Manuscript submitted for publication.
Please do not quote without permission.

Version of September 20, 2008
Comments welcome.

Alphabetic Letter Perceivability, Similarity, and Bias

Shane T. Mueller
Klein Associates Division, Applied Research Associates Inc.

Christoph T. Weidemann
University of Pennsylvania

Shane T. Mueller
Klein Associates Division, ARA Inc.
1750 N. Commerce Center Blvd, Fairborn, OH 45324
e-mail: smueller@ara.com
phone: (937) 873-8166
Version of September 20, 2008

The similarity of the letters in the Latin alphabet has been measured numerous times using a variety of methodologies and stimuli. We provide a comprehensive review of past attempts to construct similarity spaces of the Latin alphabet, and determine that accuracy has frequently been attributed to a subset of three common sources: perceivability, bias, and similarity. We present the results of two new experiments which, for the first time, allow for the simultaneous estimation of these factors. We found that perceivability was affected substantially by the visual similarity between target and mask, whereas the similarity space did not vary substantially with the mask, and the impact of bias was small. Finally, we demonstrate a method (related to choice theory) by which these three factors can be estimated simultaneously for our experiments.

Researchers have studied the perceivability and confusability of the letters of the alphabet since the early days of modern psychology and visual science (Javal, 1881; Cattell, 1886). In the intervening 130 years, letter confusability and similarity have been measured numerous times using many different techniques. One interesting subset of this research has produced and reported the similarity structure of the entire Latin alphabet, in the form of error, confusability, or similarity matrices. To understand the major theoretical conclusions reached by studying these matrices, we identified more than 70 such attempts to characterize the similarity structure of the entire Latin alphabet. In an effort to synthesize the findings we will first provide a comprehensive review of these previous studies, then report the results of two new experiments that use methods not used in these previous studies, and finally present a statistical model that decomposes performance into contributions from perceivability, similarity, and bias.

Some of the research presented in this article was carried out while both authors were affiliated with the Department of Psychological and Brain Sciences, at Indiana University, Bloomington. This research was supported by NIMH Grant #12717 to Richard M. Shiffrin, and by a post-doctoral fellowship to C.T.W. from the German Academic Exchange Service (DAAD). We would like to thank Jim Townsend, Rob Goldstone, Richard Shiffrin, Kelly Addis, Andrew Cohen, Amy Criss, Bill Estes, Krystal Klein, Angela Nelson, and Adam Sanborn for helpful comments and discussions. Address correspondence to S. T. Mueller, Klein Associates Division, ARA Inc., 1750 N. Commerce Center Blvd., Fairborn OH 45324 (e-mail: smueller@ara.com) or to C. T. Weidemann, University of Pennsylvania, Department of Psychology, 3401 Walnut St., Room 302c, Philadelphia, PA 19104 (e-mail: ctw@cogsci.info).

Overview of Prior Research Motivations

Previous attempts to produce letter similarity matrices can be characterized by three primary motivations: (1) Applied attempts to make written text more comprehensible or allow learners to acquire reading skills more easily; (2) empirical research aimed at understanding the visual system; and (3) theoretical research attempting to characterize or model how letters are represented by the visual or cognitive system.

Many early researchers were concerned with identifying typefaces, fonts, and letters that were more or less legible, with the aim of improving printing and typesetting. For example, Javal (1881), Helmholtz's students Cattell (1886) and Sanford (1888), Roethlein (1912), and Tinker (1928) all attempted to rank letters in their order of legibility, identifying letters that were especially confusable in order to allow faster reading and less error-prone communication. Javal (1881), Cattell (1886), and Sanford (1888) each made suggestions about how to modify some letters to be more distinguishable and readable. One of the most substantial efforts aimed at improving the legibility of typeset text was made by Ovink (1938), who published a book describing in detail the errors and confusions produced for letters and numbers of eleven different fonts, including detailed recommendations for how each letter should be formed to improve its legibility. Other early applied research was concerned with ophthalmological tests (including Javal, 1881, as well as Hartridge & Owen,

1922 and Banister, 1927). Similar applied research has continued in more recent years: Bell (1967), van Nes (1983) and Gupta, Geyer, and Maalouf (1983) each have dealt with practical modern applications of font face and letter confusions.

Despite the obvious practical applications for this type of research, by far the most common motivation for collecting letter similarity matrices has been to understand aspects of the perceptual system. Early researchers performed detailed psychophysical studies into the limits of letter perceivability with respect to numerous secondary variables (e.g., presentation time: Sanford, 1888; distance and size: Sanford, 1888, Korte, 1923; peripheral eccentricity: Dockeray & Pillsbury, 1910), adopting techniques that continue to be used today. Later researchers have attempted to use similarity matrices to understand other aspects of visual perception, such as representation and configurality (e.g., McGraw, Rehling, & Goldstone, 1994). Such empirical research has not only investigated the visual perception system, but has also studied tactile perception (Loomis, 1974; Craig, 1979), learning (Popp, 1964), choice behavior (Townsend, 1971a; Townsend, 1971b) and other relevant psychological phenomena.

A great number of these studies have been conducted in order to verify or test models of visual perception. These models often include a description of the visual features used to represent letters, which in turn have produced similarity matrices of their own. Occasionally, these theoretic similarity matrices have been published, albeit sometimes in the form of a representational feature set that can be used to represent all letters (Gibson, 1969, Geyer & DeWald, 1973).

Other theoretical measures of letter similarity have been developed that were not directly based on theories or models of the visual system, but rather examined the physical images representing the letters. For example, some researchers have used simple methods of letter congruency or overlap (e.g., Dunn-Rankin, Leton, & Shelton, 1968; Gibson, 1969) to measure letter similarity, whereas others have developed more elaborate techniques relying on Fourier decomposition (Coffin, 1978; Gervais, Harvey, & Roberts, 1984; Blommaert, 1988).

Overview of Methodologies

The most commonly used procedure involves presenting characters and requiring an observer to name the identity of the presented character. Confusion matrices have typically been constructed by computing the number of times each letter was given as a response for each presented letter. Typically, these letter naming procedures have produced confusion matrices with most trials being correct (along the diagonal), with most other cells empty or having just a few errors, and a few specific confusions (usually between visually similar letters) capturing most of the errors. Because letter pairs are not compared directly, these naming methods are indirect measures of letter similarity, in that errors presumably index the similarity between the presented stimulus and participants' memories for each alternative response.

