Vol.5 , No.2, 1999, pp. ***–***. / Visual Cognition

A signal detection model applied to the stimulus: Understanding covariances in face recognition experiments in the context of face sampling distributions

Alice J. O'Toole, James C. Bartlett, & Hervé Abdi

The University of Texas at Dallas

We provide a description and interpretation of signal detection theory as applied to the analysis of an individual stimulus in a recognition experiment. Despite the common use of signal detection theory in this context, especially in the face recognition literature, the assumptions of the model have rarely been made explicit. In a series of simulations, we first varied the stability of d' and C in face sampling distributions and report the pattern of correlations between the hit and false alarm rate components of the model across the simulated experiments. These kinds of correlation measures have been reported in recent face recognition papers and have been considered to be theoretically important. The simulation data we report revealed widely different correlation expectations as a function of the parameters of the face sampling distribution, making claims of theoretical importance for any particular correlation questionable. Next, we report simulations aimed at exploring the effects of face sampling distribution parameters on correlations between individual components of the signal detection model, (i.e., hit and false alarm rates), and *other* facial measures such as typicality ratings. These data indicated that valid interpretations of such correlations need to make reference to the parameters of the relevant face sampling distribution.

1. Introduction

A number of researchers have employed techniques based on signal detection theory (Green & Swets, 1966) with the aim of measuring the recognizability of *individual faces* in an experiment (e.g., Bartlett, Hurrey, & Thorley, 1984; Hancock, Burton & Bruce, 1996; Light, Kayra-Stuart, & Hollander, 1979; Light, Hollander, & Kayra-Stuart, 1981; O'Toole, Deffenbacher, Valentin, & Abdi, 1994; Valentine & Bruce, 1986; Vokey & Read, 1992). Applied to the analysis of individual faces rather than individual observers, these techniques have proven useful for understanding both theoretical and practical issues in human face processing. For example, the relationship between the rated typicality and recognizability of faces is the primary evidence for a prototype-based account of human face processing. Additionally, measurement-based approaches to the recognizability of individual faces have proven useful for understanding applied issues in forensic psychology, such as the relationship between confidence and accuracy in face recognition (Deffenbacher, 1980) and the relationship between facial description accuracy and recognition accuracy (Pigott, Brigham & Bothwell, 1990).

Despite its widespread use, the implicit assumptions underlying a signal detection theory model applied to the analysis of individual stimuli are rarely made explicit. To our knowledge, the issues have been addressed only once, in the appendix of O'Toole, Deffenbacher, Valentin, McKee, Huff, and Abdi (1998),

Correspondence about this paper should be send to Alice O'Toole, Program in Applied Cognition and Neuroscience, The University of Texas at Dallas, Richardson, TX 75083-0688, U.S.A.; e-mail: otoole@utdallas.edu. A.O. gratefully acknowledges support from NIMH grant 1R29MH5176501A4. Thanks are also due to Ken Deffenbacher, Bob Johnston, Robin Thomas, and two anonymous reviewers for helpful comments on a previous version of this manuscript; and to Greg Murphy and Jay Dowling for helpful and stimulating discussions.

and there, only briefly. In the present paper we sketch out a full description. Given the extremely common application of signal detection theory to the description of individual *observer* performance in a particular condition of an experiment, this may seem unnecessary. We would argue, however, that there are a number of crucial differences in how these signal detection measures *tend to be used* for stimulus versus observer-based performance descriptions. For example, whereas individual differences in observer performance obviously occur in an experimental condition, they are rarely considered "interesting" in a theoretical sense. In fact, they are mostly viewed as an experimental annoyance, which is dealt with by the application of inferential statistics. By contrast, although the majority of face recognition studies operate at the level of observer analysis, individual differences in face recognizability have become an important theoretical focus in the face recognition literature. For example, individual differences in face recognizability form the backbone evidence for theories of face processing that build on face space models (Valentine, 1991). Additionally, the primary evidence for prototype theory as applied to faces is the finding that the recognizability of individual faces correlates inversely with ratings of facial typicality (cf., Bruce & Young, 1986; Light et al., 1979). Individual differences in faces have been used also for assessing the categorical structure of face spaces.

The focus on studying individual stimuli has been extended recently by a number of researchers to address questions about the *kinds* of recognition errors people tend to make with individual faces. These questions have obvious practical value in the context of eyewitness identification accuracy. Thus, researchers have asked questions like, "What kinds of faces elicit many false alarms?", "What kinds of faces elicit few hits?", and "What factors underlie the relationship between hits and false alarm rates for individual faces?". Good examples of this kind of questioning can be found in recent papers by Hancock, Burton and Bruce (1996) and Lewis and Johnston (1997) who consider the theoretical significance of their finding that hit and false alarm rate errors did not correlate in their recognition experiment. Both suggest that this result indicates a dissociation of the processses and/or information used in recognizing faces as "old", and the processes and/or information used in rejecting faces as "novel". While these questions seem to have a prominent role when signal detection theory is applied to stimuli, they are ascribed theoretical status only rarely when the analysis is applied to the description of observer performance.[1] The important difference in the uses to which signal detection measures have been put for the individual observer versus individual face case motivates the present simulations. Signal detection theory is sufficiently complicated that our intuitions may be of only limited utility in predicting how the different computational components of the model interrelate under varying assumptions about the properties of the distribution of faces serving as stimuli in a particular experiment.

The aims of this paper are : 1.) to give an explicit presentation of the signal detection model implied in measuring the recognizability of individual stimuli; 2.) to measure the pattern of correlations between the components of hit and false alarm rate that are expected for face samples that vary in the mean and standard deviation of their characteristic discriminability index and criterion; and 3.) to explore the source of correlations between a single component of the discriminability index and criterion, (either hit or false alarm rate), and *another measure* such as typicality. For this latter question we look at the extent to which criterion and discrimination index variation in the face sample relate to correlations obtained between a nonsignal detection measure of faces, such as a facial rating, and hit or false alarm

[1] Readers who have followed the very active literature on the mirror effect (Glanzer and Adams, 1990) and its surrounding controversy (Hintzman, 1994; Hintzman, Caulton & Curran, 1994), will note a number of complex and subtle connections with the issues we raise here. The difference in the approach we have taken here to understanding these issues is dictated in large part by an historical difference between the face recognition literature and the more general recognition literature. In the former literature, stimulus properties (e.g., typicality) have been analyzed generally as continous variables that are measured with stimulus-based analyses in the broader context of an experiment. Among these stimulus-based measures, correlations of all sorts have been reported. Our efforts in this paper are directed at understanding these kinds of covariance measures applied *a posteriori* to samples of faces, each described by a signal detection model. In the latter literature, stimulus properties have been treated most commonly as dichotomous independent variable manipulations (e.g., high versus low-frequency words). We return to the implications of this issue in the discussion.

rates. Although it is possible to make the primary points for the second and third goals of this paper by using a more mathematical or analytical approach to the problem, we have chosen to present simulations for two reasons. First, the simulation approach can be presented in a manner that makes it accessible to a broader audience of researchers in the area of face and object recognition who are not necessarily specialists in mathematical psychology. Second, and equally important, there are a number of open parameters concerning the form of the distributions. These open parameters make the simulation approach applicable to real experiments for which the parameters are known, but do not support the assumptions required for the application of a mathematical analysis.

2. The Signal Detection Theory Model of Observers and Stimuli

In this section we make explicit the model underlying the application of signal detection theory to individual stimuli. Readers who are not very familiar with signal detection theory, or who would like a quick refresher course in the context of analyzing observer behavior in a face recognition task, are referred to the presentation in Appendix A. This appendix provides a complete description of the observer model in a face recognition experiment. We include this as a complete self-contained foundation for the analogy we develop for the stimulus model.

It is worth noting that the description we present in the next section can apply to any kind of stimulus employed in a standard yes/no recognition experiment. For concreteness and clarity, however, we use the problem of face recognition as an example.

The signal detection model in experimental psychology is applied occasionally to the task of measuring or describing the "behavior" of a *single stimulus*. Most commonly in the literature, it is the *recognizability* of a particular face that is of interest. Thus, just as some people are better at face recognition than other people, some faces are more recognizable than other faces. As is the case for other stimulus-based measures, to compute the recognizability of a face, data are collapsed across different observers. So, just as data are collapsed across face stimuli to measure the performance of a single observer, data are collapsed across observers to measure the recognizability of a single face.

Additionally, the signal detection model allows for the meaningful computational assessment of a face's criterion. While the interpretation of this criterion is somewhat less certain for the stimulus model than for the observer model, we consider cases in the discussion section for which stimulus-based criterion fluctuation occurs and can be theoretically interesting. For present purposes, we sketch the basics of the full stimulus model in this section, and leave these other interpretation issues aside until the discussion section.

