior procedure (the simultaneous lineup), not to switch to using a potentially diagnostically inferior procedure (the sequential lineup).

**Signal Detection Theory and the Confidence-Accuracy Relationship**

The data collected to plot a confidence-based ROC can be analyzed in a more intuitive way by simply examining the confidence-accuracy characteristic (Mickes, 2015). For a yes/no detection task, accuracy for yes decisions is equal to $\text{HR}/(\text{HR} + \text{FAR})$, where the subscript $c$ refers to a particular level of confidence. Note that these values are not cumulative over confidence, as they would be when plotting confidence-based ROC data. For example, if a 6-point rating scale is used (where ratings of 1
Figure 8. Simultaneous and sequential ROC data from Experiment 1a of Mickes et al. (2012). This was a mock-crime study in which subjects viewed a simulated crime and were then presented with a target-present lineup (containing six photos, one of which depicted the guilty suspect and five of which depicted innocent fillers) or a target-absent lineup (in which the guilty suspect’s photo was replaced with the photo of an innocent filler). The hit rate is the proportion of target-present lineups in which the guilty suspect was identified, and the false alarm rate is the proportion of target-absent lineups in which an innocent filler was identified (this value was then divided by 6 to estimate the probability that a single innocent suspect would be misidentified). Note that these ROCs often seem strange to those familiar with signal detection theory because the maximum false suspect would be misidentified. However, for signal detection theory and also low accuracy because such signals will be fairly common on both stimulus-present and stimulus-absent trials. For parallel reasons, signal detection theory naturally predicts a continuous confidence-accuracy relationship for yes decisions.

Confidence and Accuracy According to Signal Detection Theory

In contrast to the inherent predictions of threshold theories, Gaussian-based signal detection theory inherently predicts a continuous confidence-accuracy relationship for both yes and no decisions. For example, as illustrated in Figure 9, sensations that fall far to the right of the yes/no decision criterion—leading to high-confidence yes decisions—will be associated with high accuracy because such signals mainly occur on stimulus-present trials. By contrast, sensations that fall near the yes/no decision criterion will be associated with low confidence and also low accuracy because such signals will be fairly common on both stimulus-present and stimulus-absent trials. For parallel reasons, signal detection theory naturally predicts a continuous confidence-accuracy relationship for no decisions.

Confidence in Yes Decisions

An absolutely ubiquitous finding throughout all areas of psychology is that confidence and accuracy for yes decisions are related in continuous fashion, as predicted by signal detection theory. One apparent exception was eyewitness memory, where it was long assumed that confidence is not very predictive of accuracy, even on an initial test from a lineup (e.g., Penrod & Cutler, 1995). On the surface, this idea seems improbable given that it contrasts with a vast body of evidence supporting the signal detection perspective. As it turns out, and contrary to what many textbooks still say and what many professors still teach in their classes, on an initial uncontaminated test of memory from a lineup, eyewitness confidence is highly predictive of accuracy, in continuous fashion, just as signal detection theory predicts (Brewer & Wells, 2006; Justsen, Olsson, & Winman, 1996; Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Mickes, Dunn, Clark, & Wells, 2016; Wixted & Wells, 2017). It is later, at trial, that memory is often contaminated to the point where confidence is now dissociated from accuracy.

Unfortunately, jurors attach weight to the confidence expressed by the eyewitness at trial. Like all forms of forensic evidence (e.g., DNA, fingerprints, etc.), eyewitness evidence is unreliable when it is contaminated. Although the idea that memory can be contaminated once came as a surprise, it is now established knowledge (Loftus & Ketcham, 2003; Loftus & Ketcham, 1994; Loftus, Miller, & Burns, 1978; Loftus & Pickrell, 1995). It therefore makes sense to ask about the confidence-accuracy relationship when memory is not contaminated. A recent review of the relevant literature by Wixted and Wells (2017) shows that, on an initial uncontaminated test, the relationship between confidence and accuracy for a positive ID from a lineup (i.e., for yes decisions that land on and therefore

9 For the equal base-rate situation (i.e., an equal number of stimulus-present and stimulus-absent trials), these equations represent the Bayesian posterior probability of being correct.
imperil the suspect) could scarcely be stronger than it is. The results are summarized here in Figure 10. Based on results like these, Wixted (2018) argued that eyewitness evidence is reliable in the same way that fingerprint evidence and DNA evidence are reliable. Specifically, they are reliable when the evidence is not contaminated.

Confidence in No Decisions

An example of the strong confidence-accuracy relationship from the field of basic recognition memory is shown in Figure 11 (Mickes et al., 2007). Note that, in Figure 11, a relationship between confidence and accuracy is evident even for no decisions. As far as I can determine, that fact has received less attention than it should. It deserves attention because it is hard to reconcile with both high-threshold theory and low-threshold theory. According to those theories, below the threshold of conscious awareness, there should be no relationship between confidence and accuracy because no diagnostic information is available to the observer. Contrary to that idea, diagnostic information is available even when the subject believes that the stimulus was not presented on the current trial. The data are also not fully compatible with two-high-threshold theory, which naturally predicts a binary confidence-accuracy relationship for both no decisions and yes decisions.