The informativeness of an experiment can be enhanced when more errors are committed, and so a number of techniques have been used to induce more detection and naming errors. As reviewed above, experiments have commonly used standard psychophysical techniques (such as brief, small, peripheral, noisy, or low contrast presentations) to reduce naming accuracy and develop better estimates of letter similarity. Furthermore, some researchers have studied haptic identification of letters (Craig, 1979; Loomis, 1974; Kikuchi, Yamashita, Sagawa, & Wake, 1979; and Loomis, 1982), which tends to be more error-prone than visual identification, and others have tested subjects who naturally make errors in letter identification, even when the stimuli are presented clearly, such as children (Gibson, Osser, Schiff, & Smith, 1963, Popp, 1964, and Courrieu & de Falco, 1989), pigeons (Blough, 1985), or patients with motor output difficulties (Miozzo & De Bastiani, 2002). Because these subjects are often unable to name letter stimuli, these researchers sometimes measured performance by presenting a small set of alternatives (often just two) from which a response could be chosen. In contrast to the letter naming procedures described earlier, these are more direct measures for assessing similarity, because comparisons between the alternative letters can be made explicitly between presented stimuli, rather than requiring comparison of a stimulus to well-learned internal representations.

Other direct methods for measuring letter similarity have been used as well. For example, some researchers have measured similarity by asking participants to rate how similar each pair of letters is (e.g., Kuennapas & Janson, 1969, Podgorny & Garner, 1979; Boles & Clifford, 1989) or have otherwise elicited subjective similarity estimates (Dunn-Rankin, 1968). In addition, saccade times and accuracies (Jacobs, Nazir, & Heller, 1989), and response times from a same-different task (Podgorny & Garner, 1979) can be considered direct measures.

Across these past experiments, a wide variety of methods have been used to measure letter similarity. We have conducted an extensive review of the literature in which we have found more than 70 cases where letter similarity for the entire alphabet was measured and reported. These are summarized in Table 1.

To be included in Table 1, we required that an experiment must use most or all letters of the Latin alphabet. A substantial number of research reports have shown similarity effects of a small subset of letters, often incidental to the original goals of the research, and we did not include these.[1] Several papers we reviewed and included in Table 1 did not contain complete similarity matrices, but instead reported sets of confusable letters (e.g., Roethlein, 1912), or listed only the most confusable letters. We felt that these were sufficiently

[1] We have included several experiments that did not use the entire alphabet, but these measured nearly the entire alphabet and so warranted inclusion. Specifically, we included Kuennapas and Janson (1969) who used the Swedish alphabet which does not include the "w" but does include three additional symbols; Dunn-Rankin (1968), who used only the 21 most common letters of the alphabet; and Bouma (1971), who excluded the letter "y" in one condition of his experiment.

useful to merit inclusion. Finally, some theoretical techniques we included in Table 1 did not produce actual similarity matrices, but did report feature-based representations for letters. Because these representations can easily be used to derive theoretical similarity matrices, we have included this research as well. As a final note, we came across numerous experiments and research reports that collected, constructed, or mentioned otherwise unpublished letter similarity matrices, but did not report the actual matrices. We did not include these reports in Table 1, as these data sets are probably lost forever. Table 1 briefly describes the measurement methods, letter cases, and font faces used in the experiments.

Summary of Theoretical Conclusions

The behavioral studies summarized in Table 1 used a variety of techniques to measure letter similarity, and typically attributed accuracy to one or more of three distinct factors: visual perceivability, visual similarity, and response bias. Many of the early researchers who studied letter identification were primarily interested in the relative *perceivability* or legibility of different characters. Perceivability is a theoretical construct affecting the probability the observer forms a veridical percept from the stimulus, independent of response factors and the uniqueness of that percept. One might expect perceivability to be affected by manipulations like stimulus intensity, size, duration, etc., and it is reasonable to think some letters may be more perceivable than others (e.g., a simple letter like "l" may be more difficult to see than a complex letter like "w"). Furthermore, perceivability may be influenced by factors such as visual masking, and it is likely that some types of masks make some letters more difficult to perceive than others. The notion of perceivability has continued to be relevant in modern theories such as signal detection theory (corresponding roughly to sensitivity parameters), and was used directly in the all-or-none activation model proposed by (Townsend, 1971a).

Yet even early researchers found that perceivability alone was insufficient to describe performance in these tasks. For example, Sanford (1888) reported that participants frequently gave incorrect responses that were visually similar to the stimuli. And indeed, some components of perceivability appear to stem from a letter's similarity to other letters, rather than poor perceivability per se. Thus, it appears that another factor important in letter identification is *visual similarity*.

Effects of the visual similarity between letters are closely related to those of perceivability. A letter that is visually similar to a large set of other letters (e.g., "O" with "Q", "G", "C", etc.) is likely to be identified less accurately, on average, than a letter with fewer similar alternatives (e.g., "S"). But these errors in identification appear to stem not from the physical shape of any individual letter (i.e., perceivability), but rather from its similarity to other choice alternatives. It is not surprising that similarity and perceivability have rarely been used together to predict letter identification accuracy, given that most empirical measures of letter similarity use an unrestricted letter naming task, as was done in about 50 of the experiments we reviewed. If a naming task is used, errors stemming from perceivability often cannot be distinguished from errors stemming from similarity, because a stimulus that is perceived poorly may elicit responses of similar letters, even if the errors stem from perceivability and not similarity.

In fact, many of the experiments reviewed in Table 1 do not distinguish between perceivability and similarity, and used similarity alone to describe their data (e.g., Podgorny & Garner, 1979). However, data patterns have sometimes suggested that factors other than similarity and perceivability influenced identification accuracy. In many of these cases, researchers have assumed that these other factors include guessing or *response biases*.

Response biases are present in classical theories of detection such as high-threshold theory (cf. Macmillan & Creelman, 1990, 2005), and for letter identification such biases were noted as early as 1922 (Hartridge & Owen, 1922). However, response biases gained wider use in the analysis of letter confusion data with the development of axiomatic theories of detection, such as the so-called Bradley-Terry-Luce Choice theory (Bradley & Terry, 1952; Luce, 1959, 1963) and signal detection theory (SDT, Green & Swets, 1966). Several experiments in Table 1 attempted to account for letter identification accuracies based on similarity and bias alone (e.g., Townsend, 1971a, 1971b; Gilmore et al., 1979). According to these theories, people may have biases for or against giving certain responses. These may be pure guessing biases invoked only when a participant is uncertain (as in high-threshold theory), or there could be biases in evidence decision criteria (as assumed by SDT or choice theory.)

These three factors (perceivability, similarity, and bias), although psychologically distinct, have rarely, if ever, been combined into a single model to account for alphabetic confusion data. This is probably because in many models, these effects are not identifiable (although even Shepard, 1957, provided formulas to estimate these factors). Yet this approach denies the distinction between the useful psychological constructs of perceivability, similarity, and bias, each of which have been used in isolation.

In order to measure the joint impact of these three factors, we used a technique that was not used to measure the similarity space of the complete Latin alphabet in any of the experiments we reviewed in Table 1: two-alternative forced-choice perceptual identification (2-AFC, e.g. Ratcliff, McKoon, & Verwoerd, 1989; Ratcliff & McKoon, 1997; Huber, Shiffrin, Lyle, & Ruys, 2001; Weidemann, Huber, & Shiffrin, 2005, 2008). Variations on the 2-AFC task have been in common use since at least the 1960s for memory and perceptual experiments, and the task was used prominently in experiments testing threshold theories of perception against strength-based accounts (such as SDT and choice theory, cf. Macmillan & Creelman, 2005).