In the observer model, to be able to recognize faces at a level above chance, a particular observer should experience higher levels of "familiarity"[2] when viewing faces he/she has seen before than when viewing novel faces. In this case, each individual point in the probability density function represents the degree of familiarity elicited by a *single face* when it is viewed by the observer. In other words, the *noise distribution* on the left of Figure 1 is comprised of faces the observer has never seen before and the *signal + noise distribution* is comprised of faces the observer has seen before. By contrast, the signal detection theory model applied to stimuli is based on the assumption that the degree of familiarity experienced by observers when viewing a *particular face* for the first time is discriminable from the degree of familiarity experienced by observers when viewing this face for a second time. Thus, for a face to be recognizable at a level above chance, on the average, observers should experience lower levels of familiarity viewing the face for the first time, than when viewing it for a second time. The familiarity experienced by observers viewing the face for the first time is represented in the distribution on the left of Figure 1, whereas the distribution on the right represents the familiarity experienced by observers viewing the face for the second time. Each data point in each probability density function represents the degree of familiarity experienced by a *single observer* viewing the face. In other words, the *noise distribution* on the left is comprised of observers who have never seen the face before and the *signal + noise distribution* is comprised of observers who have seen the face before.

[2] Care must be taken in interpreting *familiarity* in this context. No connection is claimed between this rather abstract and unspecified dimension and more precise definitions offered other papers, e.g., Bartlett, Hurry & Thorley (1984) and Vokey & Read (1992).

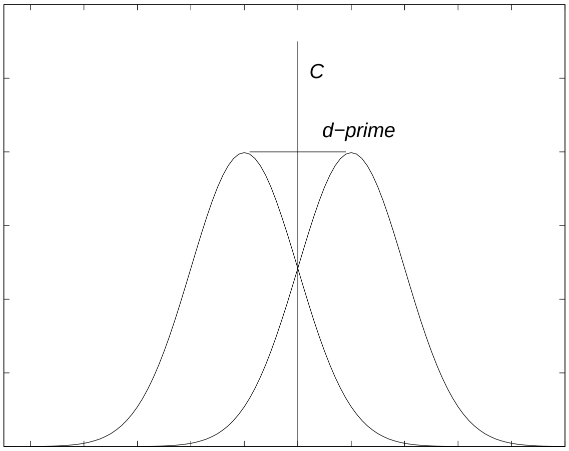

FIGURE 1. A classic signal detection model, with $d' = 2$ and $C = 0$.

The computation of the d' and C for the stimulus model proceeds analogously to the computation for the observer model. In the observer model, each hit that contributes to the hit rate, and each false alarm that contributes to the false alarm rate comes from a different face. In the stimulus model, each hit that contributes to the hit rate, and each false alarm that contributes to the false alarm rate comes from a different observer.

Applied to the observer model, d' refers to how accurately the observer recognizes faces in the particular experimental condition, whereas C gauges the bias of the observer to respond "old" versus "new". Applied to a face stimulus, this model is intuitive and interpretable as follows. The discrimination index or d' gauges the recognizablility of the face, (i.e., how good people tend to be at recognizing this particular face). Generally, d' is thought to reflect the characteristics of the individual face, such as whether or not it has distinctive features, (e.g., a mole, buck teeth, etc). The criterion gauges the tendency of the face to evoke "old" versus "new" responses from observers in a particular experimental condition. The criterion reflects both the characteristics of the individual faces and the characteristics of the experimental context. Both factors might work together as follows. A male face with long hair may evoke many "old responses" in a task in which long-haired males comprise 80 percent of the learned faces used (e.g., recognition of rock stars), but may evoke many fewer "old" responses when long-haired males represent only a small minority of the faces (e.g., Wall Street brokers).

3. SIMULATIONS

To carry out the simulations that follow, we generated a data base of "face models" that varied systematically in their characteristic d' and C. A face model is simply a signal detection theory representation of a single face in an experimental condition, and is specified by a d' and C. We then sampled from this data base in different ways and analyzed the samples to address our questions.

The construction of the data base proceeded as follows. We generated a "matrix" of signal detection models that varied systematically in d' and C. This is depicted schematically in Figure 2, which shows a sampling of signal detection models with d' increasing across rows, and criterion increasing across columns. While d' can vary computationally from negative infinity to positive infinity, in practical terms, successful psychological experiments are those that set up the task requirements to avoid ceiling and floor effects. The simulations we report here are indeed sensitive to the d' and criterion range chosen. To make these simulations as meaningful as possible, we have choosen parameter values that are commonly encountered in these kinds of experiments. For d', we chose only positive values, varying from .50 to 2.5, in increments of 0.1. The criterion varied in this matrix from -1 to $+1$, also with an increment parameter of 0.1. This yielded a 21×21 matrix similar in form to that displayed in Figure 2.

In summary, each "element" of this matrix can be thought of as an hypothetical face stimulus model that could correspond to a particular face in some experimental condition of a yes/no recognition experiment. Highly recognizable faces are represented toward the bottom part of the matrix, and less recognizable faces toward the top. Faces recognized with loose criteria are on the left side of the matrix and faces recognized with with stricter criteria are on the right side of the matrix.

From the signal detection models specified, we next computed a matrix of the hit rates these models yield, and a second matrix of the false alarm rates these models yield.

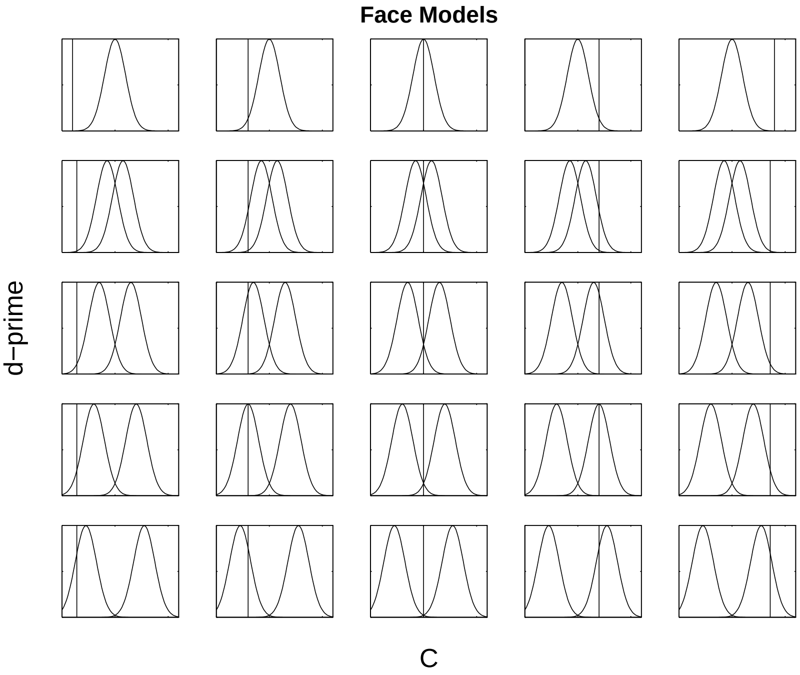

FIGURE 2. Schematic representation of the signal detection models created, with d' increasing by row, and criterion becoming stricter by column.

The first set of simulations results in variations of the kind of information captured in families of Receiver Operating Characteristic (ROC) curves. Specifically, we present the results that occur: 1.) when one limits oneself to a range of commonly found d' and C scores, and more importantly, 2.) when different sampling distributions of face models are included. We wished, first, to extract from this representation, the correlations between hit and false alarm rates that one would obtain when these different distributions underlie the face models included in a standard face recognition study.

3.1. Exhaustive Distribution Simulations.

3.1.1. Baseline Statistics. .

In this first analysis, we did the simplest thing possible. We correlated the hit and false alarm rates for all of the face models in the matrix. This is an exhaustive sample of the faces. This sample yields a correlation of 0.589 between hit and false alarm rate. The pattern of covariance is illustrated in Figure 3 and is simply a repesentation of a complete set of ROC curves for the d' and C range considered.

3.1.2. One dimensional variation of face sampling distribution.

In this exercise, we simply divided up Figure 3 into the parts caused by variation in the d' and and the parts caused by variation in the C.

Stable Criterion. .

The case that is perhaps most commonly assumed or hoped for in face recognition experiments, is a case in which the criterion remains more or less constant and only d', or face recognizability, varies meaningfully. It is obvious that if this is the case, regardless of the value of d', the correlation between hit and false alarm rate is high and negative, with r peaking at -1.0 when the criterion is stable at 0. This is easy to see intuitively if we look back at Figure 1. With the criterion stable at 0, imagine pulling apart the signal and noise distributions symmetrically about this zero point. One obtains a perfect negative correlation between hits and false alarms. Figure 4 illustrates iso-criterion functions for the range of criteria between -1 and 1. For the range we included, the correlations between hit and false alarm varied from $-.9696$ to -1.0.