Even in studies of low-level sensation and perception, where a threshold of conscious awareness seems most likely to be found, if it plays any role a continuous confidence-accuracy relationship is observed for no decisions. As an example, Koenig and Hofer (2011) investigated the ability of subjects to detect a brief flash presented in darkness. There were four kinds of trials: a strong flash, a medium flash, a weak flash, or no flash. A 6-point confidence scale was used, with the first 2 levels indicating de-

---

10 In most studies of simultaneous lineups, no decisions (lineup rejections) are not made in relation to any particular face in the lineup. Instead, all of the faces are rejected at once, and confidence in a no decision applies to the set of faces. Signal detection theory does not make a clear prediction about confidence and accuracy under such conditions. Interestingly, for such no decisions, the relationship between aggregate confidence and accuracy is often weak (Brewer & Wells, 2006).
degrees of confidence in a no decision (1 = definitely not seen, 2 = not seen but uncertain) and the next 4 levels indicating degrees of perceived brightness. From the data presented in their Table A1, I computed the average accuracy for ratings of 1 and 2 for each of the five subjects. For no decisions made with high confidence (ratings of 1), average accuracy was 75% correct, whereas for no decisions made with low confidence (ratings of 2), average accuracy was 59% correct, a significant difference, \( t(4) = 3.40, p = .027 \). Thus, in both perception and memory, and as predicted by signal detection theory, a diagnostic signal is clearly present even on trials associated with a no decision. According to signal detection theory, there is diagnostic information associated with these negative sensations because, although those sensations fall below the decision criterion, they fall above the threshold.

The continuous confidence–accuracy relationship for no decisions (indicative of diagnostic information below an ostensible threshold) is conceptually analogous to studies of \( n \)-alternative forced-choice perception tasks (with \( n > 2 \)) in which, following an error (which is theoretically indicative of a below-threshold signal), subjects are given a second choice and found to perform with above-chance accuracy (Green & Swets, 1974, p. 108). Once again, such findings indicate a diagnostic signal where no such signal should be found if a threshold exists (see related work involving recognition memory by Kellen & Klauer, 2011, 2014).

The data considered above for confidence in no decisions are especially problematic for high-threshold theory and low-threshold theory, but there is a way to rescue two-high-threshold theory. Again, however, it requires the addition of arbitrary assumptions (the same assumptions needed for this theory to accommodate a curvilinear ROC). As noted earlier, these additional assumptions are not principled in that they are not based on any coherent theory of confidence. Adopting arbitrary assumptions to bring two-high-threshold theory into line with confidence data is a useful mathematical exercise, but it seems fair to say that is not a compelling explanation for the empirical data.

### Nonintuitive Predictions About Confidence and Accuracy

Signal detection theory provides an effective guide for conceptualizing confidence over and above the confidence-accuracy relationship in ways that threshold models never could. Consider, for example, a recent study by Sanders, Hangya, and Kepecs (2016). Subjects listened to separate auditory click streams delivered independently to each ear, and the task was to indicate whether the faster clicking stream was in the left ear or the right ear. Thus, this was a 2AFC task. Following each left or right decision, confidence was rated using a 5-point scale (1 = low confidence; 5 = high confidence). The strength of the signal was determined by the balance of left and right click rates, and it varied randomly and uniformly across trials over a wide range, yielding accuracy scores that ranged from chance \( (d' = 0) \) to near perfect performance \( (d' \gg 0) \). Four main findings were reported: First, aggregated over all levels of discriminability, confidence strongly predicted accuracy. Second, and counterintuitively, the average level of confidence increased with \( d' \) for correct responses, but it decreased with \( d' \) for error responses. Third, and again counterintuitively, for the subset of trials with zero evidence discriminability (such that \( d' = 0 \)), average confidence was intermediate (approximately 3 on the 5-point confidence scale). And fourth, for a given level of signal strength (i.e., for a given \( d' \)), confidence predicted accuracy. The upper panel of Figure 12 provides the simplest signal detection interpretation of this task, and the lower panel shows that it naturally anticipates all of these findings. Critically, in making these predictions, no arbitrary assumptions needed to be added to the standard Gaussian-based signal detection model of discrimination. Instead, these are the inherent predictions of signal detection theory.

### Speeded Decisions and Sequential Sampling Models

The considerations discussed above show that signal detection theory accounts for ratings of confidence about which stimulus was presented on the current trial (e.g., Stimulus A vs. Stimulus B, or Signal vs. Noise). However, when subjects are pressured to make speeded decisions, it is often the case that they are first asked to make a binary decision about the stimulus (e.g., Stimulus A vs. Stimulus B) and are then asked to express confidence in the correctness of that initial decision. Although the distinction is subtle, expressing confidence that the stimulus was A or B (Type I) is different from expressing confidence in the correctness of a prior binary decision (Type II). A Type II rating is a metacognition judgment (e.g., can people discriminate their correct yes decisions from their incorrect yes decisions?), whereas a Type I rating is a cognitive/perceptual judgment (e.g., can people discriminate the presence of a stimulus vs. the absence of a stimulus?). As noted by Galvin, Podd, Drga, and Whitmore (2003), from a purely methodological standpoint, either a Type I decision or a Type II decision can be made using either a confidence rating scale or a binary choice.

---

11 Note that perceived brightness is not equivalent to certainty that a flash occurred (e.g., one can be certain that a dim flash occurred), so the upper end of this scale is problematic for any kind of signal detection analysis (cf. Jiang & Metz, 2010).
Sequential sampling models of speeded decision-making assume that an initial decision about the presented stimulus is made when the accumulated evidence reaches a threshold. An intriguing theoretical question is whether, at that exact moment, the subject has access to additional information that could be used to meaningfully express confidence on a more fine-grained scale. A number of sequential sampling models of Type II confidence do not speak to that issue because, for the Type I decision, they model the binary decision that occurs when a boundary is reached. With regard to confidence, a Type II rating is then based on information accumulated during a brief interval of time following the initial binary decision. (e.g., Moran, Teodorescu, & Usher, 2015; Pleskac & Busemeyer, 2010). With regard to the initial binary decision about which stimulus was presented, these models are akin to two-high-threshold models except that (a) both stimuli can reach either threshold and (b) the below-threshold state is not a relevant consideration because decisions only occur when a threshold is reached.