In the 2-AFC task, a participant is presented with a brief target stimulus, often preceded and/or followed by a mask. A mask is a convenient method for reducing accuracy of detection, although many other methods have been used to accom-

Table 1
Summary of experiments reporting letter similarity matrices.

Reference	Method	Case	Typeface
Cattell (1886)	Naming errors	B	Latin serif
	Naming errors	B	Fraktur
Sanford (1888)	Naming errors of distant stimuli	L	Snellen
	Naming errors of brief stimuli	L	Snellen
	Naming errors of brief stimuli	L	Old-style Snellen
Dockeray and Pillsbury (1910)	Naming errors of stimuli in periphery	L	10-pt. Roman old-style
Roethlein (1912)	Confusable letter sets	B	16 different fonts
Hartridge and Owen (1922)	Naming of distant stimuli	U	Green's Letter Set
Korte (1923)	Naming of distant stimuli	B	Antiqua
	Naming of distant stimuli	B	Fraktur
Banister (1927)	Naming of distant brief stimuli	U	Green's Letter Set
	Naming of distant brief stimuli	U	Green's Letter Set
	Naming of distant stimuli	U	Green's Letter Set
Tinker (1928)	Naming of brief stimuli	B	Bold serif font
Ovink (1938)	Naming of distant stimuli	B	11 different fonts
Hodge (1962)	Letter reading errors	B	Uniform-stroke alphabet
Gibson et al. (1963)	Children's matching of target to set of random choices	U	Sign-typewriter
	Children's matching of target to set of similar or dissimilar letters	U	Sign-typewriter
Popp (1964)	Forced choice confusions of children	L	Century-style
Bell (1967)	Naming errors of brief stimuli	U	Long Gothic
	Naming errors of brief stimuli	L	Murray
Dunn-Rankin (1968)	Similarity preference of letter pairs	L	Century Schoolbook
Dunn-Rankin et al. (1968)	*Shape congruency	L	Century Schoolbook
Kuennapas and Janson (1969)	Subjective similarity ratings	L	Sans serif Swedish alphabet
Uttal (1969)	Naming errors of brief masked stimuli	U	5x7 dot matrix
Laughery (1969)	*Feature Analysis	U	Roman block letters
Gibson (1969)	*Feature Analysis	U	Roman block letters
Fisher et al. (1969)	Naming errors of 200-ms stimuli	U	Futura medium
	Naming errors of 400-ms stimuli	U	Futura medium
	Naming errors of brief stimuli[2]	U	Leroy lettering set
Townsend (1971a)	Naming errors of brief unmasked stimuli	U	Typewriter font
	Naming errors of brief masked stimuli	U	Typewriter font
Townsend (1971b)	Naming errors of brief unmasked stimuli	U	Typewriter font
Bouma (1971)	Naming errors of distant stimuli	L	Courier
	Naming errors of stimuli in periphery	L	Courier
Geyer and DeWald (1973)	*Feature analysis	U	Roman block letters
Engel et al. (1973)	Naming errors of brief stimuli	L	Century Schoolbook
Loomis (1974)	Tactile letter identification	U	18x13 matrix
Briggs and Hocevar (1975)	*Feature analysis	U	Roman block letters
Mayzner (1975)	Naming errors of brief stimuli	U	5x7 dot matrix
Thorson (1976)	*Overlap values based on feature analysis	U	Roman block letters
Geyer (1977)	Naming errors of brief dim stimuli	L	Tactype Futura demi 5452
Coffin (1978)	*Fourier spectra similarity	U	128x128-pixel block letters
Podgorny and Garner (1979)	Same-different choice RT	U	5x7 Dot matrix chars.
	Subjective similarity ratings	U	5x7 Dot matrix chars.
Gilmore et al. (1979)	Naming errors of brief stimuli	U	5x7 Dot matrix chars.
Kikuchi et al. (1979)	Tactile letter identification	U	17x17 Dot matrix chars.
Craig (1979)	Tactile letter identification	U	6x18 Dot matrix chars.
Keren and Baggen (1981)	*Feature analysis	U	5x7 Dot matrix chars.
Johnson and Phillips (1981)	Tactile letter identification	U	Sans serif embossed letters
Loomis (1982)	Visual identification	U	Blurred Helvetica
	Tactual identification	U	Helvetica
Paap et al. (1982)	Naming errors	U	Terak
Gupta et al. (1983)	Naming errors of brief dim stimuli	U	5x7 Dot matrix chars.
	Naming errors of brief dim stimuli	U	Keepsake
Phillips et al. (1983)	Naming errors of small visual stimuli	U	Helvetica
	Tactile identification	U	Sans serif
Gervais et al. (1984)	Naming errors of brief stimuli	U	Helvetica
	*Similarity of spatial frequency spectra	U	Helvetica
van Nes (1983)	Naming errors of brief peripheral stimuli	L	12x10 pixel matrix–least confusable (IPO-Normal)
	Naming errors of brief peripheral stimuli	L	12x10 pixel matrix–most confusable
van der Heijden et al. (1984)	Naming errors of brief stimuli	U	Sans serif roman
Blough (1985)	Pigeon's 2-alternative letter matching	U	5x7 dot matrix
Blommaert (1988)	*Fourier spectra similarity	L	16x32 pixel matrix courier
Heiser (1988)	*Choice model analysis of confusions	U	Sans serif roman
Jacobs et al. (1989)	Saccade times to matching target	L	9x10 pixel matrix
	Saccade errors to distractor	L	9x10 pixel matrix
Boles and Clifford (1989)	Subjective similarity ratings	B	Apple-Psych letters
Courrieu and de Falco (1989)	Children identifying targets that matched uppercase reference	L	Printed script
Watson and Fitzhugh (1989)	Naming errors of low-contrast stimuli	U	5x9 pixel font (gacha.r.7)
McGraw et al. (1994)	Letter identification with keyboard	L	"Gridfont" chars.
Reich and Bedell (2000)	Naming of tiny or peripheral letters	U	Sloan Letters
Liu and Arditi (2001)	Naming of tiny crowded or spaced letter strings	U	Sloan Letters
Miozzo and De Bastiani (2002)	Writing errors of impaired patient	B	handwriting
Mueller and Weidemann (2008a, Exp. 1)	Forced choice identification of letter with distractor: @-mask	U	Courier
	Forced choice response latencies: @-mask	U	Courier
Mueller and Weidemann (2008a, Exp. 2)	Forced choice identification of letter with distractor: #-mask	U	Courier
	Forced choice response latencies: #-mask	U	Courier

Note. In the Case column, "L" indicates lowercase, "U" indicates uppercase, and "B" indicates both cases were studied. Methods denoted with an * were measures developed by analyzing the visual form of letters, and not directly based on data from observers.

plish this goal in the studies we reviewed in Table 1.[3] After the masked character is presented, the participant is shown two choices: the target, and an incorrect alternative (i.e., the foil). The participant then indicates which of the two options was presented. Identification performance is reduced by degrading the target presentation (e.g., by brief duration, masking, and/or low contrast) to avoid perfect accuracy.

