Measuring single-item identification efficiencies for letters and 3-D objects

Ami Eidels · Jason Gold

Abstract Identification thresholds and the corresponding efficiencies (ideal/human thresholds) are typically computed by collapsing data across an entire stimulus set within a given task in order to obtain a "multiple-item" summary measure of information use. However, some individual stimuli may be processed more efficiently than others, and such differences are not captured by conventional multiple-item threshold measurements. Here, we develop and present a technique for measuring "single-item" identification efficiencies. The resulting measure describes the ability of the human observer to make use of the information provided by a single stimulus item within the context of the larger set of stimuli. We applied this technique to the identification of 3-D rendered objects (Exp. 1) and Roman alphabet letters (Exp. 2). Our results showed that efficiency can vary markedly across stimuli within a given task, demonstrating that single-item efficiency measures can reveal important information that is lost by conventional multiple-item efficiency measures.

Keywords Identification · Signal detection · Ideal observer · Threshold · Efficiency

In psychophysics, a threshold in pattern perception tasks such as detection and identification is often defined as the minimal amount of stimulus energy necessary for an observer to achieve a criterion level of performance, such as 75 % correct or a d' of 1 (Cornsweet, 1970; Green & Swets, 1966). In many psychophysical tasks, it is also possible to measure the performance of a statistically optimal or "ideal" observer (Geisler, 1989; Tanner & Birdsall, 1958). The ideal observer makes optimal use of all available stimulus information by performing a computation and using a decision rule that will maximize average accuracy. The comparison of human to ideal performance allows one to measure the statistical efficiency with which a human observer uses the information that is available in a given task (Tanner & Birdsall, 1958).

Ideal-observer analysis has proven to be a useful tool for separating the effects of psychological and physical factors on the human ability to detect, discriminate, and identify different kinds of patterns (e.g., Geisler, 2003). However, one shortcoming of this approach has been that, for tasks involving multiple exemplars (e.g., detection of a set of faces or identification of a set of objects), thresholds and corresponding efficiencies are computed by collapsing the data across the entire set of stimuli. This approach assumes that all of the stimuli being tested are processed with equal efficiency. For some tasks, such as the detection of highly similar, localized patterns, this may be a reasonable assumption. But it is less clear how much the efficiency of information use will vary across more complex and dissimilar patterns.

In this article, we attempt to address this shortcoming by proposing a simple technique for measuring thresholds and efficiencies for individual items. We begin by clarifying the definition of *efficiency* within the ideal-observer approach, and then outline the proposed technique. Next, we describe two experiments in which we applied our technique to the identification of two different classes of complex patterns: 3-D computer-rendered objects and Roman alphabet letters. The technique is explicated in great detail in the Method section of Experiment 1 (especially in the procedure description), where it is conveniently presented within a concrete experimental context. We conclude by pointing out the advantages, and limitations, of the new technique.

A. Eidels (✉)
School of Psychology, University of Newcastle, Callaghan,
NSW 2308, Australia
e-mail: ami.eidels@newcastle.edu.au

J. Gold
Indiana University, Bloomington, IN, USA

Efficiency

Efficiency in any given task, η, is defined as the ratio of the squared human to the squared ideal sensitivity (d')—that is

$$\eta = \frac{d'^2_{\text{human}}}{d'^2_{\text{ideal}}}. \tag{1}$$

An efficiency of 1 implies that the human observer is using stimulus information optimally. An efficiency of less than 1 implies that, for the human processing system, information is being lost somewhere between stimulus presentation and response measurement (Tanner & Birdsall, 1958). Because comparison of the human sensitivity to that of the ideal observer controls for differences in the availability of stimulus information, variations in efficiency across experimental conditions imply variations in the ability of the human observer to use the available information. In contrast, constant efficiency implies that the ability to use the available information is invariant, regardless of changes in performance (e.g., Banks, Geisler & Bennett, 1987; Gold, Bennett & Sekuler, 1999).

It is often difficult to compute efficiency as described above, because human sensitivity tends to be much lower than ideal sensitivity at any given stimulus level. This will result in ceiling performance for the ideal observer at any stimulus level at which the human observer performs above chance. The solution to this problem has typically been to choose a criterion level of d' (e.g., $d' = 1$) or of percent correct (e.g., 75 %) and to measure both human and ideal observer thresholds at this level of performance (Pelli, 1981; Tanner & Birdsall, 1958). Using this approach, efficiency is often expressed as the ratio of ideal to human energy threshold, E, at a criterion level of performance—that is

$$\eta = \frac{E_{\text{ideal}}}{E_{\text{human}}}. \tag{2}$$

Human efficiencies have been measured in this manner for a wide range of detection, discrimination, and identification tasks involving a variety of different stimuli, such as auditory tones and noise bursts (e.g., Green, 1960), sinusoidal gratings (e.g., Geisler, 1989), simple random dot patterns (e.g., Barlow, 1978) , objects and faces (e.g., Gold et al., 1999; Tjan, Braje, Legge & Kersten, 1995), letters and words (e.g., Pelli, Burns, Farell & Moore-Page, 2006; Tjan et al., 1995), and biological motion (Gold, Tadin, Cook & Blake, 2008).