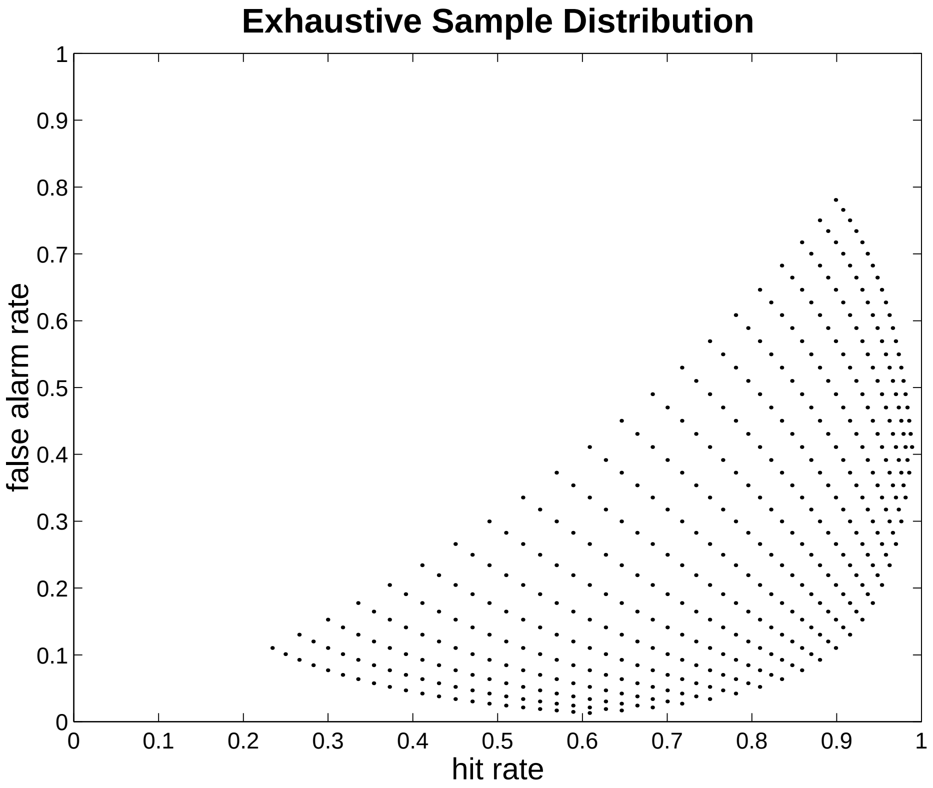

FIGURE 3. Hit versus false alarm rates in an exhaustive sample distribution of face models. Each point represents a possible combination of hit and false alarm rate. The correlation between hit and false alarm rates across the entire sample is equal to 0.589.

Stable d'. .

The complementary possibility is a stable d' with meaningful variation in only the criterion. Although theoretically possible, the extreme version of this is an unlikely situation given the commonly reported finding that face ratings and d' correlate. This indicates that d' must vary at least somewhat. Nonetheless, for completeness, we present the iso-discrimination lines in Figure 4. These iso-discrimination lines clearly produce correlations in the *opposite* direction to those obtained with a stable criterion. For the d' and C range tested, the correlations for this stable d' case ranged from 0.847 to 0.99.

3.2. **Random Normal Distribution Simulations.** In this series of simulations, we generated random normal, distributions that varied in the standard deviation of the d' and C about some mean value. We have already seen in the previous simulations that a completely stable criterion produces high negative correlations between hit and false alarm rates, and that a completely stable d' produces high positive correlations between hit and false alarm rates. In this simulation we investigate the middle ground of moderate variations of d' and C. Additionally, rather than using an exhaustive sampling distribution, in which all d' and C values in the matrix were equally probable, we sampled these values from normal distributions. This seems more realistic as a sampling assumption.

We began by generating random normal sampling distributions $n = 100$ from the face models created previously. These distributions were centered on a mean $d' = 1.0$ and $C = 0.0$. We varied only the standard deviation of the sampling distributions. These standard deviations varied from 0, i.e., completely

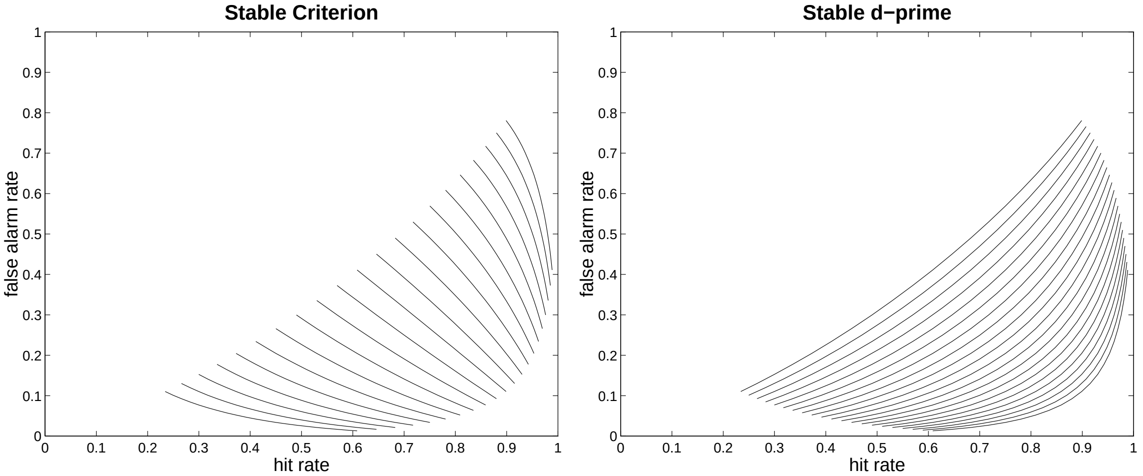

FIGURE 4. In the left figure are the iso-criterion lines, and in the right figure are the iso-discrimination lines. Within iso-criterion lines, highly negative correlations between hit and false alarm rate are found. Within iso-discrimination lines moderately high positive correlations between hit and false alarm rate are found.

stable and equivalent to the above "extreme" situations, to .5, in .1 z-score increments. The standard deviation for the d' and C components of the distribution varied in all possible combinations, yielding 36 different face sampling distributions (e.g., $\sigma_{d'} = 0$ and $\sigma_C = 0$;[3] $\sigma_{d'} = 0$ and $\sigma_C = .1$; $\sigma_{d'} = 0$ and $\sigma_C = .2 \ldots \sigma_{d'} = .5$ and $\sigma_C = .5$).

We present the results of these simulations in two figures. First, by way of summary, Figure 5 illustrates the correlations between hit and false alarm rates for the face samples as a function of the standard deviation of d' and C. The figure does not include the line for a stable criterion at 0, because we have already illustrated that this condition yields a perfect correlation of -1.0.[4] As can be seen from Figure 5, the correlation between hit and false alarm rate varies widely as a function of the stability of d' and C in the sampling distribution of faces in the experiment. Further, it can be seen clearly that the correlations are generally lower (i.e., less strongly positive, more strongly negative) for relatively smaller variability of the C, and generally larger for relatively smaller variability of the d'.

The same data are illustrated more graphically in Figure 6, which shows scatter plots for hit and false alarm rates as a function of the standard deviations of d' and C in the sample. Each individual scatter plot represents a face sampling distribution, with hit rate on the x-axis and false alarm rate on the y-axis. As can be seen, the manipulation of face distribution parameters has a strong effect on the covariance relationship between hit and false alarm rate.

The simple conclusion from these data is that the mechanics of signal detection theory, in conjunction with the mean and variability of the face sample parameters, can yield widely different correlations between hit and false alarm rate. The range of expected correlations found across the samples we examined here spans from a perfect inverse correlation to a perfect positive correlation. Also, while we consider this issue more thoroughly in the discussion, with the exception of the extreme cases of zero variability, we consider all of the points on the graph in Figure 5 to be quite plausible face sampling distributions for a standard face recognition experiment. Thus, in the absence of other evidence, these factors should be considered to provide the most parsimonious account of any particular obtained correlation between hit and false alarm rate.

[3] As noted in Figure 5, this first variation is meaningless because it defines only one face model.

[4] We do not display the line for $C = 0.5$, but it continues the pattern.