Several previous experiments reviewed in Table 1 have used forced-choice procedures, but all have differed substantially from the 2-AFC task we will report next. For example, the children in Popp's (1964) experiment were shown a target, and then given the choice of two letters (the target and a foil). However, errors occurred because the children had not learned letter discrimination perfectly, and probably not because of any perceptual deficiencies. Dunn-Rankin (1968) also showed participants a letter followed by two comparison letters, but in that experiment the two choices did not always include the target, and participants were instructed to select the most visually similar option. Blough (1985) conducted an experiment similar to Popp (1964), but used pigeons instead of children. Finally, Jacobs et al. (1989) used a choice task to measure saccade accuracies and latency: participants were shown an uppercase target and then presented with two lowercase letters in the periphery; and were instructed to move their eyes to the lowercase version of the target.

The 2-AFC procedure has some potential advantages over letter naming techniques. Two of these advantages were mentioned explicitly by Macmillan and Creelman (2005): it tends to reduce bias, and produces high levels of performance. Consequently, the procedure may mitigate some of the effects of guessing and response biases that can be introduced in naming procedures. It also provides a more direct measure of similarity, because every pairing of letters is measured explicitly, rather than using the low-probability naming confusions as an index of similarity. Thus, it has the potential to measure differences in similarity between letter pairs that are only rarely confused. Because of these advantages, a 2-AFC task may enable better estimation of the three factors typically associated with letter detection: perceivability, bias, and similarity.

Experiment 1

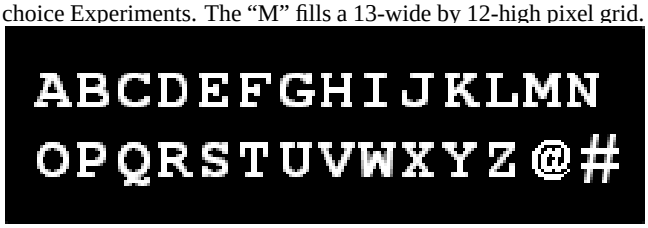
To collect letter similarity data that allows for the simultaneous estimation of the joint effects of perceivability, bias, and similarity, we carried out an experiment involving a 2-AFC perceptual letter identification task. In this task letters were presented briefly and flanked by a pre- and post-mask allowing us to also investigate how similarities between the targets and masks impact these factors.

Method

Participants. One hundred and eighteen undergraduate students at Indiana University participated, in exchange for introductory psychology course credit.

Materials, Equipment and Display. All 26 upper-case letters of the Latin alphabet served as stimuli. Letters were presented in 16-point Courier New Bold. All letters except "Q" were 12 pixels high and all letters were between 8 and 13 pixels wide. An "@" was adjusted in font and size ("Arial Narrow Bold", 14 pt.) to cover the display area of the letters. A depiction of the stimuli and mask, enlarged to show the anti-aliasing and pixelation present on the display terminal, appears in Figure 1. The "#" character depicted in Figure 1 was not used in the current Experiment.

Figure 1. Depiction of the stimuli and mask used in the forced-choice Experiments. The "M" fills a 13-wide by 12-high pixel grid.



All stimuli were displayed on PC monitors with a vertical refresh rate of 120 Hz. The display was synchronized to the vertical refresh using the ExpLib programming library (Cohen & Sautner, 2001). This provided a minimum display increment of 8.33 ms, but due to the occasional unintentional use of different software driver settings, the display increments for a few participants were as high as 10 ms.

The stimuli were presented in white against a black background. Each subject sat in an enclosed booth with dim lighting. The distance to the monitor (controlled by chin rests positioned approximately 60 cm from the screen) and font size were chosen such that the to-be-identified letter encompassed less than 1° of visual angle.

Responses for the 2-AFC test were collected through a standard computer keyboard. Participants were asked to press the "z"-key and the "/"-key to choose the left and right alternative respectively.

Procedure. Each trial began with the presentation of an "@"-sign pre-mask (300 ms) immediately followed by the target letter (for an individually adjusted duration as described below). Immediately after the offset of the target letter an "@"-sign post-mask was presented and remained until 600 ms after the first pre-mask was presented (regardless of how long the stimuli was presented). The post-mask was immediately followed by two choices presented to the right and left, with the position of the correct choice randomly determined on each trial.

The first block of 96 trials of the experiment was used to adjust the display time of the target presentation such that performance was roughly 75%. Target letters and foils for

[3] Despite the claim that the specific choice of a mask can limit generalizability (cf. Eriksen, 1980), the effect of specific masks across the entire alphabet needs to be better understood. In terms of the theoretical perspective we put forth in this paper, it is important to know whether similarity between the mask and the target impacts estimates of perceivability, bias, and similarity.

these calibration trials were randomly chosen (with replacement) from the alphabet.

Across participants, the mean presentation time obtained by using this procedure was 54 ms, but as is typical for studies using a 2-AFC perceptual identification paradigm (e.g. Huber, Shiffrin, Lyle, & Quach, 2002; Huber et al., 2001; Huber, Shiffrin, Quach, & Lyle, 2002; Weidemann et al., 2005, 2008), there were large individual differences: The minimum, 25th-percentile, median, 75th-percentile, and maximum mean target presentation times were 10, 39, 50, 64, and 150 ms, respectively.

Following the block of 96 calibration trials, there were five blocks with 130 experimental trials each. Each block was preceded by three additional practice trials which were discarded (targets and foils for these practice trials were randomly chosen, with replacement, from the alphabet). Target and foil letters were assigned to test trials randomly with the restriction that all 650 possible combinations of targets and foils needed to be presented exactly once in the test trials of the experiment.

Feedback was given after every trial. A check-mark and the word "correct" appeared in green when the answer was correct and a cross-mark ("X") and the word "incorrect" were presented in red when the answer was incorrect. The feedback stayed on the screen for 700 ms and was immediately followed by the presentation of the pre-mask for the next trial (unless the current trial was the last trial in a block).

After each block, participants received feedback providing the percentage of correct trials in the last block and the mean response time (this was the only time when feedback about response time was given, and the instructions emphasized accuracy rather than response speed). Between blocks, participants were encouraged to take short breaks and only resume the experiment when they were ready to continue. The entire experiment took about 45 minutes.

Results

Our experiment provides two measures by which letter similarity can be assessed: accuracy and response latencies. Accuracy presumably serves as an index of similarity because physically similar letters become attractive choices when partial information about letter identity has been obtained. Although participants were not encouraged to respond quickly, response latencies may also serve as an index of similarity: participants may deliberate on trials whose choices are very similar to one another.