Thresholds and the corresponding efficiencies in experiments such as those described above are typically measured by combining data across all items to be identified and computing a single summary threshold and the corresponding efficiency. However, this kind of analysis ignores variations in how efficiently a human observer makes use of information when processing individual items. That is, efficiency could

vary quite dramatically across items in a set of patterns, but this variation is lost in the standard measure of efficiency. This complication makes it somewhat difficult to interpret variations in efficiency across tasks and stimuli, because individual stimuli that are processed with relatively high or low efficiency within a set (i.e., outliers) could lead to a distorted and unrepresentative measure of efficiency. For example, Tjan et al. (1995) tested how efficiently humans use visual information to recognize simple 3-D objects. Computer-rendered images of a wedge, a cone, a cylinder, and a pyramid were presented as shaded objects, line drawings, small silhouettes, or large silhouettes. The average calculated efficiencies were 3.28, 2.69, 7.84, and 4.51, respectively, for each of these rendering conditions. However, from these data, one cannot tell whether the variations in efficiency across the different rendering conditions were produced by individual items within each stimulus set that were processed with relatively greater or lesser efficiency, or whether the efficiency measures are representative of all of the items with each condition. This same complication applies to most experiments that involve the measurement of efficiencies for sets of dissimilar stimuli (e.g., Gold et al., 1999; Liu, Knill & Kersten, 1995; Pelli et al., 2006).

Outline of the proposed technique

In many psychophysical studies, the information necessary to measure the contributions of individual items to overall efficiency is simply not available, due to the experimental design. Typical threshold-finding experiments often employ an adaptive staircase procedure, which finds an observer's threshold by systematically adjusting the stimuli across trials in order to locate the stimulus level that yields a criterion level of performance. In principle, it is possible to use a staircase for each individual item and to measure the identification thresholds for separate items. However, to obtain a given level of performance, items that are easily detected (or identified) require a lower signal-to-noise ratio than do items that are more difficult to detect/identify. Consequently, if such a procedure were used to adjust the contrast of each item individually, the amount of contrast energy necessary to reach threshold performance would become informative and might serve as a cue for an item's identity. For example, in a hypothetical letter identification experiment, the amount of contrast energy required to correctly identify the letter "A" might be smaller than that required to identify the letter "B." The observer can then correctly identify the presented item on the basis of its contrast energy, rather than its form.

Here, we develop a different approach for measuring single-item efficiencies in identification tasks. The technique is based on the method of constant stimuli, rather than a staircase, and is therefore immune to the problem articulated

above. For a given task (here, object recognition or letter recognition), we use a staircase procedure to initially measure a psychometric function for the entire set of n items (a "calibration" phase). We then use this information to generate a fixed set of s stimulus levels that span the threshold range across all n items. We present all n items for t trials at each of the s stimulus levels, which are randomly intermixed throughout testing. The same stimulus levels are used for all items, to eliminate stimulus level as a possible cue to item identity. We then fit individual-item psychometric functions and compute corresponding thresholds for each stimulus by conditionalizing the analysis according to item identity. Individual item efficiencies are then computed by comparing the human and ideal thresholds for each item.

We demonstrate this technique in two experiments. In the first, we used a set of six 3-D rendered objects. In the second, we used a complete set of 26 Roman alphabet letters.

Experiment 1: 3-D objects

Method

Observers Three observers (two male, one female) participated in the experiment. Two (B.F. and R.S.) were naïve to the purpose of the experiment; A.E. was an author. All had normal or corrected-to-normal vision. Each observer completed the experiment within an approximately 1.5-h session.

Apparatus The stimuli were displayed on a Sony Trinitron Multiscan G520 monitor controlled by an Apple G4 computer running Mac OS 9.2.2. The monitor had a resolution of 1,024 × 768 pixels, subtending 16.4° × 12.4° of visual angle at the viewing distance of 130 cm. The frame rate was set to 85 Hz. The experiment was conducted in the MATLAB programming environment using the Psychophysics Toolbox extensions (Brainard, 1997). A Minolta Luminance Meter LS-100 photometer was used to calibrate the monitor, and a 1,792-element look-up table was built from the calibration data in order to linearize the display, as was described by Tyler, Chan, Liu, McBride and Kontsevich (1992). Luminance ranged between 0.7 and 103.6 cd/m², with an average luminance of 34.1 cd/m².

Stimuli Six different geometric objects were generated for the experiment: a sphere, a cube, a pyramid with a square base, a cylinder, a cone, and a square pyramidal frustum (a square pyramid truncated by a plane parallel to its base; see Fig. 1). To minimize reliance on size cues, the objects were constructed to have similar heights and widths. For instance, the lengths of the sides of the square bases were equal to the diameters of the circular bases. One two-dimensional projection was used for each object throughout the experiment. All of the objects

had the same angle of rotation away from the observer. Rotating the objects enhanced the illusion of 3-D depth.

All images were represented in contrast values such that the contrast (c_i) at pixel location i in an image was given by

$$c_i = \frac{l_i - L}{L}, \tag{3}$$

where L is the background luminance and l_i is the pixel luminance at location i in the image. The integrated contrast of a given image was computed using the root-mean-squared (RMS) contrast C_{RMS}, a quantity that is proportional to stimulus energy when squared (i.e., RMS^2 contrast). C_{RMS} is defined as

$$C_{\mathrm{RMS}} = \sqrt{\frac{1}{p} \sum_{i=1}^{p} c_i^2} \tag{4}$$

where p is equal to the total number of image pixels.

Noise Gaussian white contrast noise (with contrast defined as in Eq. 3) was added to each pixel of the signal shown on each trial. The value for each pixel in the noise field added to the signal was obtained from a Gaussian pseudorandom number generator with a mean of 0 contrast and a variance of 0.0625 (noise spectral density of 1.67×10^{-5} deg²).