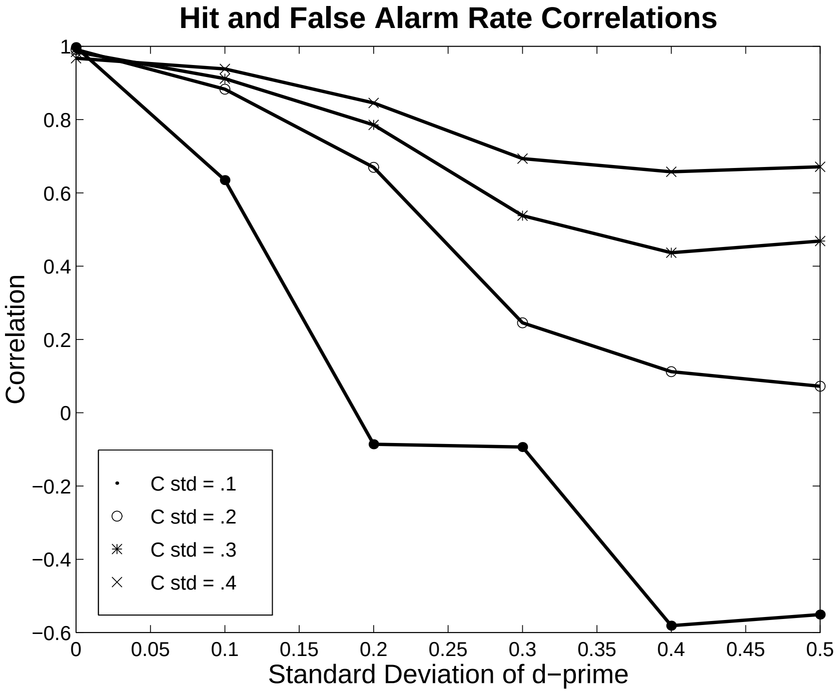

FIGURE 5. Hit versus false alarm rate correlations as a function of the variation in the standard deviation of the face sampling distribution about the mean $d' = 1.0$ and mean $C = 0.0$. As can be seen, increasing the standard deviation for C produces increasing correlations between hit and false alarm rates, whereas increasing the standard deviation for d' produces decreasing correlations between hit and false alarm rate.

3.3. **Nonmodel Measure Simulations.** It is common practice in the face memory literature to correlate particular kinds of errors, such as false alarms or hits with other, non-signal detection-based measures of the stimuli (e.g., rated typicality, rated attractiveness). We have argued elsewhere (O'Toole, et al., 1994) that the interpretation of such correlations can be problematic. Primarily, the problem stems from the fact that a false alarm or hit rate by itself is uninterpretable without knowing the d' and C of the face model. So, for example, if a false alarm rate is greater for face A than for face B, three interpretations are possible: 1.) face A is less recognizable than face B; 2.) observers tend to use less conservative criteria with face A than with face B; and finally, 3.) some unknown combination of 1.) and 2.). Although often the researcher may not care theoretically which of these interpretations is correct in a particular condition of an experiment, very serious interpretation problems can occur when the correlations are made and compared among two or more conditions of an experiment that differ in either their mean d' or C, or in the standard deviations of these. Often such systematic differences in the recognizability of the faces in the different conditions of an experiment are both predicted and obtained.

We carried out two kinds of simulations here. In the first, we work from synthesized correlations between an "other measure" and hit (false alarm) rate, with the aim of examining how these synthesized correlations constrain the expectations for : a.) correlations between this other measure and d'; and b.) for correlations between this other measure and C. For convenience, we will refer to this other measure, generically, as "facial-rating". In the second kind of simulation, we work from synthesized correlations between a facial rating measure and d' with the aim

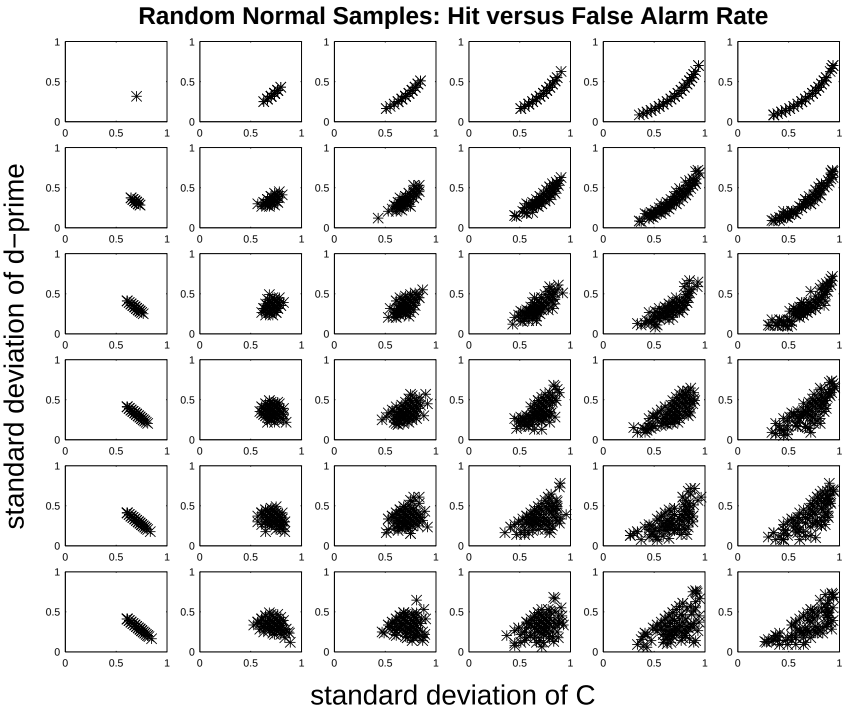

FIGURE 6. Scatter plots for hit and false alarm rates as a function of the standard deviations of d' and C in the sample. The plots show all 36 combinations of the d' and C standard deviations, with the top left corner showing the meaningless case of zero variability for both d' and C and the lower right corner showing the hit and false alarm rate relationship for standard deviations of 0.5 for both d' and C.

of examining how these synthesized correlations constrain the expectations for correlations between the rating measure and hit rate and between the rating measure and false alarm rate.

Correlations between Facial Ratings and Hit or False Alarm Rate. In this simulation, we used random normal samples of faces, and synthesized a facial-rating measure that correlated with either the hit or false alarm rate of the sampled data.

The simulation proceeded as follows. For each correlation level[5], we sampled 100 faces from a random normal distribution with a mean $d' = 1.0$ and a mean $C = 0.0$, (see Section 2.2). Based on the hit and false alarm correlation data presented in Figure 5, we set the standard deviation for the d' sampling distribution to .3 and the standard deviation for the C sampling distribution to .1, values which yielded approximately zero correlation between the hit and false alarm rates.[6]

Next, we generated a facial-rating measure for the 100 faces. Facial-rating vectors were created to correlate to varying degrees with false alarm rate by : 1.) sampling 100 numbers from a normal distribution with a mean of zero and a standard deviation of σ, where σ varied probabilistically from 0 to .495 in 0.005 steps;[7] and by 2.) adding one of these 100 sampled values to each one of 100 false alarm rates for the faces sample. For each standard deviation condition, this yielded 100 pairs of false alarm rates

[5] Note that these levels vary probabilistically, not in precise intervals.

[6] In fact, the correlation was a bit more negative as indicated in Figure 5. We chose the approximately zero correlation as a correlation that has been reported previously in a recent face recognition study (Hancock et al., 1996).

[7] By probabilistically, we mean that distributions with mean zero and each of the tested σ values were created and that these distributions were sampled randomly.

and facial-rating values. Across standard deviation conditions, the correlation between false alarm rate and facial rating varied probabilistically as a function of σ, with larger σ's yielding lessor correlations and smaller σ's yielding stronger correlations. We then computed the associated correlations between d' and facial-rating and between C and facial-rating.

The results of this false alarm analysis appear on the left half of Figure 7. These results are somewhat difficult to unpack intuitively. As a guide, however, it is perhaps easiest to start at the extremes. For example, beginning at the right hand side of the figure on the left, the far right points in this figure indicate the case where there is a nearly perfect positive correlation between the facial-rating and false alarm rate. In this case, because the facial rating value equals the false alarm rate, the asterisks at the right extreme of the x-axis reduce to the correlation between false alarm rate and d', which is strong and negative. Specifically, high false alarm rates generally indicate low d's, though again we must recall that we are operating above ceiling and below floor in the performance ranges considered. Likewise, the open circles on the right extreme of the graph represent the correlation between false alarm rate and C, which is also strong and negative. Specifically, high false alarm rates generally indicate low or loose C.

To aid in interpreting the less extreme data points, it is convenient to pretend that our facial rating is face typicality. Many studies have found a moderately positive correlation of approximately .50 between typicality and false alarms. To obtain such a correlation with the present face sample, (i.e., operating within the specified ranges of d' and C), the data displayed in the left side of Figure 7 indicate that the source of the correlation is very likely to comprise *both* a negative covariance relationship between typicality and d' and a negative covariance relationship between typicality and C. In short, at the .50 correlation label on the x-axis both the correlation between d' and facial-rating and the correlation between C and facial-rating are non-zero. In other words, it is very unlikely *with this face sample* that a false alarm rate-typicality correlation of this magnitude could be due *only* to variations in the recognizability of the faces — criterion variation must also play a role. Concommitently, it is very unlikely with this face sample that a false alarm rate-typicality correlation of this magnitude could be due *only* to variations in face criterion — recognizability variation must also play a role.