Other sequential sampling models are more akin to signal detection theory in assuming that, at the moment the Type I decision made, continuous evidence is available upon which to base a confidence judgment (i.e., without the accumulation of additional information). For example, some models involve two accumulators that race to different thresholds, with the decision (“A” or “B”) being based on the winner and confidence being determined by the separation of the two accumulators at that moment (Merkle & Van Zandt, 2006; Van Zandt, 2000; Vickers, 1970, 1979). Alternatively, the decision itself might be based on a continuous evidence signal (Ratcliff & Starns, 2009, 2013). For example, the RTCON2 model proposed by Ratcliff and Starns (2013) begins with a standard signal detection representation for confidence (illustrated earlier for yes decisions in Figure 9) and assumes that the area under the Gaussian distribution for each level of confidence determines the drift rate associated with independent accumulators (one for each confidence rating). The first one to reach a decision boundary determines the rating.

RTCON2 is an extension of the influential diffusion model (Ratcliff, 1978), and it allows the model to account for both RTs and Type I confidence on both detection tasks and 2AFC discrimination tasks. According to this model, if subjects are asked to make binary Type I decision, they are essentially being asked to make a decision using a 2-point confidence scale (e.g., 1 = A and 2 = B). However, because continuous information is available even at that stage, the rating scale could instead be 1 = Sure B, 2 = Probably B, 3 = Maybe B, 4 = Maybe A, 5 = Probably A and 6 = Sure A. Because its quantitative details differ from standard signal detection theory, RTCON2 offers a qualitatively similar but quantitatively different interpretation of ROC data (e.g., Starns & Ratcliff, 2014). Nevertheless, of the various sequential sampling models, it is the most akin to signal detection theory in that it models Type I decisions in terms of a continuous evidence variable. In addition, it inherently predicts a continuous confidence accuracy relationship for both yes and no decisions for the same reason that signal detection theory does.

**Ideal Observer Theory (1980s–Present Day)**

Thus far, I have illustrated how signal detection theory effectively guides thinking about general trends in performance (e.g., it predicts curvilinear ROCs and a continuous relationship between confidence and accuracy for both yes and no decisions). However, a more rigorous approach uses signal detection theory to precisely specify what *optimal performance* would be given the inherent statistical uncertainties associated with the task (Green & Swets, 1966; Hecht et al., 1942). When the signal and noise distributions overlap, as they do by design in most lab tasks, perfect performance is not possible, but many different levels of imperfect performance are possible. The *ideal observer* would respond in such a way as to maximize some criterion definition of optimality, such as overall utility or overall percent correct.

As an example, in the simplest signal detection scenario (i.e., equal variance, equal priors), the Bayesian posterior probability of
being correct is maximized when the decision criterion is placed midway between the signal and noise distributions (i.e., at the point where the two distributions intersect). At that point, the odds are even that a given sensation, $x$, was generated by signal ($s$) or noise ($n$). Formally, the yes/no criterion would be placed at the point where $P(x|s)/P(x|n) = 1$. If humans always placed their decision criterion at that point, then they would be behaving as ideal observers. Moreover, when $d'$ decreased, the hit rate would decrease and the false alarm rate would increase. In studies of recognition memory, this predicted phenomenon is so universally observed that it is considered a lawful regularity, one known as the mirror effect (Glanzer & Adams, 1985, 1990; Glanzer, Adams, Iverson, & Kim, 1993). Thus, in this regard, humans behave a lot like (though not exactly like) ideal observers, and that fact is now incorporated into most models of recognition memory (McClelland & Chappell, 1998; Osth, Dennis, & Heathcote, 2017; Shiffrin & Steyvers, 1997). What is true of the yes/no decision criterion also happens to be true of the full range of confidence criteria. For example, the criterion for making a high-confidence yes decision might be placed at the point on the decision axis where the odds are 10-to-1 that a given sensation, $x$, was generated by signal ($s$) versus noise.

**Figure 12.** Upper panel: A signal detection interpretation of the 2FC task used by Sanders, Hangya, and Kepecs (2016). On each trial, subjects indicated whether the faster clicking stream was in the left ear or the right ear, and confidence was rated using a 5-point scale ($1 = low confidence; 5 = high confidence$). The strength of the signal was determined by the balance of left and right click rates, and it varied randomly and uniformly across trials over a wide range. Here, three levels of signal strength are depicted (including no signal, which is when the balance is equal). Lower panel: Predictions derived from the model depicted in the upper panel. Aggregated over all levels of discriminability, confidence strongly predicted accuracy (left). In addition, the average level of confidence increased with $d'$ for correct responses, but it decreased with $d'$ for error responses (middle). Finally, for a given level of signal strength (i.e., for a given $d' > 0$), confidence predicted accuracy in that accuracy was higher for confidence ratings of 4 and 5 than for confidence ratings of 1 – 3.
(\(n\)). Formally, the criterion might be placed where \(P(x|y)/P(x|n) = 10\). If human observers always placed their high-confidence yes criterion at that point, then, when \(d'\) changed, its location would also change in such a way as to maintain that likelihood ratio (thereby ensuring that high confidence yes decisions are associated with high accuracy). Although, in practice, observers do not adjust their criteria as much as they should (i.e., they are not actually ideal observers), in studies of recognition memory, they behave more like ideal observers than is predicted by other models (Stretch & Wixted, 1998). This is true even in studies of eyewitness identification (Semmler, Dunn, Mickes, & Wixted, 2018).