Procedure Viewing was binocular, and a combination forehead- and chinrest stabilized the observer's head. The monitor supplied the only source of illumination during the experiment. A 1-of-6 identification task was used to measure performance, which was completed in one 1.5-h session. On each trial, observers were presented with a signal plus an external noise field mask on a background of average luminance. The display duration was approximately 500 ms. After signal presentation, the display was reset to the average luminance, and a selection screen with images of the possible signals was presented. Observers used the mouse to select the signal that they thought had been presented. After a selection was made, auditory feedback indicated whether the response was correct, and the display was reset to the average luminance prior to the beginning of the next trial.

The experiment consisted of two phases: an initial calibration phase, in which we used an adaptive staircase procedure, and a second, testing phase, based on the method of constant stimuli. The purpose of the initial calibration phase was to quickly measure a threshold for the set of stimuli as a whole. This threshold was then used as a starting point for generating a series of fixed contrast levels to be used with the method of constant stimuli in the second testing phase. Recall that adaptive staircases could not be used to estimate thresholds for each individual signal because of the possibility that observers
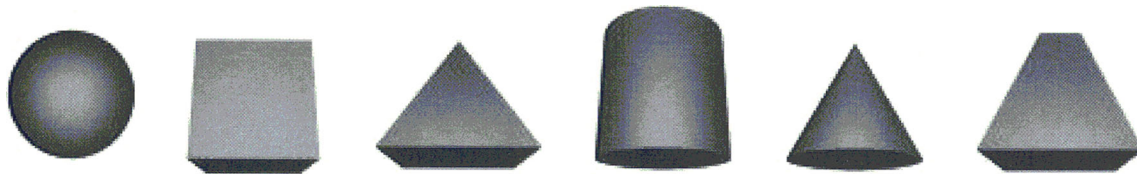
Fig. 1 Computer-rendered objects used in Experiment 1

might make use of the signal contrast itself to identify the signals.

At the beginning of the session, during the calibration phase, by means of the staircase procedure we manipulated the contrast of the signals across trials according to the observer's responses for 150 trials. The signal shown on each trial was randomly chosen from the set of six possible signals. The staircase tracked the 55 %-correct point on the psychometric function with a 1-down, 1-up rule (six alternatives were available, yielding a guessing rate of 16.7 % correct). A contrast threshold was estimated by fitting a Weibull function to the data. Threshold was defined as the RMS^2 contrast level corresponding to 55 % correct responses.

For the second, testing phase, the threshold measured during calibration was used as an anchor for generating six fixed contrast levels, equally spaced in log units ±1 log unit above and below the calibration threshold. The testing phase involved presenting each signal at each of the six contrast levels for 20 trials, yielding a total of 720 trials. All of the signals and contrast levels were randomly intermixed throughout the testing phase. The resulting data were sorted conditional upon signal identity, and Weibull functions were fit to the data for each signal in order to estimate the individual item thresholds. The reliability of each threshold was estimated by carrying out bootstrap simulations (Efron & Tibshirani, 1993). Specifically, we generated data for 200 simulated experiments by repeatedly drawing t random trial samples with replacement and fitting psychometric functions to each simulated set of data. We then computed the mean and standard deviation of the threshold estimates generated from this procedure.

Ideal observer The ideal decision rule for our tasks and stimuli can be derived using Bayes's rule (Geisler, 2003; Gold et al., 2008; Green & Swets, 1966; Tjan et al., 1995). For any given signal type (shapes or letters), observers were asked to determine the individual signal S_k (where k refers to the kth signal in the set of n possible signals) that was most likely to have appeared within the noisy stimulus data D. Note that for our stimuli, both S_k and D were vectors of contrast values. According to Bayes's rule, the a posteriori probability of S_k having been presented, given D, can be expressed as

$$P(S_k|D) = \frac{P(S_k)P(D|S_k)}{P(D)}. \tag{5}$$

For our tasks and stimuli, the prior probability of seeing any given signal, $P(S_k)$, and the normalizing factor $P(D)$ are both constants, and thus can be removed without changing the relative order of the probabilities. Thus, the ideal observer would choose the signal that maximized $P(D \mid S_k)$. In our experiment, we used the method of constant stimuli to measure observers' thresholds (staircases were only used to initially gauge roughly where the thresholds would be). Thus, the stimulus could be set to s different contrast levels on each trial, with equal probabilities. Given the conditions above, the ideal observer must compute this probability for all s possible contrast levels for each signal within the set and compute the summed probability across contrast levels, resulting in the following likelihood function:

$$P(D|S_k) = \sum_{j=1}^{s} \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(D_i - S_{ijk}\right)^2}, \tag{6}$$

where p is the total number of pixels in the stimulus and σ is the standard deviation of the Gaussian distribution from which the external noise was generated. The ideal decision rule would be to choose the signal S_k that maximized this function. Monte Carlo simulations of 6,000 trials per signal (30 contrast levels and 200 trials per level) were run to measure the ideal observer's contrast threshold for each signal, as well as the global threshold for the entire set of signals.