Both of these types of data are shown in Table 2, with accuracy in the top half of the table and mean response time in the bottom half of the table. For the response latencies, we eliminated the 89 trials (out of 76,700) on which the response took longer than five seconds. Otherwise, both correct and incorrect trials were included.

Correct responses were made on average 110 ms faster than incorrect responses (542 ms vs. 652 ms), which was highly reliable ($t(117) = 9.05$, $p < .01$). Because of this correlation between speed and accuracy, and because the task was designed to measure response accuracy, we performed all subsequent analyses using only the accuracy values found in Table 2.

The data shown in Table 2 is perhaps too complex to easily make sense of. We have therefore plotted mean accuracies from Table 2 in Figure 2, with respect to either the target (top panel) or foil (bottom panel) present on each trial.

For each target letter, the mean accuracies across the 25 foils would be expected to have a binomial distribution, if there were no impact of bias or similarity. For a binomial distribution with mean .77 and 118 observations (as in our experiment), the standard error of the estimate is $\sqrt{\frac{.75\times.25}{118}} = .0387$, which indicates that 96% of the observations[4] should fall within roughly 8% on either side of the mean (because a 96% confidence interval for the *t* distribution with 118 degrees of freedom is 2.07 standard units, and $2.07 \times .0387 = .0802$). The grey boxes in each column of Figure 2 show these 96% confidence ranges of the mean estimates. Means that fall outside these grey boxes can be considered outliers, and may indicate that particular target/foil pairs are recognized especially well or poorly. Similar reasoning can be used to estimate the expected range across column means. With 2950 observations (118×25), the estimated standard error is .0086 for a binomial distribution with mean .77 (as occurred in our experiment). The 96% confidence bounds range between .75 and .79. Column means (depicted as horizontal line segments in the figure) outside that interval (marked by horizontal lines in the figure) indicate substantial deviations from the mean for a single target or foil across the experiment.

With respect to the factors of perceivability, bias, and similarity, reliable differences supporting each of these can be seen in Figure 2. The mean accuracies for many of the targets differ reliably from the average (for example, "A", "J", "Q", and "X" fall below, and "B", "M", "R", "S", and "V" lie above), falling well outside the range expected by chance alone. These differences are effects that the target has across all foils, and are likely to be most impacted by perceivability. Furthermore, many points fall considerably below the expected range for each target (such as "O", "Q", and "D" for a number of similarly shaped targets), and these inaccuracies might be attributed to similarity. Finally, the accuracies associated with some foil letters appear to be uniformly higher (e.g., "I", "L", and "Z") or lower (e.g., "D", "O", "U") than average. Because these effects occur when these letters were foils, they likely stem from response biases rather than target perception.

This analysis also suggests several phenomena related to the post-stimulus mask. First, the target that was least accurate was the "A". This may have occurred because the "@" mask, which contains a lowercase "a", somehow interfered with correct identification of the "A". We propose a hypotheses for why this happened in the General Discussion. Another provocative result revealed by Figure 2 is that when

[4] We used a 96% confidence range because there were 25 points per column, and because 1-1/25=.96, one would expect one mean or fewer to fall outside this range if they were sampled from the same distribution.

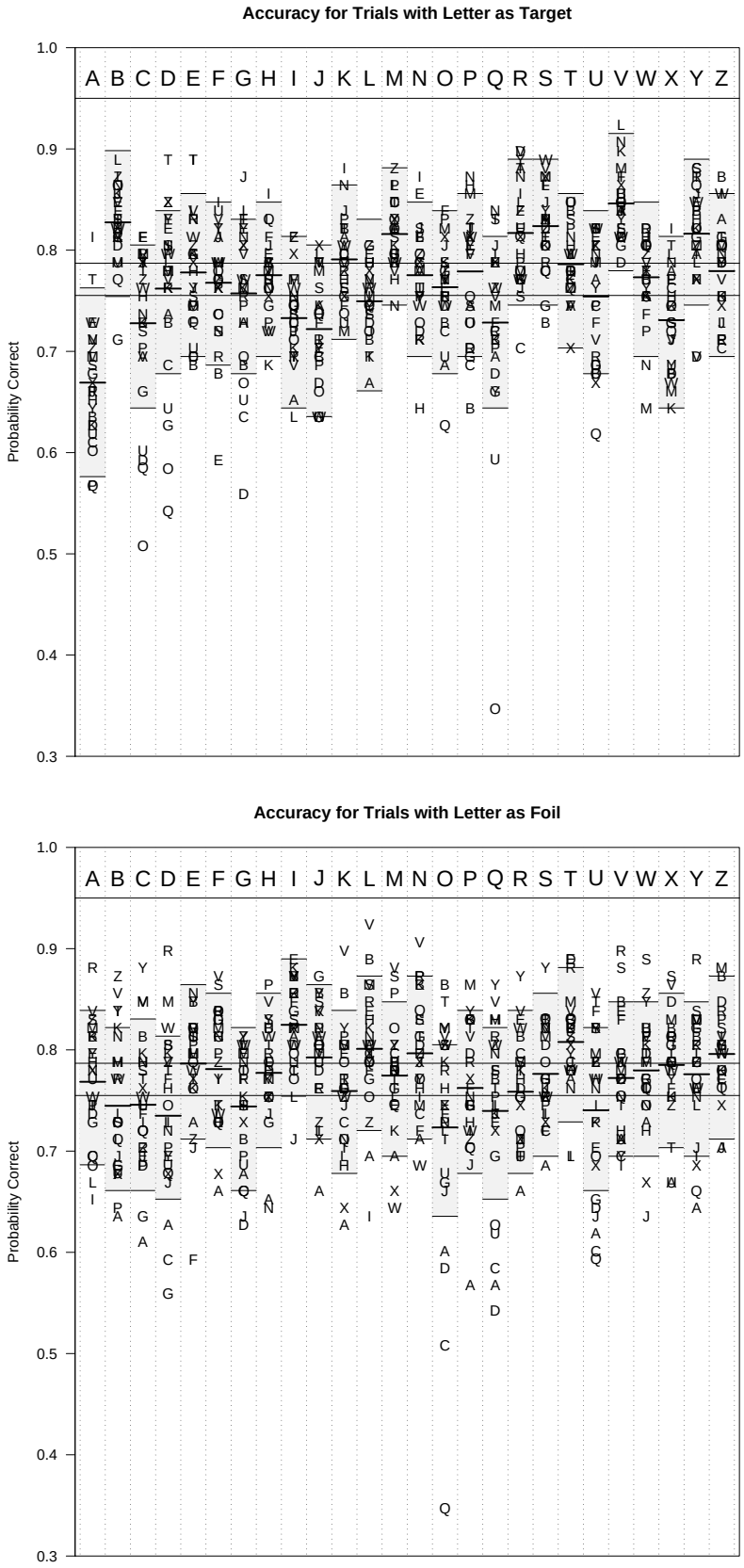


Figure 2. Accuracy for letter combinations in Experiment 1. Top panel shows accuracy for letter pairs sorted by target, and bottom panel shows accuracy for letter pairs sorted by foil. The two horizontal lines in each panel between .7 and .8 depict a 96% confidence range for the overall means of each target (which are shown by horizontal line segments in each column.) The grey boxes in each column depicts the 96% confidence range for each target-foil combination, assuming that all conditions in a single column came from the same mean. Conditions well outside these bounds correspond to conditions where the impact of similarity is strong. Exact values for point are shown in Table 2.