The important point made by this illustration is that the constraint is the nature of the face model sample. It is highly unlikely to obtain a correlation of .50 between false alarm rate and "typicality" that has, as its sole source, recognizability. This is because C varies in the face model sample too much relative to the variation of d' in the sample. A very different result can be obtained by changing the sampling parameters to tighten the standard deviation of C relative to d'. To illustrate, we re-ran this simulation changing σ_C from 0.1 to 0.02. The results of this simulation appear on the right side of Figure 7. Here we see that it quite likely to get recognizability as the sole source for the correlation between false alarm rate and facial rating. Again, although this scenario of a stable C is the case that researchers may hope for and perhaps implicitly assume in face recognition experiments, it is perhaps not a realistic assumption.

For the complementary part of this simulation, we repeated the above methods but synthesized the facial rating correlation from the hit rate, rather than from the false alarm rate. The results appear in Figure 8. Not surprisingly, the pattern of results is rather different. Again, however, the main point is that moderate correlations between hit rate and a facial rating cannot generally be due to variations in only the recognizability or criteria of faces — the variation of both factors is likely to be involved. We leave more detailed interpretations of these data to the reader, and proceed to the more basic conclusions.

First, correlations between our synthesized facial-rating measure and hit (false alarm) rate constrain, in a probabilistic fashion, the magnitude and direction of correlations between the facial-rating measure and both d' and C. This is not surprising in that all of these measures are codependent. The particular form of these constraining functions is not as important as the point that the mechanics of the signal detection model mandate that such functions exist. Indeed the form of these functions will vary with the parameters of the face sample.[8] Second, the shape of

[8] We carried out this set of simulations also with a standard deviation for the d' sampling distribution to 0.3 and the standard deviation for the C sampling distribution to .3, and got similarly shaped functions with somewhat different slopes.

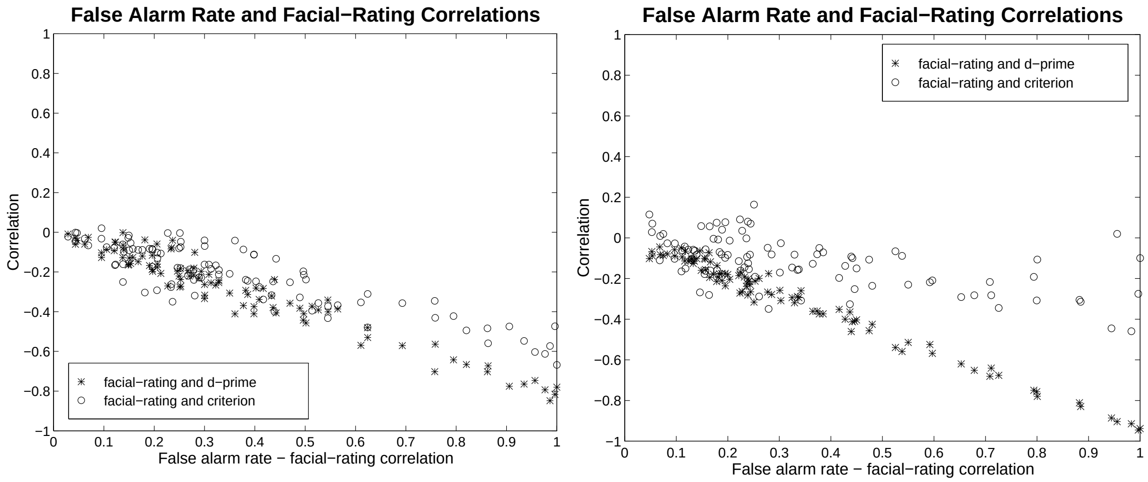

FIGURE 7. Both figures show the correlations between face rating and d', and face-rating and C that are obtained with correlations of varying strengths between false alarm and face rating. The right and left figures differ only in the standard deviation of the C value used, (left figure $\sigma_C = .1$; right figure $\sigma_C = .02$.)

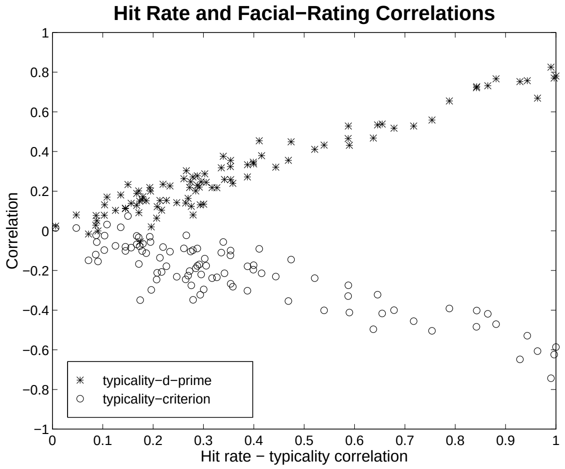

FIGURE 8. The figure shows the correlations between face rating and d', and face-rating and C that are obtained with correlations of varying strengths between hit rate and facial rating.

the constraining functions differs markedly for correlations between facial-rating and hit rate and for correlations between facial-rating and false alarm rate. Finally, viewed in terms of the source of the correlations, (i.e., d' and/or C), correlations measured on hit (false alarm) rate, can be less ambiguously interpreted in the context of the signal detection model.

Correlations between Facial Ratings and d'. In this final simulation, we used the procedures described previously, with only the following change. We created the synthetic correlation between d' and facial rating. Additionally, before proceeding, we made a small change to the σ_C and $\sigma_{d'}$ distribution parameters to enable us to make a more direct comparison to some questions raised by the Hancock et al. (1996) paper. They found correlations between a facial rating and both hit and false alarms, in the absence of a correlation between hits and false alarms. As noted, the parameters previously chosen from Figure 5 actually produce slightly negative correlations (see Figure 5). We adjusted these parameters to values not discretely on the Figure 5 graph, to obtain correlations between hit and false alarm rate as close to zero as possible. One set of suitable parameters turned out to be $\sigma_C = .09$ and $\sigma_{d'} = .18$.

The results of these analyses appear in Figure 9. We mention only a few basic points. First, we think these graphs offer a way out of the apparent paradox Hancock, et al. (1996) considered. They found that the rated facial distinctiveness correlated both with hit and false alarm rate, but that hit and false alarm rate did not correlate with each other. Such a situation is clearly theoretically possible with correlation, but confounds intuition a bit. This situation arises throughout Figure 9. Specifically, at virtually all points along the x-axis, there is a nonzero correlation between false alarm rate and the facial rating and between hit rate and facial rating, and yet, as noted, no correlation between the hit and false alarm rate (as specified via the simulation standard deviation parameters).

Although the results of this simulation are difficult to unpack intuitively, we again looked to a more extreme case to make the point. On the right side of the x-axis, are simulations with very high synthesized correlations between d' and facial rating. As such, the individual asterisks and open circles on this side of the graph reduce to the correlations between hit rate and d' and false alarm rate and d', respectively. To get a closer look, we isolated one of these simulations, which yielded a relatively high correlation between false alarm rate and facial-rating, and hit rate and facial-rating, but no correlation between hit and false alarm rate. We then looked at a three-dimensional scatter plot of the hit rate, false alarm rate, and facial rating values for the sampled faces. These data are illustrated in Figure 10. The different graphs in the figure display the same cloud of points from 6 different views. Beginning in the first row, we are viewing this cloud from directly overhead so that we can see the hit verus false alarm data. It is evident from this viewpoint that the three-dimensional points "projected" onto these two dimensions form a circular structure that yields a 0 correlation between hit and false alarm. This is due, again, to the face sampling distribution parameters chosen (i.e., the relative variability of d' and C). The center graph in the second row displays these data viewed from a lower elevation. Moving rightward, the viewer can get a better look at the positive correlation between hit rate and facial rating, with the best view in the right-most graph, which hides the false alarm rate dimension. Moving leftward, the viewer can see the negative correlation between false alarm rate and facial rating, with the best view in the left-most graph, which hides the hit rate dimension.

These data, thus, illustrate a relatively simple scenario for resolving the apparent paradox reported in Hancock et al. (1996). When the face sampling distribution is such that the variations of C and d' are somewhat balanced, even substantial correlations between hit rate and distinctiveness, and false alarm rate and distinctiveness, are possible in the absence of a correlation between hit and false alarm rate. Interestingly, the intuitions of Hancock et al. are indeed correct *when the variability of d' far outweighs the variability of C*. As noted previously, this seems to be the case most commonly hoped for, and sometimes implicitly assumed in face recognition studies. To illustrate the important difference this assumption makes, we repeated this last focused simulation with parameters from Figure 5 that yield a high negative correlation between hit and false alarm rate (i.e., cases where the variability of d' substantially outweighs the variability of C). Specifically, we took values $\sigma_C = .1$ and $\sigma_{d'} = .35$, which yielded a correlation of $-.42$ between hit and false alarm rate. We next synthesized a "distinctiveness" rating that correlated with d' to a degree of .61. In summary, in this relatively stable criterion scenario, we obtained : 1.) a positive correlation between hit rate and distinctiveness, $r = 0.51$; 2.) a negative correlation between false alarm rate and distinctiveness, $r = -0.51$; 3.) a positive correlation between d' and distinctiveness, $r = 0.61$; and 4.) a negative correlation between hit and false alarm rate, $r = -0.42$. In summary, when the variability of d' outweighs the variability of C, it is highly likely that negative correlation between hit and false alarm rate will result, and concommitantly, that these quantities will correlate in opposing directions with a stable entity such as a facial-rating.