Ideal observer analysis is by no means limited to recognition memory and is perhaps most well developed in vision science (Geisler, 1989, 2003, 2011; Seidemann & Geisler, 2018). As an example, Eckstein (1998) measured target detection accuracy in a search task involving briefly presented displays in which set size varied from 2 to 12. The target might be an open ellipse with the distractors being filled ellipses (feature display), or the target might be a tilted open ellipse, with the distractors being tilted filled ellipses or upright open ellipses (conjunction display). On each trial, subjects judged whether or not the display contained the target at a cued location (indicated by a rectangular box). The performance of a signal-detection-based ideal observer model declines with set size because of the increased chance that noise features from distractors will be mistaken for target features, but the predicted decline is much greater for serial processing models than for parallel processing models. The results corresponded closely to an ideal observer model in which (a) each feature dimension is processed independently (in parallel) with inherent neural noise and (b) information is combined linearly across feature dimensions. This was true for both feature and conjunction searches, which are usually assumed to involve different processing mechanisms.

Beyond recognition memory and vision science, ideal observer analysis has been employed in areas such as medical imaging (Kupinski, Hoppin, Clarkson, & Barrett, 2003), neuroscience (Christison-Lagay, Bennur, & Cohen, 2017), and categorization (Erev, 1998), to name a few. In each case, a signal-detection-based ideal observer model was specified that performed the relevant task at the optimal level, given the available information and specified constraints (Geisler, 2011).

In a related vein, in the 1980s, unidimensional signal detection theory was expanded to include the multidimensional scenario with the development of General Recognition Theory (GRT: Ashby & Townsend, 1986). GRT is an influential detection framework that applies when decisions are made about stimuli that vary along more than one dimension (e.g., faces that vary in emotional expression and age) and the goal is to determine how those perceptual dimensions interact (e.g., Maddox & Ashby, 1996; Soto, Vucovich, Musgrave, & Ashby, 2015). Like the unidimensional examples considered above, GRT has often been used in conjunction with ideal observer theory (e.g., Ashby & Gott, 1988; Ashby & Perrin, 1988; Soto & Waterman, 2011; Sridharan, Steinmetz, Moore, & Knudsen, 2014). These widely used methodologies (i.e., ideal observer theory and GRT) further underscore the broad scientific utility of signal detection theory.

### The Neuroscience of Signal Detection Theory (1980s–Present Day)

As noted earlier, Fechner was mainly a threshold theorist. However, like Helmholtz, he also believed that neural noise, especially in the visual domain, could generate conscious sensations (contrary to the threshold view), and, at times, he entertained the idea that sensations produced by a physical stimulus are superimposed on neural noise. These ideas correspond to modern-day signal detection theory, and, as reviewed next, they are both consistent with a considerable body of evidence from neuroscience.

#### Visual Sensation in Complete Darkness

Some studies have investigated visual sensations that occur in complete darkness (something that, according to high-threshold theory, does not happen). In the eye, photons from a flash that are absorbed by rods trigger photoisomerization (i.e., a structural change) of the rhodopsin molecule, resulting in a neural response of the rod photoreceptor and, ultimately, the sensation of a flash of light (Hecht et al., 1942). Several studies have found that noise signals also occur in the retina as a result of random thermal isomerizations (i.e., a structural change of the rhodopsin molecule from warmth alone) that are indistinguishable from photoisomerization signals (Ashmore & Falk, 1977). If so, then experiential sensation could arise in the absence of stimulation (i.e., even in complete darkness). In fact, conceivably, Fechner’s dark light may be partly attributable to spontaneous activation of rhodopsin in rods.

To investigate this issue, Aho, Donner, Hyden, Larsen, and Reuter (1988) observed that the rate of thermal isomerizations in the retina increases with temperature, which means that sensitivity to a flash of light should decline as temperature increases because there would be more sensory noise to overcome. Thus, all else equal, the measured threshold of detection should be higher in species with higher body temperatures (e.g., humans) compared with those with lower body temperatures (e.g., frogs). Aho et al. (1988) tested this prediction by measuring detection thresholds as a function of the rate of thermal isomerizations across species. As predicted, the threshold increased with body temperature, perhaps explaining why frogs are better at detecting extremely low-intensity flashes of light than humans are. Thus, apparently, neural noise can give rise to the sensation of light even in complete darkness (consistent with the notion of a noise distribution in signal detection theory).

#### Single-Unit Recording Studies and Signal Detection Theory

Events happening in the retina may contribute to neural noise, but it seems likely that events happening in the brain do so as well. Quite a few studies in the fields of perception and memory have measured single-unit activity in regions of the brain thought to support behavioral decisions in a detection or discrimination task. The large majority of these studies support the signal detection perspective. Indeed, as Crapse and Basso (2015) put it “Without question, a significant breakthrough in efforts to understand the relationship between neuronal activity and perception came with the application of signal detection
theory in psychophysics (Green & Swets, 1966) to neurophysiology” (p. 3039).

Perception. Tolhurst, Movshon, and Dean (1983) argued that stimulus-driven neural activity is superimposed upon spontaneous activity that fluctuates across trials. Articulating a fundamental concept of signal detection theory, they noted that the variability of the discharge of visual cortical neurons limits the reliability with which such neurons can relay signals about weak visual stimuli (cf. Rose, 1942). As an example, Britten, Newsome, Shadlen, Cebalbrini, and Movshon (1996) took advantage of the fact that the responses of neurons in area MT are strongly determined by the direction and strength of visual motion signals. For example, the stimulus might consist of a field of moving dots, with a certain fraction carrying a unidirectional motion signal (e.g., 50% of the dots moving upward, the rest moving randomly). Some MT neurons are attuned to upward motion whereas others are attuned to downward motion. Britten et al. (1996) found that signals carried by MT neurons were associated, trial by trial, with the monkeys’ behavioral decisions. That is, on a given trial, a monkey was more likely to make a decision consistent with the preferred direction of a neuron when that neuron was firing more vigorously in response to a stimulus delivered to its receptive field. Critically, this effect was observed even on noise trials in which the stimulus contained no net motion signal.