Results and discussion

The identification thresholds and efficiencies from Experiment 1 are summarized in Fig. 2. The top panel shows the individual-item identification thresholds for each of the human observers, as well as that for the ideal observer. Several interesting things can be noted about these data. First, the ideal observer's performance is not the same for all items. Instead, two of the shapes (the pyramid and wedge) were particularly difficult for the ideal observer, and one of the shapes (the sphere) was particularly easy. Because the ideal observer's performance was solely determined by the relative amount of information carried by each item, this means that the sphere carried the most information of all of the items and that the pyramid and wedge carried the least. It is worth emphasizing that the information carried by any item in the set is not a
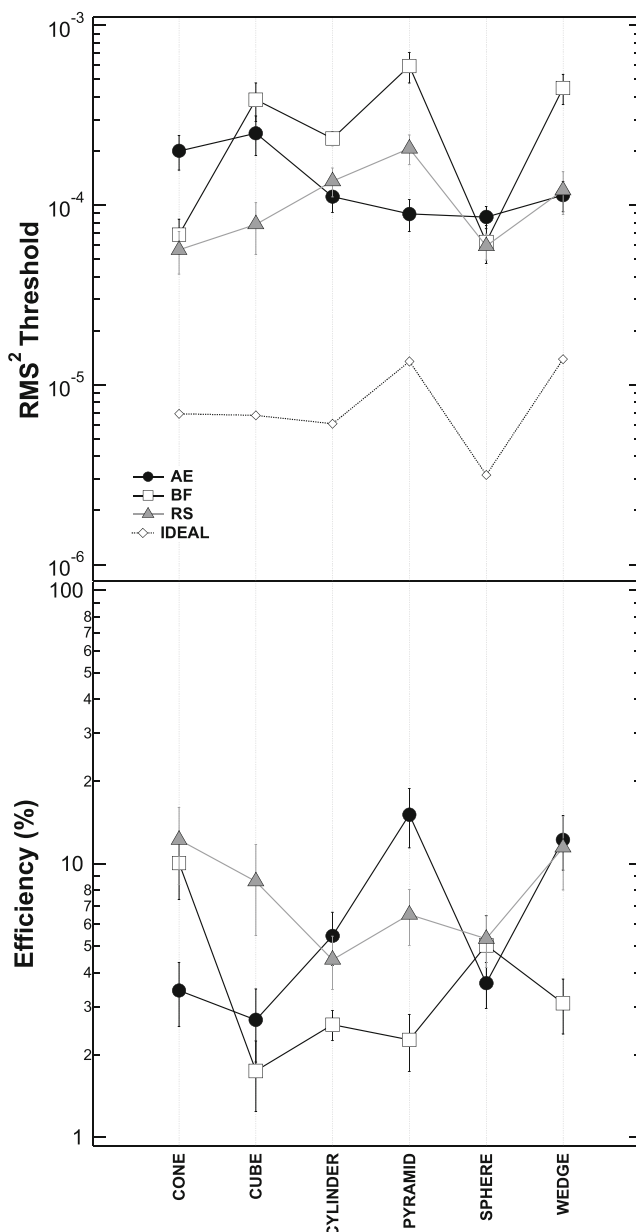
Fig. 2 Results of Experiment 1, presented on a log scale. The identification thresholds (top panel) and efficiencies (bottom panel) for each observer are presented separately for each of the stimulus items. Identification efficiencies were obtained by dividing the ideal observer threshold by the corresponding human threshold. Error bars represent ±1 standard deviation, estimated through bootstrap simulations (see the text for details)

context-independent quantity, but rather is determined by how discriminable it is from the other items in the set. That is, the amount of information carried by an item is defined by its physical dissimilarity to the rest of the items in the set, and is therefore an entirely context-dependent measure (Tjan & Legge, 1998).

Second, the human performance also varied markedly across items and was surprisingly idiosyncratic. Although two of the three observers (B.F. and R.S.) exhibited a pattern

that partially followed the ideal observer's performance (both had thresholds that were relatively high for the pyramid and wedge, and low for the sphere), the third observer showed a very different pattern of performance, with the pyramid, wedge, and sphere being nearly equal (and lowest) in threshold. Note that individual differences in thresholds cannot be accounted for by differences in the physical availability of information, and therefore must reflect differences in the ability to make use of the available information.

The corresponding efficiencies (ideal/human thresholds) are plotted in the bottom panel of Fig. 2. Computing efficiency factors out the differences in information content across stimuli and leaves a pure measure of the ability to make use of the information for each item. If the information content carried by the stimuli were the sole determinant of the variations in human thresholds, we would expect efficiency to be equal across items. But, as could be expected on the basis of the individual differences in thresholds, each observer exhibited her or his own distinctive pattern of efficiency across items. This result nicely demonstrates the utility of a single-item efficiency analysis: Individual differences in efficiency across the items in a set that would normally be obscured by a single, summary measure of efficiency are easily detected when each item is analyzed individually.

One potential concern about these data is that the variations in efficiencies across items and observers could easily be the result of response biases rather than true differences in processing efficiency. If an observer were biased to choose a particular item more frequently than others, this would erroneously inflate the percent correct for that item, as well as decrease threshold and increase efficiency. Similarly, it would reduce the percent correct, increase threshold, and decrease efficiency for the other items in the set. For example, consider the data from observer B.F. Constantly responding "cone" (whether B.F. was certain or uncertain that a cone was actually presented) would erroneously lead to the correct identification of a cone in low contrast, consequently resulting in a low identification threshold and high efficiency for this item.