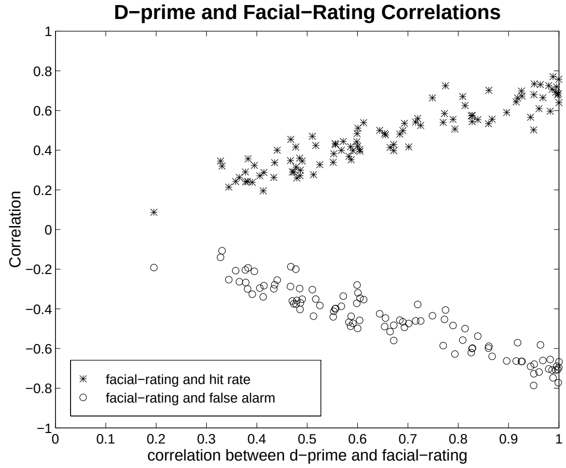

FIGURE 9. The figure shows the correlations between face rating and hit rate, and face-rating and false alarm that are obtained with correlations of varying strengths between face rating and d'.

4. DISCUSSION

The stimulus model that most researchers in face recognition use is based on signal detection theory, for which hit and false alarm rates are assessed to define the recognizability and response bias associated with individual faces in an experimental condition. Considered across all of the stimuli in a particular experimental condition, the nonlinear mechanics of signal detection theory limit the utility of intuition for determining the expected covariations of hits and false alarms with each other, and also, the expected covariation of hits and false alarms with *other* stimulus-based measures. Notwithstanding, the mechanics of signal detection theory in combination with stimulus sampling parameters constrain these expected covariations in a more or less knowable way. The present results indicate that the interpretation of these covariation data is critically dependent in a number of ways on the sampling distribution parameter constraints.

The present simulations remind us that with a signal detection model applied to describing the behavior of individual stimuli in an experiment, both d' and C comprise the sampling parameters of the face distributions. Although the variation in the recognizability of faces is generally the focus of stimulus-based research hypotheses (e.g., prototype and face-space models), the variation of C has equally potent consequences for the interpretation of a number of kinds of commonly reported data in face recognition studies. In this discussion, we consider, in turn, the implications of variation of the stimulus sampling parameters at two levels of analysis in psychological experiments : a.) simple sampling variability of d' and C for the faces selected to serve *within* a particular condition of an experiment, and b.) systematic variations of d' and C *between* the conditions of an experiment.

There are at least two concrete implications of the present data for the *within* condition variation of d' and C. First, no specific correlation between hit and false alarm rate should be expected in any given experimental condition without reference to the sampling distributions of d' and C associated with the condition. It follows, therefore, that any particular obtained correlation between hit and false alarm rate, cannot be considered sufficient evidence for a theoretical claim. Thus, while there is intuitive theoretical appeal in believing that face errors of different kinds should be related, (e.g., faces that evoke high levels of hits should also evoke low levels of false alarms), the stimulus model used in the vast majority of face recognition experiments does not support such intuitions. In fact, the statement that "faces that evoke high levels of hits should also evoke low levels of false alarms" is synonymous with the statement that "the criterion for all faces in an experimental condition is the same". As noted previously, we believe that this

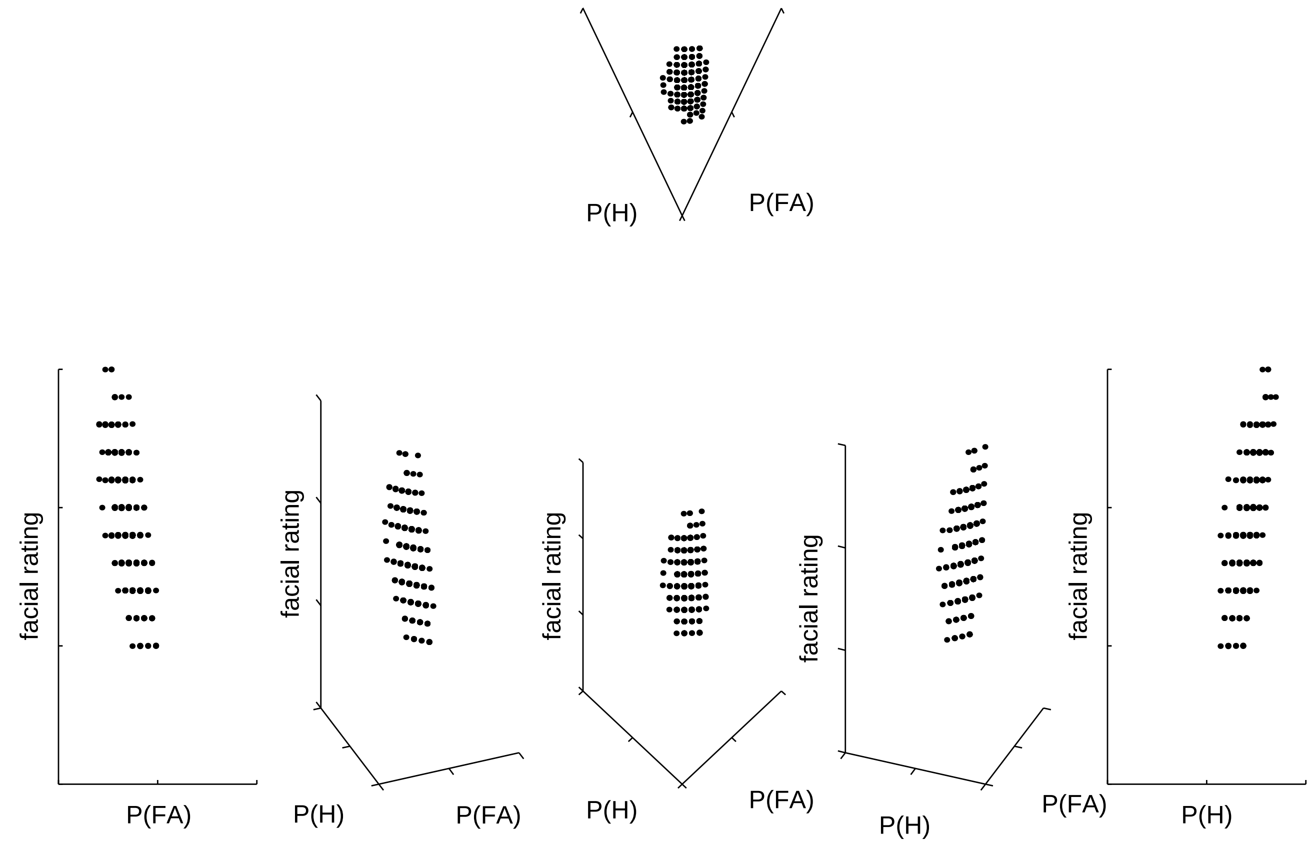

FIGURE 10. Five views of the *same* three-dimensional scatter plot of hit rate, false alarm rate, and facial rating. In the first row center, the view is from overhead and shows the lack of correlation between hit and false alarm rate. The second row center is a view from a lower elevation. Across the second row, the view changes from right to left of center, with the left-most plot showing the negative correlation between facial rating and false alarms, and the right-most plot showing the positive correlation between facial rating and hit rate.

assumption is implicit in the reasoning of many researchers computing the recognizability of individual stimuli in the context of a signal detection model.

Applied to past work, these findings indicate that the lack of correlation between hits and false alarms reported by Hancock et al. (1996) and Lewis and Johnston (1997) may not *need* an explanation. One would need to look at the distribution of both the *d'* and *C* values to know what to expect for these correlations and then to verify the extent to which the correlations deviate from expectations. This however, does not indicate that we cannot find out more about how individual faces "act" in the context of an experiment when they are known (potentially a hit or miss) versus novel (potentially false alarm or correct rejection). Interestingly, the data of Lewis and Johnston are informative in that they demonstrate quite clearly that there are additional important experimental-context factors that may contribute to these covariances. Lewis and Johnston show that the consistency of false alarms depends not only on the faces used as distractors but on those used as *targets* as well. The single stimulus model of signal detection theory is only a measurement tool, and thus cannot by itself be used to explain these context effects.

The second concrete implication of the within condition variation of d' and C is that although it is a very common practice in the face literature to correlate false alarm and hit rate with "other" face measures, such as typicality, attractiveness, and memorability, the interpretation of such correlations is not straightforward. For example, Lewis and Johnston (1997) found a significant correlation between distinctiveness and hits but no correlation between distinctiveness and false alarms. At the same time, the authors found that a measure of personal familiarity of faces, if analyzed on an individual subject basis, correlated reliabily with false alarms but not hits. Understandably, given this apparent "double dissociation", the authors considered the possible information processing mechanisms selectively underlying the hits and false alarms, respectively. This was done by assessing other aspects of the hits and false alarms that might give insight into these processes. Though this may be a reasonable strategy in some cases, we argue below that there may be a more parsimonious explanation of these kinds of findings.