Uka and DeAngelis (2004) reported similar findings with monkeys trained to perform a depth-perception task while also eliminating the possibility that stimuli on noise trials contained subtle depth signals that may have driven performance. They found that even on physically identical no-signal (noise) trials, the activity of MT neurons predicted trial-to-trial trial choice behavior (consistent with signal detection theory). Such findings are correlational, but they are compatible with earlier causal research showing that electrical stimulation of motion-selective MT neurons can bias perceptual judgments of depth (DeAngelis, Cumming, & Newsome, 1998). Thus, the authors concluded that MT transmits sensory signals that the monkeys relied upon to make decisions in their depth discrimination task. Conceptually similar results have been obtained using a variety of decision-making tasks (e.g., Carandini, 2004; Dodd, Krug, Cumming, & Parker, 2001; Niemborg & Cumming, 2006; Purushothaman & Bradley, 2005; Roelfsema & Spekreijse, 2001; Uka, Tanabe, Watanabe, & Fujita, 2005; van Vugt et al., 2018).

The studies discussed above are compatible with an essential assumption of signal detection theory, namely, that false positives reflect genuinely experienced false sensations. At the same time, they are hard to reconcile with either high-threshold theory or two-high-threshold theory. Still other studies support the signal detection interpretation of confidence. For example, in one intriguing study, Kiani and Shadlen (2009) measured the relationship between a monkey’s confidence in a decision about the direction of moving random dots and the activity of neurons in parietal cortex. On half the trials, the monkey indicated its direction choice by making an eye movement to one of two direction-choice targets. Critically, on the other half of the trials, the monkey was given the option to abort the direction discrimination and instead choose a small but certain reward by making a saccade to a third target. Choosing that third option would make sense on trials in which the money was uncertain about the true direction of motion. The key finding was that neurons in parietal cortex represented the formation of the direction decision and the degree of certainty underlying the decision to opt out (i.e., the mechanisms leading to decision formation and the establishment of a degree of confidence). This conclusion is compatible with the RTCON2 sequential sampling model: confidence is not just a metacognitive process based on a variable that differs from the one that underlies the binary decision; instead, they are based on the same variable, namely, the strength of the sensation. According to this view, a binary decision is, essentially, a decision made using a 2-point confidence scale.

Memory. In the memory domain, Rutishauser, Aflalo, Rosario, Pouratian, and Andersen (2018) recently reported conceptually similar findings with two human tetraplegic subjects implanted with microelectrode arrays while they performed a recognition task for a list of previously presented images. The subjects classified each test image as old or new using a 6-point confidence scale (1 = sure new, 6 = sure old). Single neuron activity was recorded from posterior parietal cortex, a region that has been previously implicated in memory retrieval. The key findings were that some neurons exhibited elevated activity for old decisions (whether correct or incorrect), whereas other neurons exhibited elevated activity for new decisions (again, whether correct or incorrect). However, these were not simply old-versus-new binary choice neurons because the degree of activity was modulated in continuous fashion by confidence in the memory decision, which, as noted by Rutishauser et al. (2018), is consistent with signal detection theory. Importantly, the continuous variation of the neural signals for neurons associated with new decisions is not compatible with either high-threshold theory or low-threshold theory.

Neuroimaging and Signal Detection Theory

In studies that measure brain activity, it is not just single unit recording research that has supported the signal detection perspective. A substantial body of neuroimaging research—in both perception and memory—has as well.

Perception. In an elegant study of perception, Ress and Heeger (2003) used functional MRI to measure activity in early visual cortex during a contrast-detection task. Each trial consisted of either a background pattern alone or a low-contrast target pattern superimposed on that background, and the subjects pressed a button to indicate whether or not the target was present. Critically, both hits and false alarms were associated with elevated activity in visual cortex compared with correct rejections and misses, directly supporting one of the basic tenets of signal detection theory. As they put it: “That false alarms evoked more activity than misses indicates that activity in early visual cortex corresponded to the subjects’ percepts, rather than to the physically presented stimulus” (p. 414). Although not all neuroimaging studies have detected elevated neural signals in sensory pathways associated with false detection (e.g., Hulme, Friston, & Zeki, 2009; Mostert, Kok, & de Lange, 2015), many have (e.g., Pajani, Kok, Kouider, & De Lange, 2015; Ress & Heeger, 2003; Watkins, Shams, Tanaka, Haynes, & Rees, 2006; Vililaite, Marsh, & Baker, 2019). In addition, still other neuroimaging studies have provided evidence that activity in higher (nonsensory) regions supports the false subjective experience of perception (e.g., Lloyd, McKen-
zie, Brown, & Poliakoff, 2011; similar conclusions about the role of higher brain regions were reached in single-unit work by de Lafuente & Romo, 2005, 2006).

Memory. Neuroimaging studies of memory have added to the already considerable body of neuroscience evidence supporting the signal detection perspective. For example, Slotnick and Schacter (2004) asked subjects to memorize a list of shapes. Later, they made recognition memory decisions (old or new) for old shapes and new-but-similar shapes. Using fMRI, they found that late visual regions were similarly active during true recognition and false recognition, concluding that: “It is possible that late visual processing regions (BA19, BA37) contribute to the conscious experience of remembering, thus supporting ‘old’ responses during both true and false recognition” (p. 667).