To test whether response bias made a significant contribution to the observed efficiencies, we computed the correlation between the proportions of trials on which an observer made a particular item response and their corresponding efficiency for that item. The data for all three observers are shown in the right panel of Fig. 3, along with the best-fitting linear function. We found a moderate correlation between response frequency and efficiency ($r = .56$, $p < .01$), indicating that response biases were in fact present in the data. Given this result, we also computed the correlation between response frequency and threshold for the ideal observer, to gauge whether we might expect to find a systematic relationship between performance and response frequency that was intrinsic to the task and stimuli. Although the ideal observer's response frequencies varied far less across items that did the human observers',
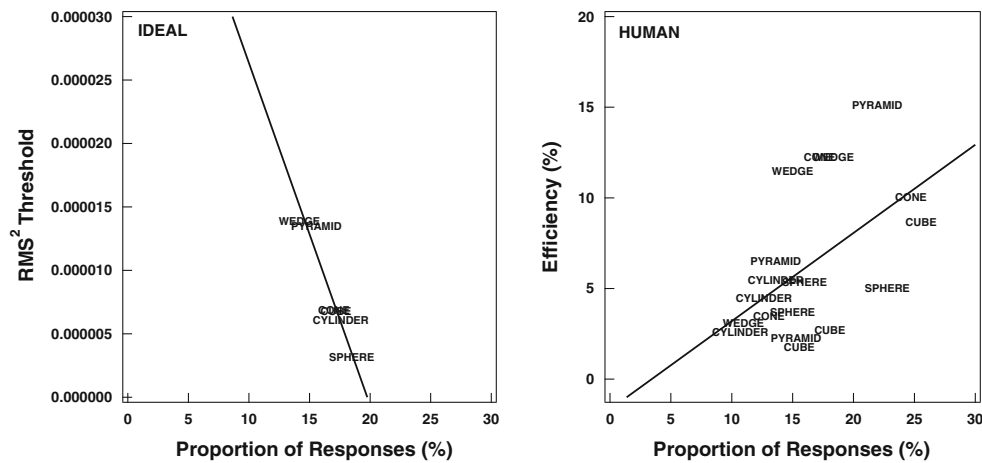
Fig. 3 Scatterplots showing the relationship between proportions of responses and either thresholds (for the ideal observer, left panel) or efficiencies (for the human observers, right panel) in Experiment 1. The solid line in each plot corresponds to the least-squares linear fit

we found a highly significant relationship between ideal response frequency and threshold ($r = -.96$, $p < .00001$). The negative relationship between response frequency and threshold for the ideal observer simply reflects the fact that threshold is inversely related to performance. The high correlation between response frequency and threshold for the ideal observer is most likely due to the fact that, for any set stimulus contrast values, the ideal observer will on average make more correct responses (and thus, more responses overall) to those items that are more intrinsically discriminable from the other items (i.e., the ones that have the lowest thresholds). Thus, the ideal-observer analysis indicates that we might expect response frequencies to vary across items in a systematic fashion for our human observers, due to the demands of the task and the experimental design. However, the ideal-observer analysis also indicates that we would expect the range of variation in response frequencies across items to be far narrower than we found for our human observers.

To address the possible problem of response bias in our data, we recalculated each observer's threshold for each item using a bias-free measure of sensitivity (d'), rather than percent correct performance (Green & Swets, 1966; MacMillan & Creelman, 1991). In signal detection theory, d' is a bias-free measure of sensitivity that typically represents the perceptual separation between the internal responses to two stimuli, relative to their variances. To compute d', one needs *hit* and *false alarm* rates. Here, we defined hits as correct identifications (say, responding "cone" when a cone was actually displayed) and false alarms as incorrectly identifying another item as item k (e.g., responding "cone" when a "cube" or "sphere" was presented). In computing d' in this way, we made an implicit assumption that comparisons amongst items were carried out along a single, unitary dimension (i.e., a single dimension that corresponded to "similarity to item k"). We computed d' in this fashion as a function of RMS

contrast, in order to generate psychometric functions that would not be contaminated by response frequency. For many tasks, d' has been found to vary approximately linearly with RMS contrast (Green & Swets, 1966; MacMillan & Creelman, 1991), and our data also followed this trend. As such, we fitted linear psychometric functions to each participant's data for each individual item and computed the contrast energy necessary to obtain a d' of 1.

The results of this reanalysis are plotted in Fig. 4. These data show that, even after correcting for stimulus complexity and possible response biases, efficiency still varied markedly across items and observers in our object recognition task. These results show that despite the presence of response biases, some items were in fact processed more efficiently than others, and that this variation in efficiency across items was highly idiosyncratic.

In Experiment 2, we decided to apply the same technique to the identification of a different class of stimuli, one that was highly overlearned and familiar: Roman alphabetic letters. We used Bookman uppercase font, since it has been found to be processed with relatively high efficiency, relative to other fonts (Pelli et al., 2006).