It is beyond the scope of this paper to delve into the details of the particular processing mechanisms advanced in the Lewis and Johnston (1997) paper. Rather, we present only two interpretive/methodological caveats concerning the treatment of correlations involving hits and false alarms. First, before theorizing about apparent process dissociations involving the hit and false alarm measures, one should examine the pattern of correlations using the discrimination and bias measures directly. We expect that in many instances, correlations involving d's and C's will support more coherent theoretical accounts than the hit/false alarm-based correlations. Correlations done at this level, however, may not be significant if the simultaneous variation of both d' and C contributed to the orignally significant correlation between the "other" measure and either the hit or false alarm rate. This is not necessarily problematic if one views the signal detection measures as an analytical tool for identifying the sources of interesting and applicable results concerning, for example, false recognitions (Schacter, Norman, & Koutstaal, 1998). As such, the covariance of a facial rating and either hit or false alarm rate may be relevant in the eyewitness identification literature, where the effects are important regardless of their source. We would argue, however, that in many cases, including applied studies, it is still important to understand the extent to which these effects are due to discrimination versus criterion changes.

Second, our simulations in this paper lead us to worry that the apparent dissociations involving hit and false alarm rate measures may prove unstable across experiments due to irrelevant variations in the stimulus sets used. It follows then that the correlations between the false alarm or hit rate and other measure may be similarly unstable. For example, contrary to the majority of other work in the field (Bartlett, et al. 1984 ; Hancock et al. 1996; and Light et al, 1979), Lewis and Johnston (1997) find that distinctiveness was correlated more strongly with hits than with false alarms. This might be due to random differences in the items, as the procedure used by Lewis and Johnston was similar to the procedures used in the previous studies. In any case, the simulations presented here lead us to *expect* differences of this sort when the recognizability and bias properties of the stimulus set are changed.

The major implication of variation of d' and C *between* experimental conditions concerns comparisons of correlations obtained in the different conditions, e.g., correlations between false alarm rate and typicality in two conditions of an experiment. Specifically, there are serious interpretation difficulties for correlations between an "other" measure and either hit or false alarm rate when they are compared *across conditions* in an experiment. Any sort of systematic effect of condition on d' or C strongly compromises the validity of these kinds of false alarm or hit rate comparisons. This includes manipulations that affect either the mean or variability of either d' or C. Thus, it may be possible to obtain no main effects of the independent variable and still be at risk in interpreting across condition correlations (e.g., when the variability of d' or C are not comparable in the conditions compared).

Condition-based effects on d' are frequently the basis of experimental predictions. And, although not often predicted *a priori*, condition-based effects on C have been found also (O'Toole, Edelman, & Bülthoff, in press; Valentin, 1996) and have been interpretable in the context of the experiment. In the experiments of O'Toole et al. and Valentin, when the observer's task was to recognize faces across a large change of viewpoint (90 degrees), both studies found that observers used very strict criteria (much stricter than they used with smaller changes in viewpoint). This yields stricter criteria for the face models in this large viewpoint change condition relative to those in the smaller viewpoint change conditions.[9] One interpretation of the strict criteria used in this "difficult" condition is of a meta-memorial nature. Observers, perhaps, did not feel very confident in their own abilities, which may have made them more generally conservative on this task. (See also Hintzman, 1994, for a discussion of the potential complexity of the effects of meta-memorial influences on recognition data). The main point is that systematic variations of both d' and C across the conditions of an experiment are possible, and may occur for a number of theoretical reasons. In the presence of these between condition variations, comparisons of false alarm or hit rates (or any correlation measure that includes these measures), confound the role of d' and C in the comparisons.

Mixed Metaphors. We suggest that a number of the confusions in face recognition studies that consider individual stimulus measures can be traced to some tempting, but unsupported, "mixed metaphors" of the observer and stimulus models. As noted, the model construct of signal detection theory can be applied equally validly to the analysis of either stimuli or observers in a particular condition. However, there is no valid, formal way to relate the results of these two analyses to each other. Some common ways of thinking about this, nonetheless, are evident in a number of papers in the face literature. We argue here that these implicit arguments contain various mixed metaphors of the observer and stimulus models. Some of these mixed metaphors contain a grain of truth, and others lead to circularities in reasoning that cannot be supported with empirical data. We try here to sketch out common advantages and pitfalls in trying to conceptualize the stimulus and observer analyses together.[10]

First, correlation has occasionally slipped into the literature as an attempt to relate the stimulus and observer models implicitly, as follows. When one finds a correlation between face recognizability (i.e., d') and perceived typicality, it is tempting to imagine that the certain kinds of faces (e.g., highly typical faces) tend to "hang out together" in the old and/or new distributions (i.e., perhaps on the right side of both distributions or on the right side of one distribution and on the left side of the other distribution). This might explain why certain faces kinds of faces attract more false alarms than hits. Specifically, one might imagine that certain kinds of faces are sitting at some particular place in the distribution, and hence have differential probabilities of being hits or false alarms. The flaw in this reasoning is that a signal detection model comprised of individual faces is *necessarily* a single observer model, and yet, the correlation we are trying to understand (between d's and typicality) is based on *many stimulus models.* More to the point, using the whole set of observers in the experimental condition, each face in the observer's old and new distribution can be said to have its own d' and a C, as well as its own hit and false alarm rate. But, there is no *formal* relationship between the position of the faces in the observer distribution and the face's d' as computed across the observers. Indeed different data contribute to these computations. So, even if it is tempting to imagine that the correlation one obtains between d' and typicality on *face models* is based on the face's clustering in the *observer distributions*, it is tenuous at best and can lead to empirically unsupportable conclusions.

[9] We will discuss this observer-face link in the next section as part of mixed metaphors problem.

[10] It is perhaps worth a brief reference to the classic paper of Clark (1973) on the "language as a fixed-effect fallacy", which is relevant by analogy to the present issues. Clark argues convincingly that not only the observers, but also the stimuli in an experiment vary meaningfully. Clark's (1973) concern, however differs from ours in that he was interested in the implications of the nature of stimulus variation for the validity of the inferential statisics applied to the data. Our concern is in relating descriptive measures on a single side of the analysis — the stimulus side.

It is worth noting explicitly that it is never possible to figure out exactly where a particular face "sits" in a signal detection-based observer model; nor, is it possible to figure out exactly where a particular observer "sits" in a signal detection-based face model. A face in an observer's model has exactly one of four possible states: it is either a hit, false alarm, correct rejection or miss. This allows one to locate it into one of these four regions of the model, but no more precisely than that.[11] This is completely counter-intuitive for many of us who are used to dealing with both stimulus and observer-based data from face recognition experiments for the following mixed metaphorical reason. A face *can* be said to have a hit and a false alarm rate, but it is impossible to find the face in the *signal* or *signal + noise* distributions because again we are (mentally) in the wrong model. The face's hit and false alarm rate define its *stimulus-based* signal detection model, not its position in the observer model.

Second, an alternative method to correlating facial rating measures and d' is often implicit in the reasoning advanced in some ways of discussing facial rating data and face recognizability. For example, the standard interpretation of a correlation between d' and typicality is that "distinctive faces are recognized more accurately than typical faces". This correlation is based on the computation of a d' for each face (across a large number of observers) and the assessment of each face's typicality (again across a large number of observers). It is equally easy to imagine face typicality (e.g., as pre-assessed by observers) as an independent variable in an experiment with two discrete levels : "typical" and "unusual". In this case, the signal detection model for observers, but not for faces, changes entirely from the model we have been considering. Specifically, whereas individual face models are the same as those we have described previously, for the observer model, depending upon one's theoretical assumptions, two models are possible. The simplest would consist of a single *noise distribution* comprised of novel faces, and two *signal + noise distributions*: one comprised of typical faces and the other comprised of unsual faces.

A second possible observer model emerges by implementing some interesting assumptions about the way a single prior exposure differentially affects subjective familiarity for typical versus unusual faces (cf., Bartlett et al., 1984; Mandler, 1980). The assumptions, as stated in Bartlett et al. are as follows: a.) all novel faces elicit non-zero levels of familiarity, but this familiarity is greater for typical as opposed to unusual faces; and b.) the increment in familiarity that results from a single prior exposure is greater for unusual than for typical faces. Combined, these assumptions indicate that it is theoretically invalid for the typical and unusual faces to *share* a *noise distribution*. It is further worth noting that under this model, the analysis of the observer data are analogous to those that have been well-studied in reference to the mirror effect (Glanzer & Adams, 1990), though to our knowledge, no one has ever published a mirror effect for face recognition.