In a related study, Dennis, Bowman, and Vandenkerckhove (2012) used the Remember-Know procedure to isolate brain activity associated with true recollection (Remember judgments made at retrieval to old images) and false (or phantom) recollection (Remember judgments made at retrieval to new-but-related images). They found that true and phantom recollection were associated with a largely overlapping retrieval network. As they put it: “Results showed that both true and phantom recollection were mediated by a largely overlapping network, previously shown to support true recollection and memory-related reconstruction processes. Finding common activity associated with true and phantom recollection supports the theory that false retrieval can be based on erroneously triggered recollection processes” (p. 2991). This conclusion accords with the signal detection (and low-threshold) interpretation of recollection-based false alarms and contrasts with the high-threshold and two-high-threshold view that false recollections are random guesses.

More recently, Karanian and Slotnick (2017) used a spatial location memory task with humans in which abstract shapes were presented to the left or right of fixation during encoding. During encoding, subjects were instructed to remember each shape and its spatial location. During retrieval, the shapes were presented again, this time at central fixation, and subjects were asked to classify them as having been previously presented on the left or right. Focusing on recollection decisions made with high confidence, they found that both true and false memories for spatial location were associated with similarly elevated activity in parahippocampal cortex (a region of the medial temporal lobe thought to support spatial recollection). This is, of course, the pattern naturally anticipated by signal detection theory.

Given the intimate connection between perception and memory, and given the purely visual nature of the task used by Karanian and Slotnick (2017), it seems reasonable to suppose that true and false memories might also be associated with elevated activity in early sensory regions of the brain. In a subsequent report, Karanian and Slotnick (2018) noted that many prior memory studies using fMRI have detected greater neural activity in early visual cortex (V1) associated with true memories compared with false memories, but these studies had yet to provide convincing evidence of false memory-related activity in V1. Using the purely visual abstract shape memory procedure and focusing their analysis on decisions made with high confidence (i.e., focusing on the strongest memory signals according to signal detection theory), they found that both true and false memories for spatial location were associated with elevated activity in V1 (in accordance with signal detection theory).

Conclusion

Fechner’s (1860/1966) book The Elements of Psychophysics presented a novel approach to measuring the psychological distance between physically measurable stimuli (e.g., weights) using the 2AFC task, with the data interpreted in terms of a theory of invisible Gaussian error (Link, 1994). That theory endures—literally unchanged—to this day. Thurstone (1927) adopted—or possibly rediscovered—Fechner’s 2AFC methodology and used it to scale the psychological distance between stimuli that cannot be easily scaled in physical space (e.g., handwriting samples that differ in subjective beauty). However, it was not until 1953 that a breakthrough idea occurred: statistical noise in sensory neurons can give rise to consciously experienced sensation even in the absence of physical stimulation. This ingenious idea was anticipated by Fechner and Helmholtz in the mid-19th century, but it did not fully mature for another 100 years. Once it did, the importance of stimulus-absent trials became clear, and the field of perception (and shortly thereafter, the field of memory) replaced the notion of a fixed threshold of conscious awareness with the notion of an adjustable decision criterion. That idea, in turn, immediately led to ROC analysis, a state-of-the-art methodology that has advanced basic (theory-driven) research and transformed applied fields ranging from diagnostic medicine to eyewitness identification. As I have tried to demonstrate throughout this article, the breadth of basic and applied research that is naturally accommodated by signal detection theory (but not by any threshold theory) is simply phenomenal.

Signal detection theory, whether applied to the 2AFC task or the yes/no task, is a radical departure from where intuition usually leads. The fact that intuition virtually never leads to the path pioneered by Fechner helps to explain the headwinds that every teacher of signal detection theory faces when attempting to explain it to students. Although almost everyone is an intuitive threshold theorist (which explains why there is no fascinating history of threshold theory to tell), no one is an intuitive signal detection theorist. Its intuitive nature explains why high-threshold theory has been repeatedly reinvented under different names (e.g., see Wixted, 1993, for an example from animal learning) and why it likely will continue to be reinvented many times in the future. Even so, the effort required to understand and fully appreciate signal detection theory is worth it. After all, signal detection theory revolutionized our field as long ago as 1860 and continues to do so to this day in ways that no version of threshold theory has ever come remotely close to doing.

In commenting on the influence of signal detection theory, William Estes had this to say:

Over ensuing decades, the SD model, with only technical modifications to accommodate particular applications, has become almost universally accepted as a theoretical account of decision making in research on perceptual detection and recognition and in numerous extensions to applied domains (Swets, 1988; Swets, Dawes, & Monahan, 2000). This development may well be regarded as the most towering achievement of basic psychological research of the last half century. (Estes, 2002, p. 15)
I am hard-pressed to disagree, and it is why, in my view, it should not be possible to earn a Ph.D. in experimental psychology without having some degree of proficiency in signal detection theory and ROC analysis. Somehow, over the years, that seems to have become a minority point of view. Given recent developments in eyewitness identification, an argument could be made that the time has come for the field to reevaluate its training priorities.