Table 2
Accuracy and response time matrix for Experiment 1.

Target Letter	Foil Letter																								
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		.636	.610	.627	.729	.661	.678	.653	.814	.661	.627	.695	.695	.712	.602	.568	.568	.661	.686	.771	.619	.712	.729	.669	.644	.703
B	.814		.814	.805	.847	.839	.712	.822	.873	.856	.856	.890	.788	.864	.864	.831	.771	.814	.831	.831	.788	.847	.822	.822	.814	.873
C	.695	.729		.593	.814	.814	.661	.754	.780	.788	.729	.797	.797	.737	.508	.703	.585	.797	.720	.788	.602	.695	.763	.788	.788	.771
D	.737	.729	.686		.822	.831	.627	.780	.805	.780	.763	.788	.780	.805	.585	.797	.542	.763	.805	.890	.644	.771	.797	.847	.831	.847
E	.797	.686	.695	.695		.729	.797	.780	.890	.763	.797	.839	.746	.831	.746	.746	.729	.831	.754	.890	.703	.839	.814	.788	.763	.805
F	.814	.678	.737	.771	.593		.771	.788	.847	.814	.763	.780	.788	.720	.737	.763	.780	.695	.720	.822	.839	.831	.788	.763	.822	.771
G	.729	.686	.636	.559	.822	.831		.729	.839	.873	.763	.771	.763	.814	.669	.746	.695	.754	.771	.831	.653	.797	.771	.805	.822	.805
H	.788	.788	.788	.763	.797	.814	.746		.856	.805	.686	.831	.780	.763	.763	.729	.831	.771	.771	.780	.771	.720	.720	.754	.788	.780
I	.653	.737	.729	.729	.814	.771	.746	.754		.720	.703	.636	.771	.754	.712	.720	.746	.695	.737	.695	.746	.686	.763	.797	.695	.814
J	.746	.695	.695	.669	.703	.729	.636	.737	.712		.746	.797	.780	.788	.661	.686	.737	.712	.763	.788	.636	.788	.636	.805	.703	.703
K	.814	.822	.797	.797	.763	.746	.754	.771	.881	.839		.822	.720	.864	.788	.831	.737	.780	.763	.822	.729	.788	.805	.754	.805	.780
L	.669	.712	.746	.729	.797	.737	.805	.788	.754	.720	.695		.754	.771	.720	.746	.746	.763	.737	.695	.788	.771	.763	.780	.746	.805
M	.822	.788	.847	.847	.797	.822	.797	.771	.822	.797	.805	.864		.746	.822	.864	.831	.788	.814	.847	.797	.780	.788	.831	.831	.881
N	.780	.814	.788	.720	.856	.814	.788	.644	.873	.822	.712	.814	.780		.729	.754	.797	.712	.822	.763	.763	.754	.746	.788	.754	.797
O	.686	.729	.720	.746	.763	.839	.780	.754	.771	.805	.788	.754	.822	.771		.831	.627	.754	.788	.780	.695	.771	.746	.814	.771	.780
P	.746	.644	.686	.703	.805	.797	.695	.864	.822	.822	.814	.805	.856	.873	.729		.754	.703	.746	.822	.729	.797	.814	.814	.805	.831
Q	.695	.712	.720	.678	.788	.729	.661	.788	.797	.805	.712	.788	.746	.839	.347	.703		.720	.831	.831	.593	.754	.763	.788	.661	.763
R	.881	.771	.703	.898	.822	.839	.771	.797	.856	.763	.771	.847	.780	.873	.780	.788	.814		.754	.881	.822	.898	.771	.814	.890	.839
S	.831	.729	.780	.805	.814	.864	.746	.831	.831	.847	.805	.864	.873	.831	.805	.831	.780	.788		.814	.822	.881	.890	.873	.839	.822
T	.746	.839	.771	.780	.771	.746	.780	.805	.780	.797	.771	.797	.763	.814	.847	.822	.763	.780	.831		.847	.746	.797	.703	.797	.763
U	.771	.678	.746	.686	.814	.729	.686	.822	.788	.788	.805	.805	.788	.797	.678	.746	.619	.695	.822	.822		.712	.822	.669	.763	.805
V	.839	.856	.847	.788	.822	.873	.805	.847	.873	.839	.898	.924	.881	.907	.814	.814	.847	.839	.822	.839	.856		.814	.864	.831	.814
W	.754	.771	.754	.822	.780	.737	.805	.814	.805	.814	.754	.805	.644	.686	.805	.720	.805	.822	.754	.780	.771	.788		.780	.763	.797
X	.780	.678	.763	.678	.771	.678	.729	.754	.822	.712	.644	.797	.661	.788	.746	.771	.720	.746	.729	.805	.686	.712	.669		.686	.746
Y	.797	.839	.881	.695	.847	.771	.814	.831	.873	.856	.822	.788	.805	.771	.822	.839	.864	.873	.881	.797	.822	.695	.847	.771		.805
Z	.831	.873	.703	.788	.712	.788	.814	.754	.856	.729	.754	.729	.805	.797	.805	.712	.805	.712	.754	.814	.788	.771	.856	.746	.788	
A		618	586	599	641	577	567	671	643	605	571	612	637	626	604	596	626	592	652	602	588	605	587	654	669	613
B	561		551	612	520	501	575	546	511	505	502	535	510	504	551	544	540	577	605	538	530	564	539	531	514	532
C	597	632		623	589	589	650	573	552	570	595	585	569	552	534	647	684	631	657	557	627	556	577	580	552	614
D	586	609	619		587	576	567	595	573	602	569	566	560	601	573	571	651	639	572	575	623	537	576	581	569	553
E	550	573	525	527		581	507	541	507	566	534	569	549	518	496	560	638	559	568	525	507	521	557	545	538	585
F	520	560	567	535	629		557	568	530	588	580	529	584	534	492	542	564	547	560	546	540	562	556	573	546	602
G	630	615	709	631	603	596		608	537	658	653	572	588	616	587	569	628	574	613	611	687	601	617	632	585	603
H	543	495	502	569	551	574	541		472	538	557	552	606	612	495	526	536	576	549	511	574	594	585	528	526	546
I	516	495	512	558	545	577	522	521		553	570	578	521	518	483	518	456	494	477	583	563	526	546	518	578	530
J	525	582	559	538	586	553	560	597	607		591	598	569	568	530	538	572	584	571	604	543	538	583	610	599	585
K	515	503	541	548	572	538	504	560	468	528		532	622	585	557	491	543	524	537	528	547	487	588	569	532	599
L	546	517	512	518	532	611	489	530	562	591	535		535	491	500	530	560	576	507	556	508	510	592	514	560	529
M	498	515	497	506	493	604	483	617	500	483	504	524		567	506	497	496	524	581	514	528	534	603	517	506	465
N	536	506	523	554	579	520	557	586	483	534	576	551	585		537	500	506	535	493	491	497	572	615	596	559	549
O	586	677	646	622	597	572	555	602	621	622	643	598	640	600		595	598	598	574	530	667	611	565	599	593	609
P	507	552	555	567	546	553	541	556	504	512	544	521	517	556	531		548	572	612	548	511	471	564	528	507	552
Q	647	603	670	619	628	572	784	640	587	566	642	597	641	624	672	637		599	609	661	697	662	654	632	641	621
R	550	564	574	543	555	535	533	546	500	507	556	511	574	523	542	558	539		552	575	512	531	567	489	493	570
S	533	591	546	519	600	547	506	486	507	505	518	498	529	518	584	557	530	554		555	479	490	552	537	504	538
T	491	462	452	458	567	509	491	526	540	559	499	540	491	526	518	487	530	534	476		487	560	570	567	523	558
U	541	624	582	685	544	545	620	549	604	635	561	595	592	552	525	612	633	655	581	523		610	551	573	572	608
V	457	478	464	494	495	515	480	544	494	476	520	537	525	531	501	462	535	483	505	506	558		501	495	536	510
W	554	525	515	507	542	556	543	575	499	511	609	593	635	631	493	533	504	495	572	548	542	560		593	551	545
X	615	556	519	533	574	575	516	551	587	541	570	542	567	614	494	533	531	557	596	587	581	634	617		589	581
Y	534	494	467	491	571	488	456	586	531	497	508	570	451	495	496	494	509	493	484	556	516	579	583	544		518
Z	567	542	566	562	563	566	553	609	580	602	611	590	576	607	493	547	532	573	490	571	568	545	527	518	570	