The two cases just described are completely valid observer models, but dissociate the stimulus and observer models in an important way. This dissociation has to do with the assumption of a discrete categorical structure for typical versus unusual faces. We would argue, however, that these models may be somewhat less appropriate than the single continuous model, due to the likelihood that faces are distributed in a continuous rather than in a discretely bimodel fashion with respect to the dimension of face typicality. In any case, when stimulus and observer models dissociate in this fashion it becomes even more difficult to reason back and forth between the stimulus and observer model perspectives.

Finally, although there is no formal way to put together data from the stimulus and observer models, at least one mixed model metaphor is not only valid, but worth keeping in mind. In all experiments, we all know that observers actually do the responding. So, even if one analyzes a particular stimulus, one is actually only measuring something about observers' response patterns to this stimulus. In embedding this stimulus into the context of an experimental condition comprised of like stimuli (e.g., upside-down faces), we have the possibility, and indeed hope, that all or most observers who participate in this condition will behave in a similar fashion. For example, in

[11] It is well-worth noting that our limited ability to locate a face in an observer's signal detection distribution (i.e., as only a hit, false alarm, miss or correct rejection), has no implications whatsoever for locating a face in an observer's face space. The methods required to compute an observer's face space are based on similarity judgements (cf., Johnston, Milne, Willams & Hosie, 1997) among pairs of faces and not on recognition data. Thus, the methodological points made in this paper are not by themselves relevant for evaluating face space models.

a "difficult" experimental condition, observers may lack confidence in their ability to perform well and may respond in a conservative or cautious fashion for all stimuli. This will yield strict criteria for *both* the individual observer measures and for the individual face measures in this condition. This is a good thing for the inferential analysis of the experimental data, which contrary to the advice of Clark (1973), nearly always proceeds only on observer measures. On the other hand, as we will argue below, variations in the "performance" of individual faces, in the context of representational issues, may be worthy of study in their own right.

We suggested in the introduction that a major difference between observer- and stimulus-based measures in the face recognition literature concerns the use to which the measures are put. This leaves us with a somewhat different perspective on the issue of stimulus-based measures than that evidenced in the Clark (1973) paper. Specifically, in recent years, the study of the stimulus has taken on a great deal of importance as a way of trying to understand the perceptual constraints imposed on human information processing by the richness of the environment. What is the information that is available in the human face for specifying its gender, race, age, and identity? How is this information represented in the brain? Are faces represented by their two-dimensional image-based features or by object-centered three-dimensional features?

Different models of the information in faces and of the human representation of this information make different predictions about which *individual* faces should be easy to recognize; easy to classify by gender, etc. Studying the recognizability and classifiability of individual faces by human observers, in conjunction with a computational model of the representation of faces, provides a very much under-explored reserve of constraints on theories of face processing. These analyses have been undertaken in recent years and have provided a number of useful insights into the complexity of the information in human faces and the ways in which observers make use of this information under various task demand situations (Hancock et al., 1996, O'Toole, et al., 1994; O'Toole et al., in press). Such analyses hold out the possibility of sorting through questions about the nature of representations of faces and objects that cannot be similarly tackled by relying only on observer measures. Thus, in spite of the methodological pitfalls involved in reporting and interpreting data on individual faces, we believe that these analyses are well-worth the trouble.

References

[1] Bartlett, J. C., Hurry, S. & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition, 12*, 219-228.

[2] Clark, H. H. (1973). The Language-as-a-Fixed-Effect Fallacy : A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

[3] Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence : Can we infer anything about their relationship? *Law and Human Behavior, 12*, 41-55.

[4] Glanzer, M. & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology : Learning, Memory & Cognition, 16*, 5-16.

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

[6] Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing : Human perception and principal components analysis. *Memory & Cognition, 24*, 26-40.

[7] Hintzman, D. L. (1994). On explaining the mirror effect. *Memory & Cognition, 20*, 201-205.

[8] Hintzman, D. L., Caulton, D. A. & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology : Learning, Memory, & Cognition, 20*, 275-289.

[9] Johnston, R. A., Milne, A. B., Willams, C. & Hosie, J. (1997). Exploring the structure of multidimensional face-space : The effects of age and gender. *Visual Cognition, 4*, 39-57.

[10] Lewis, M. B. & Johnston, R. A. (1997). Familiarity, target set and false positives in face recognition. *European Journal of Cognitive Psychology, 9*, 437-459.

[11] Light, L. L., Hollander, S., & Kayra-Stuart, F. (1981). Why attractive people are harder to remember. *Personality and Social Psychology, 7*, 269-276.

[12] Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Memory and Learning, 5*, 212-228.

[13] Mandler, G. (1980). Recognizing : The judgment of previous occurence. *Psychological Review, 87*, 252-271.

[14] O'Toole, A. J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & Cognition, 26*, 146-160.

[15] O'Toole, A. J., Deffenbacher, K. A., Valentin, D. & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition, 22*, 208-224.

[16] O'Toole, A. J., Edelman, S. & Bülthoff, H. H. (in press). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research.*

[17] Pigott, M.A., Brigham, J.C., & Bothwell, R.K. (1990). A field study of the relationship between quality of eyewitnesses' descriptions and identification accuracy. *Journal of Police Science and Administration, 17*, 84-88.

[18] Schacter, D. L., Norman, K. A. & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*, 289-318.

[19] Valentin, D. (1996). *How come when you turn your head I still know who you are? Evidence from computational simulations and human behavior.* Unpublished doctoral dissertation, The University of Texas at Dallas.

[20] Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A*, 161-204.

[21] Valentine, T., & Bruce, V. (1986). Recognizing familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology, 40*, 300-305.

[22] Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition, 20*, 291-302.

Appendix A Observer-based Signal Detection Model in a Face Recognition Experiment

The signal detection model in experimental psychology is, by far, applied most commonly to the task of measuring or describing the behavior of a *single observer*. Applied to the description of a human observer's performance on a face recognition task, the model is based on the assumption that the observer can discriminate learned and novel faces based on some abstract subjective dimension. In the case of face recognition, this dimension is often thought to represent the degree of recollection or familiarity[12] an observer experiences when viewing each face in the recognition test. This dimension is symbolized by the x-axis on the sample signal detection model that appears in Figure 1.

For an observer to perform a face recognition task at a level above chance, known faces must, on the average, elicit higher levels of familiarity than novel faces. Thus, novel versus known faces are represented by the left and right distributions, respectively, in Figure ??. The *noise distribution* represents the familiarity values elicited when the observer views a population of faces for the first time (see distribution on the left of Figure 1). The *signal + noise distribution* represents the familiarity values elicited when the observer views a population of faces known to him/her (see distribution on the right of Figure 1). Thus, each data point in each probability density function represents the degree of familiarity experienced when the observer views a *single face.*

Under this model, the "performance" of each observer in a face recognition task can be described completely in terms of a discrimination index, termed d', and a response bias or criterion, termed C. Both of these measures are computed directly from the hit and false alarm rate of the observer in the face recognition task. The hit rate is defined as the proportion of old or learned faces to which the observer correctly responds "old". The false alarm rate is defined as the proportion of novel faces, to which the observer incorrectly responds "old". More formally, the discrimination index is defined as :

$$d' = z(P_H) - z(P_{FA}) \tag{1}$$

where $z(P_{FA})$ denotes the *z-score* for the false alarm rate and $z(P_H)$ denotes the *z-score* for the hit rate. This discrimination index measures the degree of overlap between the two distributions. More precisely, d' is simply the distance, in *z-score* units, between means of the the *noise* and *signal + noise* distributions (see Figure 1).

The response bias measure is defined as a different function of the $z(P_{FA})$ and $z(P_H)$:

$$C = -\frac{1}{2}[z(P_{FA}) + z(P_H)] \tag{2}$$

In practical terms, observers respond "old" to faces that elicit familiarity levels higher than the criterion, (i.e., to the right of the criterion, see Figure 1). Negative values of C indicate loose or liberal criteria, or a bias to respond "old", whereas positive values of C indicate strict or conservative criteria, or a bias to respond "new".

Applied to face recognition by a human observer, this model is intuitive and readily interpretable as follows. The discrimination index refers to the observer's ability to discriminate known from unknown faces and is considered to be a response bias-free measure of face recognition accuracy. This discrimination index is thought to reflect the *characteristics of the individual observer* such as his/her visual

[12] Care must be taken in interpreting *familiarity* in this context. No connection is claimed between this rather abstract and unspecified dimension and more precise definitions offered other papers, e.g., Bartlett, Hurry, & Thorley (1984) Vokey & Read (1992).

and perceptual abilities, memory capacity, motivation, and experience with the task. The criterion measure is thought to reflect both the *characteristics of the individual observer and the characteristics of the experimental situation.* The former include inherent aspects of the observer's personality (e.g., liberalness/conservativeness of their guessing strategy), and the latter include task demands. Task demands might include factors like different reward contingencies for hits versus false alarms, whereas experimental context might include the proportion of faces that are actually old versus new in the recognition test.