References


von Helmholtz, H. (1962). *Beiträge zur Theorie der Sinneswahrnehmung (Con-...
Appendix A

2AFC Versus Yes/No $d'$

Scaling Versus Discriminability

A potentially confusing subtlety is that the estimated $d'$ between the mean sensations generated by Weight A and Weight B corresponds to the sensations that each weight generates before computing any difference score (i.e., before deciding which of the two weights is heavier). This is the $d'$ value illustrated in the upper panel of Figure 1 of the main text. This is also the $d'$ score that would theoretically be obtained if Fechner had used a yes/no task (instead of a 2AFC task) in which only one weight was presented on each trial and the observer had been asked to indicate whether or not it was the heavier of the two. Because there would be only one sensation to assess at a time, the decision would have to be made in relation to a criterion (illustrated in the upper panel of Figure A1). If the sensation exceeded that criterion, the decision would be heavier; if not, the decision would be lighter. In Figure A1, with the criterion is placed midway between the distributions in the upper panel, the hit rate would be .773 and the false alarm rate would be .227. Thus, using the standard computational formula, $d'$ would come to $z(.773) - z(.227) = 1.5r$.

Those familiar with modern-day signal detection theory know that, according to the simplest signal detection model, $d'_{2AFC} = \sqrt{2} d'_{yes/no}$, but that relationship did not show up in Fechner’s analysis. To appreciate why performance should be better on a 2AFC task compared with a yes/no task, consider the 2AFC problem from the observer’s point of view. The observer does not know which test item is Weight A and which is Weight B. Thus, a $b - a$ difference score cannot be computed on every trial. The analysis presented in Figure 1 predicts how often $b - a$ will be negative (leading to an error), but that formal analysis cannot reflect the subtraction that occurs in the head of the observer. Therefore, instead of computing $b - a$, the observer’s computation across trials might be the left sensation minus the right sensation (left – right).

For the subset of trials in which the heavier weight happens to be in the left hand, the mean of the distribution of left – right difference scores will be positive. For the remaining trials in which the heavier weight is in the right hand, the mean of the distribution of left – right difference scores will be negative. As illustrated in the lower panel of Figure A1, these distributions would give rise to two distributions that are mirror images of each other. The rightmost distribution is identical to the distribution shown in the lower panel of Figure 1, whereas the leftmost distribution is its mirror image. This way of conceptualizing the difference between a yes/no and 2AFC task illustrates why $d'_{2AFC} > d'_{yes/no}$. That is, for the equal-variance case, $d'_{yes/no} = \frac{\sqrt{2}a - \sqrt{2}b}{\sqrt{\sigma^2}} = 1.5$, whereas $d'_{2AFC} = \frac{\sqrt{2}(a - b)}{\sqrt{\sigma^2}} = 3\sqrt{2} = 2.12$. In the general equal-variance case, $d'_{2AFC} = \sqrt{2} d'_{yes/no}$.

Fechner was not interested in comparing 2AFC to yes/no tasks in terms of how well observers could tell the difference between (i.e., how well they could discriminate) the heaviness of two weights. Instead, he used the 2AFC task to scale the psychological distance between the mean sensations generated by different weights. Doing so yielded an estimate of the distance between subjective means that, theoretically, would also have been obtained had he used a yes/no task. But even if he had, his yes/no task would have still fundamentally involved a comparison of two stimuli (“is this the heavier of the two weights or not?”). Another theoretically informative task—one that could shed light on the threshold of conscious awareness that informed Fechner’s Law—would involve stimulus-present trials vs. stimulus-absent trials (e.g., “did I just place an extremely light weight in your hand or not?”). This detection task could be used to investigate the fascinating question of whether neural noise alone is capable of generating false sensations. Research along those lines would not be performed in earnest for nearly a century after Fechner’s famous book was published (not until the early 1950s).

(Appendices continue)
Appendix B

Measures of Discriminative Performance

The Algebra of Threshold Theory

The formal algebraic specification of high-threshold theory is straightforward. Let \( g \) represent the probability of guessing that a stimulus was presented on below-threshold trials. On stimulus-present (i.e., signal) trials, let \( p \) represent the probability that the signal generates an above-threshold sensation. The probability of a hit on a signal trial is the probability that the signal generated an above-threshold signal plus the probability that, if not, the observer guessed correctly anyway. In other words, the hit rate (HR) is given by:

\[
HR = p + (1 - p)g
\]

(A1)

If the inclusion of catch trials pushes \( g \) to essentially 0, then the hit rate alone provides the measure of interest, \( p \) (i.e., the proportion of stimulus-present trials in which the stimulus was consciously detected). If \( g \) is not equal to 0, a few more steps are needed to estimate \( p \).

The probability of a false alarm—that is, the false alarm rate (FAR)—provides a direct estimate of the guessing rate because an above-threshold signal never occurs on stimulus-absent trials. Thus, all false alarms are pure guesses, and the rate at which they occur is what \( g \) represents. Thus,

\[
FAR = g
\]

(A2)

Substituting FAR for \( g \) in Equation A1 yields:

\[
HR = p + (1 - p)FAR
\]

(A3)

Solving for \( p \) yields the following result:

\[
p = \frac{HR - FAR}{1 - FAR}
\]

(A4)

Thus, for example, if \( HR = .80 \) and \( FAR = .20 \), then the probability of detecting a stimulus on signal trials (\( p \)) comes to \((.80 - .20)/(1 - .20) = 60/80 = .75 \). Equation A4 is known as the standard correction for guessing, and it was used for many years and is occasionally used still today.