Note. Values in top half of table indicates the proportion of participants who responded correctly for each target-foil combination. Values in the bottom half indicate the mean response times (in ms) for correct and incorrect responses.

round letters (i.e., "O", "D", "Q", "U", "G", or "C") appeared as a foil, accuracy suffered. These letters are visually similar to the "@" mask, and this similarity may lead people to choose the foil more often when it was round, resembling the mask. Finally, accuracy for these round letters was not especially improved when they appeared as targets, indicating that the visually similar mask interfered with perceptual identification, despite participants' increased tendency to choose them. Several foil letters led to above-average accuracy (i.e, "I", "T", and "L"). These letters stand out as being very dissimilar to the mask and other letters, indicating that people may have been more easily able to eliminate this option and select the target correctly.

These initial results of Experiment 1 indicate that the shape and identity of the mask may affect letter identification accuracy in important ways. To better investigate this influence, we carried out an (otherwise identical) experiment using a different commonly-used mask.

Experiment 2

Method

Participants. Ninety-six undergraduate students at Indiana University participated in exchange for introductory psychology course credit.

Materials, Equipment, and Display, and Procedure. The procedure used in Experiment 2 was identical to Experiment 1 in every way, except for the use of the"#" symbol instead of the "@" symbol as a post-stimulus mask. The mask was adjusted in font and size ("Arial Bold", 17 pt.) to cover the display area of the letters. A depiction of this mask is shown in Figure 1.

Results

As with Experiment 1, the similarity of pairs of stimuli can be assessed by examining both the accuracy and response latencies of trials in which one stimulus was a target and the other was the foil. These data are shown in Table 3, with accuracy in the top half and mean response time in the bottom half of the table. For the response latencies, we eliminated the 262 trials (out of 62,400) on which the responses took longer than five seconds. Otherwise, both correct and incorrect trials were included.

In Experiment 2, correct responses were made on average 93 ms faster than incorrect responses (581 ms vs. 674 ms), which was highly reliable ($t(95) = 7.4$, $p < .01$). As with Experiment 1, we performed all subsequent analyses on the accuracy data only.

Figure 3 depicts the accuracies for each letter combina-

Table 3
Accuracy and response time matrix for Experiment 2.