Experiment 2: Roman letters

Method

The participants and method were the same as those described in Experiment 1, with the following exceptions: The stimuli were the full set of 26 uppercase Bookman letters. The letters were all negative in contrast (i.e., darker than the background), 50 pixels tall, and placed in the center of a 128 × 128 pixel background. Each signal was presented 20 times at each of six contrast levels. With 26 letters, this yielded a total of 3,120
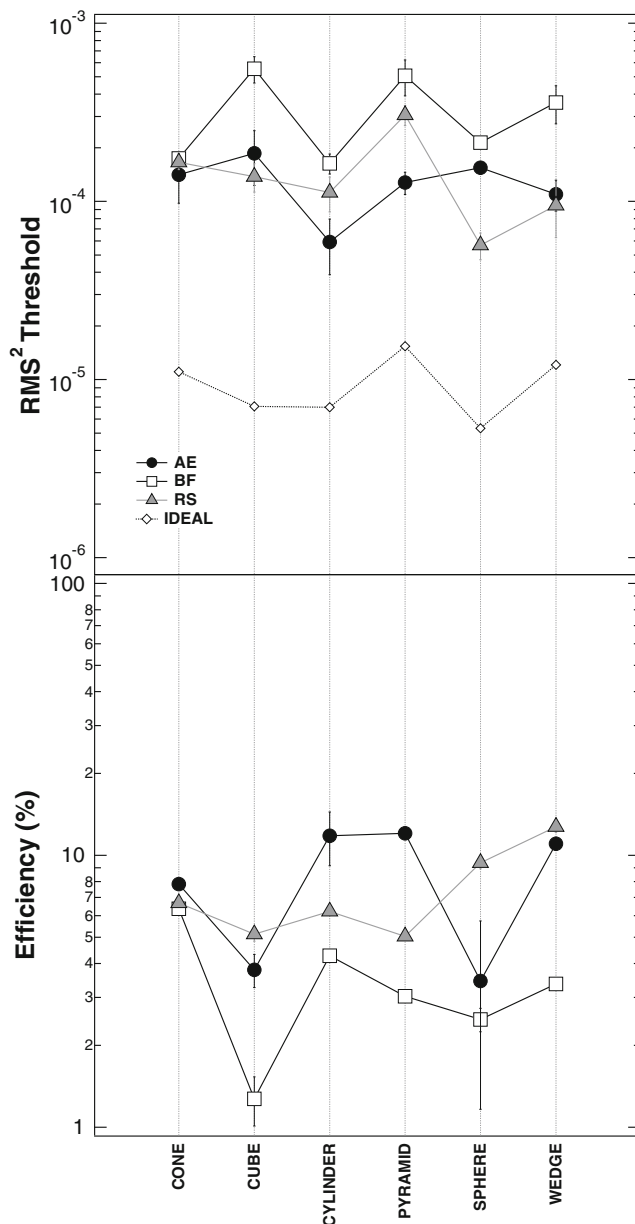
Fig. 4 Bias-corrected results from Experiment 1. The identification thresholds for individual observers, for each item, at $d' = 1$ are presented in the top panel. Efficiencies, obtained by dividing the ideal observer threshold by the corresponding human threshold at $d' = 1$, are presented in the bottom panel. See the text for further explanation of how d' was used to correct for bias

trials, so that each observer participated in four 1.5-h sessions instead of just one. Following the presentation of the signal, observers used the mouse to select the signal that they thought had been displayed.

Results and discussion

The results of Experiment 2 for uppercase Bookman font are presented in Figs. 5 and 6. Figure 5 plots human efficiency

(right panel, computed using non-bias-corrected thresholds for each observer) and ideal thresholds (left panel) as a function of the response proportions for each individual letter. As with the objects, we found a significant correlation between proportion of responses and both threshold for the ideal observer ($r = -.94$, $p < .00001$) and efficiency for the human observers ($r = .81$, $p < .0001$). As such, Fig. 6 plots only the biased-corrected thresholds (top panel) and the corresponding efficiencies (bottom panel). These data show that identification thresholds varied markedly across letters, as they had with the computer-rendered object displays. For example, across observers, the letters "C" and "I" had relatively low thresholds, as compared with the other letters, whereas "E" and "K" had relatively high thresholds. Note, however, that the ideal observer's thresholds also varied across letters, indicating that some letters in the Bookman font carry more discriminative information than others. For example, the ideal observer had a particularly low threshold for the letter "I." When we used the ideal observer's thresholds to measure efficiency and correct for the amount of information carried by each letter, we observed a somewhat different pattern of results: Letters such as "O," "S," and "Y" now stood out as being processed with relatively high efficiency, whereas the letter "I" actually had the lowest overall efficiency for two of the three observers. This result nicely illustrates how relatively better performance by a human observer (in this case, with the letter "I") does not necessarily imply a better ability to use the available information.

Inspection of Fig. 6 reveals that efficiency varied markedly across items for each observer. To confirm the main effect of item, we subjected the efficiency scores to a within-subjects analysis of variance,[1] and a significant effect of item (letter) emerged, $F(24, 48) = 2.221$, $p < .01$, $\eta_p^2 = .53$.

A close examination of Fig. 6 also suggests that similar patterns of relative between-item efficiencies might be present across observers. Therefore, we calculated the correlations between the efficiency scores across human observers, and found moderate relationships for all of the human pairwise comparisons. Table 1 shows the Spearman rank-order correlation coefficients for efficiency across observers (note that calculating correlations using Pearson's product–moment correlation coefficient yielded similar results). Thus, despite individual differences in performance, at least some of the cross-item variability in efficiency could be explained by the items' identities.