Following similar logic, the simplest version of two-high-threshold theory warrants a dependent measure given by \( HR = FAR \). As in high-threshold theory, the \( HR \) is given by Equation A1. Now, however, a different equation applies to the FAR:

\[
FAR = (1 - p_2)g_2
\]

(A5)

where \( p_2 \) is the probability of exceeding the noise-detection threshold on stimulus-absent trials (in which case a correct no decision is made) and \( g_2 \) is the probability of guessing yes on below-threshold noise trials. If, for simplicity, we assume that \( p_2 = p \) and that \( g_2 = g \), then we can write Equation A1 as:

\[
HR = p + FAR
\]

(A6)

From Equation A6, it follows that:

\[
p = HR - FAR
\]

(A7)

Equations A4 and A7 perfect sense given the assumptions of high-threshold theory and tow-high-threshold theory, respectively, which underscores a critical point that is too often overlooked: whatever measure of performance an experimenter chooses to use on a detection task or a discrimination task, that measure necessarily embraces specific theoretical assumptions. Thus, no matter how the \( HR \) and \( FAR \) are combined to yield a dependent measure, the experimenter should be cognizant of the theory that is implicitly embraced by virtue of choosing to use that measure. Although every experimenter should know this, in my experience, many do not. This seems especially true of domains that are at least one step removed from basic experimental psychology (e.g., cognitive neuroscience, applied psychology, etc.).

Signal Detection Theory and the Gaussian Assumption

To many scientists, the Gaussian assumption of signal detection theory is, a priori, more defensible than any other assumption, which is also true of many ordinary statistical analyses. For example, Green and Swets (1966) defended the Gaussian assumption as follows:

One general result is that if the random variables of the sum are independent and all have the same distribution, then, whatever this distribution, the sum of such variables approaches a Gaussian distribution as the number of variables increases indefinitely. . . . Since we often think that sensory events are composed of a multitude of similar, smaller events, which are by and large independent, the central limit theorem might be invoked to justify the assumption of a Gaussian distribution of net effects. (p. 58)

Obviously, there is no guarantee that the underlying distributions are Gaussian in form, but if the goal is to measure the mind, some distributional form must be assumed. The measure of discriminability derived from Gaussian-based signal detection theory (namely, \( d' \)) is a more defensible measure than most.

Then again, even if one does assume that the underlying distributions are Gaussian in form, the \( \text{equal-variance} \) assumption can be reasonably questioned. Indeed, certain tasks, such as recognition memory, are known to be better characterized by an unequal-variance signal detection model than by the standard equal-variance model that justifies the use of \( d' \) (Egan, 1958; Rotello, 2017; Wixted, 2007). In that case, a better choice for a dependent measure would be to use \( d'' \), which is a \( d' \)-like measure that takes into account the fact that the signal and noise distributions do not have the same variance (Macmillan & Creelman, 2005; Rotello, 2017). Conceptually:

(Appendices continue)
\[ d_a = \frac{\mu_{signal} - \mu_{noise}}{\sqrt{(\sigma_{signal}^2 + \sigma_{noise}^2)/2}} \]

Note that this equation reduces to the standard equation for \( d' \) when \( \sigma_{signal}^2 + \sigma_{noise}^2 = \sigma^2 \):

\[ d' = \frac{\mu_{signal} - \mu_{noise}}{\sigma} \]

Computationally, \( d_a = \sqrt{\frac{z[HR] - z[FAR]}{\sqrt{\frac{1}{s^2} + \frac{1}{z^2}}}} \), where \( s \) is the slope of the z-ROC (Macmillan & Creelman, 2005, p. 370, Equation 3.5). Thus, to compute this measure, one needs to collect confidence rating and fit the z-ROC data with a straight line. Note that in the equal-variance case, \( s = 1 \), in which case this computational formula reduces to the standard equation for computing \( d' \):

\[ d' = z[HR] - z[FAR] \]

\( d_a \) is an especially useful measure for recognition memory, where, usually, \( s < 1 \) (Egan, 1958).

**Area Under the Curve**

What about area under the curve (AUC), which is literally a measure of the geometry of the ROC curve that makes no assumptions about the mind whatsoever? AUC provides the only true nonparametric (theory-free) measure of discriminability, so it seems attractive for that reason, but it is a mistake to believe that it relieves a scientist of the burden of making an assumption about the distributional form of the psychological variable under investigation. AUC is the measure to use when you care nothing about measuring the mind and your only goal is to measure empirical performance. In eyewitness identification, for example, the police care about using the procedure that maximizes discriminability regardless of what any theory says about underlying discriminability. But for a scientist interested in measuring the mind, the opposite is true. Indeed, under some conditions, Gaussian-based \( d' \) and theory-free AUC can yield opposite conclusions (Wilson, Donnelly, Christenfeld, & Wixted, 2019; Wixted & Mickes, 2018).

For example, because criterion variability impairs performance (Benjamin, Diaz, & Wee, 2009), in a condition with higher \( d' \) but also higher criterion variability, the AUC could be lower than a condition with lower \( d' \) and lower criterion variability (Wixted & Mickes, 2018). In cases like that, it is \( d' \) (estimated by fitting a model that also estimates criterion variability), not AUC, that provides the theoretician with a measure of the mind.

Although it is always a good idea to use the most theoretically sensible measure of performance, a variety of commonly used measures (\( A' \), \( HR - FAR \), proportion correct, the standard correction for guessing, \( d' \), \( d_a \), AUC, etc.) will usually agree with each other so long as response bias does not differ appreciably across conditions (Snodgrass & Corwin, 1988). Thus, even a measure that implicitly embraces a ridiculous theory (e.g., \( A' \)) can often yield the correct interpretation of the data. That said, the various measures will not always agree, and even when they do, the statistical conclusion will not always be the same. For example, the two measures might agree on the direction of an effect, but the \( p \) value might be .21 using \( d' \) and .02 using a less sensible measure. The fact that the less sensible measure yields a significant (and potentially publishable) result would not be a reason to use it.

Received October 24, 2018
Revision received April 20, 2019
Accepted April 22, 2019