---

[1] The subsequent analyses take into account the efficiency scores for 25 items, rather than the entire 26 letters. We had to discard the efficiency scores for the letter "S," since it was not available for observer B.F. due to noisy data (for this item only).
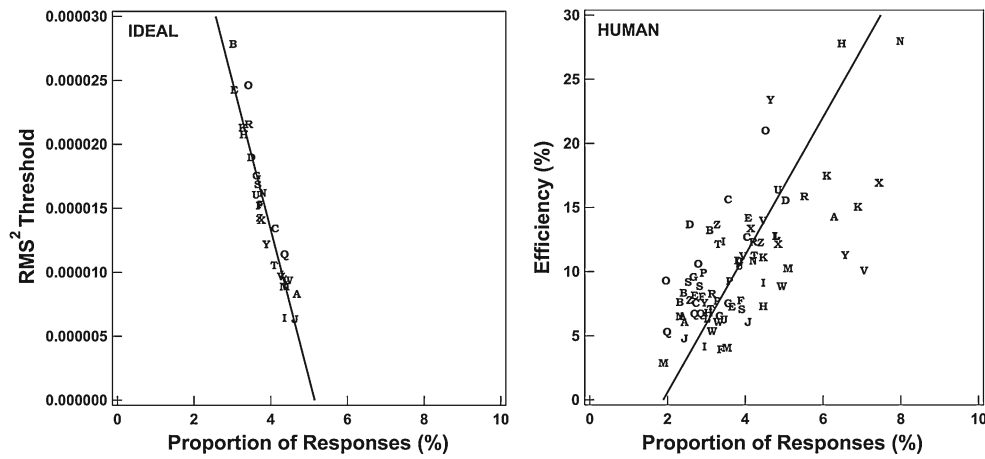
Fig. 5 Scatterplots showing the relationship between proportions of responses and either thresholds (for the ideal observer, left panel) or efficiencies (for the human observers, right panel) in Experiment 2. The solid line in each plot corresponds to the least-squares linear fit
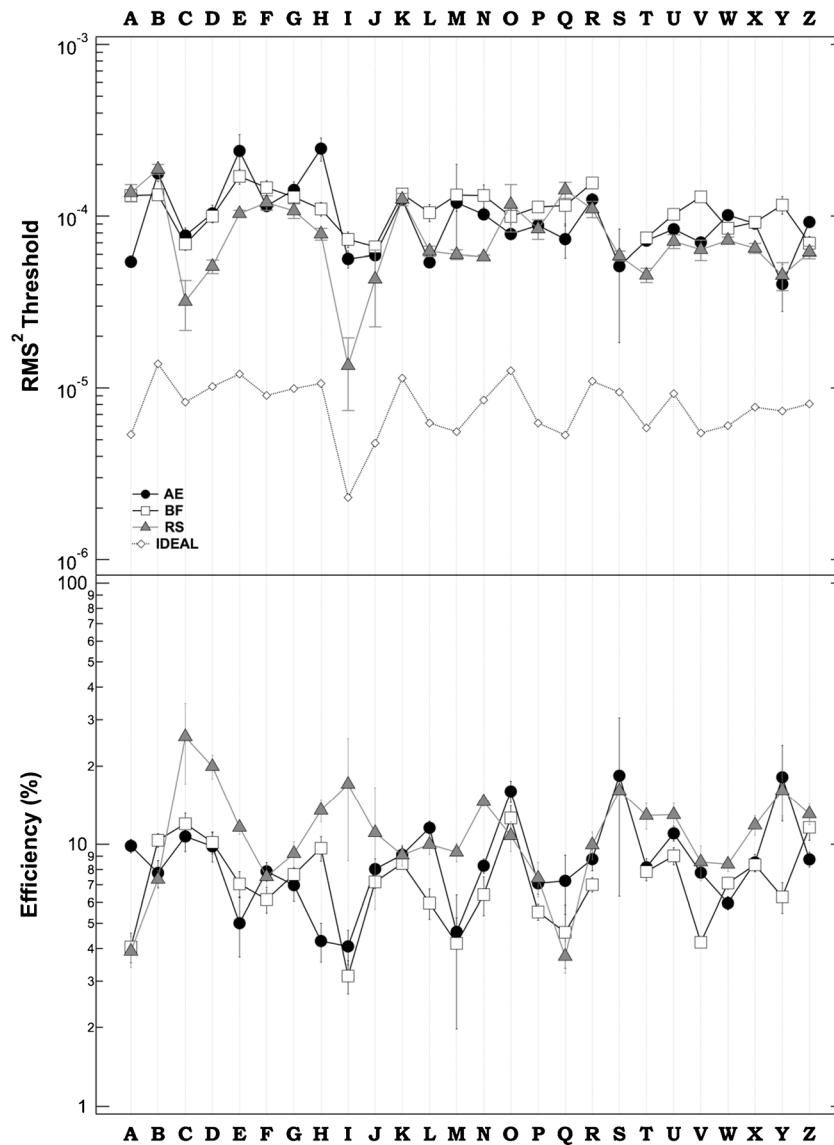

Fig. 6 Bias-corrected results from Experiment 2 (calculated at $d' = 1$; see the text for details)

Table 1 Spearman rank-order correlation coefficients for efficiency across observers

Observer	A.E.	B.F.	R.S.
A.E.	—	$r = .33^*$	$r = .30^†$
B.F.		—	$r = .38^*$
R.S.			—

† $.05 < p < .10$; * $p < .05$

The fact that efficiency for each observer was not constant across letters shows that physical similarity across letters cannot entirely account for the variations in human thresholds. However, we wished to explore whether physical similarity (as indexed by the relative performance of the ideal observer across letters) had any correlation at all with the patterns of human thresholds across letters. To do this, we calculated Spearman rank-order correlation between the thresholds of each human observer and the ideal observer. We found significant correlations for all three human observers [for A.E., $r = .64$, $p < .001$; for B.F., $r = .40$, $p < .05$; for R.S., $r = .41$, $p < .05$], showing that letter identification by our human observers was at least partially predicted by physical similarity across letters.

General discussion

In an identification task with several items, some stimuli may be processed more efficiently than others, and such differences are not captured by the conventional multiple-item threshold measurements. In this report, we have presented and applied a novel technique for measuring "single-item" identification efficiencies in object and letter identification tasks. The resulting single-item efficiencies describe the ability of the human observer to make use of the information provided by a single stimulus item, within the context of the larger set of stimuli. Our results show that efficiency can vary markedly across the stimuli within a given task (and across observers), demonstrating that single-item efficiency measures can reveal important information that is lost by conventional multiple-item efficiency measures.

So why, then, are certain items processed more efficiently than others? Many possibilities exist, ranging from a variety of low-level aspects of the stimuli, to higher-level aspects of the object features and identities. Although an extensive investigation of this issue is beyond the scope of this report, for the purposes of illustration, we considered two possible factors with respect to our letter identification data.

The first factor that we considered was the complexity of the form of the letters. More specifically, we considered *perimetric complexity*, which is defined as a letter's inside–outside perimeter, squared, divided by its ink area (Pelli et al.,
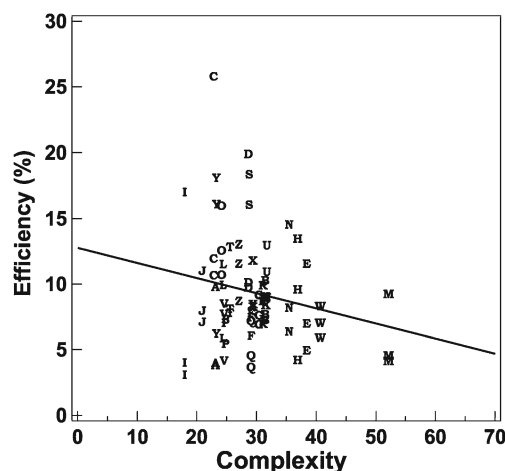

Fig. 7 Letter identification efficiency, plotted as a function of perimetric complexity (see the text for details). The solid line corresponds to the least-squares linear fit

2006). When averaged across characters within an alphabet, this measure has been shown to be highly negatively correlated with character identification efficiency across different alphabets and fonts, with more-complex character sets generally producing lower efficiencies (Pelli et al., 2006). Thus, we might expect that perimetric complexity might also be highly negatively correlated with the efficiency of processing individual characters within an alphabet. Figure 7 plots the efficiency for each letter in our letter stimulus set for our three observers as a function of perimetric complexity.[2] We found a small but significant negative correlation between efficiency and perimetric complexity ($r = -.2$, $p < .05$), suggesting that perimetric complexity may have played a role in determining relative letter identification efficiency for our observers.

The second factor that we considered with respect to our letter data was the relationship between efficiency and a letter's frequency of appearance in the English language. That is, high-frequency letters (e.g., "E," "T," and "A") may be identified more efficiently than low-frequency letters (e.g., "X," "Q," and "Z"). Frequency-related effects of this sort have been demonstrated in a variety of word recognition tasks, in which participants have been found to respond more rapidly and/or more accurately to more-common words than to less frequently occurring words (e.g., Grainger, 1990). To explore this possibility with our letter data, we calculated rank-order correlations between the letter identification efficiencies measured for each of the observers in our study and the frequency of letter occurrences in English (Lewand, 2000). We found no significant relationship for two of our three observers (for

---

[2] We computed the perimeter of each letter by identifying the edges within the letter image (using the Canny (1986) method of edge detection) and counting the number of pixels in which an edge occurred. We computed ink area by counting the number of pixels over which the letter image differed from the background of zero contrast. Perimetric complexity was then defined as perimeter²/ink area.

A.E., $r = .08$, n.s.; for B.F., $r = -.06$, n.s.), and only a marginally significant relationship for the third observer (R.S.: $r = .33$, $p = .05$). Thus, letter frequency does not appear to be a reliable predictor of efficiency.

The analyses described above are not meant to provide an exhaustive exploration of our individual efficiency data. Rather, they are designed to demonstrate how commonly used global threshold and efficiency measures can obscure potentially interesting insights about how perceptual information is being processed and represented. The technique and analyses that we have described in this report offer a principled method for going beyond global measures of threshold and efficiency and dissecting human and ideal performance at a more microscopic level.

References

Banks, M. S., Geisler, W. S., & Bennett, P. J. (1987). The physical limits of grating visibility. *Vision Research, 27,* 1915–1924.

Barlow, H. B. (1978). The efficiency of detecting changes of density in random dot patterns. *Vision Research, 18,* 637–650.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8,* 679–698.

Cornsweet, T. N. (1970). *Visual perception*. New York: Academic Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96,* 267–314.

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Cambridge: MIT Press.

Gold, J. M., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Research, 39,* 3537–3560.

Gold, J. M., Tadin, D., Cook, S. C., & Blake, R. (2008). The efficiency of biological motion perception. *Perception & Psychophysics, 70,* 88–95. doi:10.3758/PP.70.1.88

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29,* 228–244.

Green, D. M. (1960). Detection of a noise signal. *Journal of the Acoustical Society of America, 32,* 121–131.

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

Lewand, R. (2000). *Cryptological mathematics*. Washington: Mathematical Association of America.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research, 35,* 549–569.

MacMillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Pelli, D. G. (1981). *Effects of visual noise*. Cambridge: PhD thesis, Cambridge University.

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46,* 4646–4674.

Tanner, W. P., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *Journal of the Acoustical Society of America, 30,* 922–928.

Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-D objects in luminance noise. *Vision Research, 35,* 3053–3069.

Tjan, B. S., & Legge, G. E. (1998). The viewpoint complexity of an object-recognition task. *Vision Research, 38,* 2335–2350.

Tyler, C. W., Chan, H., Liu, L., McBride, B., & Kontsevich, L. (1992). Bit-stealing: How to get 1786 or more grey levels from an 8-bit color monitor. In B. E. Rogowitz (Ed.), *Human vision, visual processing, and digital display III: 10–13 February 1992 San Jose, California* (Vol. 1666, pp. 351–364). Bellingham: SPIE.