Target Letter	Foil Letter																								
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		.810	.845	.724	.741	.759	.707	.724	.879	.845	.707	.828	.759	.810	.793	.828	.914	.741	.793	.862	.862	.707	.707	.810	.741	.810
B	.724		.776	.638	.724	.724	.672	.690	.862	.776	.828	.828	.793	.776	.672	.672	.672	.724	.879	.793	.759	.741	.862	.690	.810	.759
C	.845	.845		.759	.845	.862	.672	.845	.828	.793	.845	.776	.862	.879	.741	.828	.759	.862	.845	.845	.724	.862	.862	.897	.759	.862
D	.862	.862	.862		.810	.776	.845	.828	.828	.914	.897	.897	.828	.862	.828	.810	.793	.879	.828	.828	.759	.810	.845	.793	.879	.897
E	.810	.638	.776	.707		.603	.759	.638	.862	.759	.690	.621	.724	.690	.759	.603	.810	.655	.638	.810	.741	.810	.741	.741	.810	.707
F	.707	.759	.793	.759	.603		.741	.776	.776	.810	.793	.862	.741	.879	.724	.724	.776	.707	.793	.621	.759	.724	.741	.741	.776	.655
G	.793	.724	.603	.759	.793	.828		.845	.845	.828	.707	.810	.810	.828	.776	.776	.690	.741	.879	.845	.793	.828	.776	.897	.776	.845
H	.552	.724	.707	.621	.638	.569	.741		.655	.603	.655	.724	.586	.586	.690	.517	.741	.534	.776	.690	.638	.655	.603	.655	.603	.569
I	.517	.431	.448	.586	.448	.414	.655	.569		.448	.431	.414	.466	.552	.638	.500	.586	.466	.466	.534	.638	.379	.466	.466	.466	.414
J	.741	.569	.690	.638	.776	.672	.690	.759	.879		.759	.724	.638	.724	.672	.690	.586	.724	.707	.759	.672	.690	.638	.707	.707	.707
K	.655	.759	.810	.793	.672	.776	.741	.741	.810	.828		.793	.690	.759	.810	.776	.776	.621	.776	.724	.828	.759	.621	.707	.828	.707
L	.707	.655	.638	.724	.672	.603	.810	.655	.741	.655	.655		.552	.707	.655	.707	.793	.603	.690	.621	.603	.586	.621	.586	.655	.707
M	.776	.724	.793	.845	.828	.741	.793	.707	.897	.862	.707	.828		.724	.862	.810	.741	.707	.828	.931	.845	.759	.621	.741	.845	.690
N	.793	.724	.862	.690	.690	.759	.845	.707	.776	.810	.759	.690	.672		.741	.810	.759	.707	.759	.810	.776	.655	.741	.690	.724	.776
O	.776	.879	.845	.810	.879	.828	.897	.776	.914	.828	.810	.879	.948	.862		.793	.741	.793	.931	.897	.862	.897	.810	.828	.810	.845
P	.793	.672	.793	.672	.810	.793	.845	.672	.759	.690	.759	.862	.793	.776	.862		.776	.793	.828	.828	.741	.776	.672	.707	.828	.793
Q	.828	.707	.724	.621	.828	.759	.810	.828	.776	.793	.793	.862	.879	.793	.483	.828		.776	.741	.810	.690	.845	.845	.724	.828	.793
R	.810	.793	.724	.759	.776	.879	.759	.810	.828	.793	.793	.828	.724	.793	.810	.793	.879		.724	.828	.810	.741	.793	.759	.776	.810
S	.793	.638	.724	.707	.741	.690	.724	.810	.759	.741	.879	.759	.793	.776	.776	.724	.741	.672		.759	.793	.741	.845	.845	.793	.828
T	.638	.638	.741	.638	.672	.483	.690	.621	.741	.638	.552	.741	.638	.638	.638	.655	.672	.621	.672		.621	.707	.707	.552	.707	.638
U	.828	.776	.793	.707	.741	.828	.707	.776	.810	.828	.759	.759	.793	.741	.672	.776	.672	.776	.828	.862		.793	.759	.828	.810	.828
V	.586	.845	.810	.724	.845	.776	.828	.845	.828	.793	.776	.810	.759	.741	.759	.724	.690	.810	.793	.810	.828		.741	.776	.828	.845
W	.759	.828	.724	.690	.586	.690	.655	.603	.776	.810	.759	.741	.638	.707	.845	.776	.828	.690	.810	.793	.655	.741		.724	.690	.690
X	.638	.603	.655	.655	.586	.569	.741	.621	.707	.655	.466	.690	.603	.621	.724	.621	.655	.707	.707	.655	.776	.621	.621		.552	.690
Y	.672	.552	.638	.776	.690	.655	.672	.638	.776	.810	.586	.569	.603	.741	.707	.672	.672	.621	.707	.741	.655	.586	.603	.655		.655
Z	.672	.741	.793	.810	.759	.759	.810	.793	.776	.690	.690	.845	.845	.845	.810	.793	.810	.793	.655	.776	.724	.655	.707	.586	.897	
A		509	570	644	510	581	550	590	558	528	572	582	528	522	511	474	594	594	571	508	531	573	547	561	524	533
B	597		547	594	607	553	638	582	559	586	623	569	526	562	578	569	618	597	551	536	633	519	614	584	591	585
C	508	539		521	521	560	646	518	521	487	490	493	528	509	587	557	590	469	527	505	579	524	529	538	566	568
D	524	553	540		532	521	556	474	499	555	499	499	535	524	589	520	597	505	517	550	564	560	492	495	504	545
E	610	617	613	557		590	617	580	618	555	552	594	596	525	606	575	560	631	632	631	645	506	595	604	567	716
F	596	584	541	565	634		527	585	613	665	632	564	585	598	532	602	522	585	588	577	542	604	625	608	584	590
G	544	581	755	583	555	570		546	541	517	534	509	517	521	576	551	693	534	539	567	589	604	518	584	541	562
H	655	651	694	692	707	714	623		584	602	685	644	726	711	638	653	647	710	610	673	679	638	658	715	593	619
I	664	621	571	624	730	721	678	690		625	589	711	724	633	626	629	630	698	660	630	672	647	660	721	647	688
J	586	586	534	666	601	578	623	668	560		638	565	617	561	529	598	626	662	581	620	576	594	637	604	607	582
K	628	531	516	604	672	587	557	615	572	584		599	537	608	523	555	607	572	564	593	649	637	674	634	593	630
L	573	576	516	589	589	582	626	519	622	615	570		552	542	592	545	533	563	571	581	531	553	558	657	609	608
M	635	601	571	627	645	585	582	729	586	527	568	567		611	552	589	560	627	541	577	567	596	778	621	614	617
N	594	557	542	561	611	644	546	669	540	571	610	536	713		632	615	563	593	611	622	549	660	680	663	589	599
O	457	506	547	477	479	466	483	447	464	476	488	511	506	506		504	605	492	495	449	505	472	479	497	554	534
P	536	596	619	487	571	543	551	548	484	532	592	503	564	564	508		525	605	577	499	537	474	480	555	484	501
Q	558	576	612	688	567	569	682	556	587	504	552	545	510	562	654	599		577	559	616	601	558	616	598	604	620
R	571	547	515	533	596	509	540	547	523	624	551	528	551	548	512	581	583		572	506	515	573	561	501	536	545
S	598	612	579	546	535	540	570	544	559	640	559	571	564	559	523	608	642	631		619	571	550	584	637	567	615
T	607	609	557	621	617	603	573	656	650	612	680	642	626	600	606	650	581	596	635		633	655	634	646	671	578
U	619	572	552	583	537	587	623	540	606	573	508	570	523	581	611	541	627	615	629	543		552	560	570	548	560
V	595	540	525	569	544	551	548	553	595	617	587	546	574	596	601	510	575	576	547	591	614		603	591	614	573
W	593	674	582	618	554	606	538	675	627	611	717	614	688	620	589	601	589	595	591	688	617	723		629	662	617
X	616	709	649	628	667	655	630	647	644	774	670	634	725	728	575	692	633	611	660	671	573	672	623		749	691
Y	639	621	615	624	687	613	592	727	661	698	607	736	670	599	576	598	577	573	588	641	637	671	651	662		718
Z	582	510	563	552	582	575	526	548	474	560	557	549	558	522	570	565	565	595	524	587	581	620	538	601	594	

Note. Values in top half of table indicates the proportion of participants (out of 96) who responded correctly for each target-foil combination. Values in the bottom half indicate the mean response times for correct and incorrect responses.

tion, sorted by target and foil. As we found in Experiment 1 (cf. Figure 2), there were consistent deviations in accuracy when letters appeared as targets, and also when letters appeared as foils. The 96% confidence range for each target is shown as a shaded region in each column, and the confidence range for the means are also shown by the horizontal lines.

One surprising result, shown in Figure 3, is that participants were extremely poor at identifying the target letters "I" and "T". Accuracy was above average on trials when either was presented as a foil, indicating a bias against making these responses. Overall, many letters containing vertical and horizontal features ("I", "T", "H", "X") appear to be identified less accurately when they are targets, whereas letters with round shapes ("O", "D", "C", "G", "R") appear to be identified better. Unlike Experiment 1, the letter "A" was identified with above average accuracy.

In the next section we will describe a formal statistical model that allows us to estimates the impact of perceivability, similarity, and bias across the two experiments. By considering data from both experiments simultaneously this model also allows us to better assess the extent to which different masks impact these factors.

Modeling the sources of letter detection accuracy

Our review of the literature suggests a conceptual model for the sources of letter detection accuracy that incorporates perceivability, bias, and similarity. To estimate these factors for our data, we need a more formal modeling approach. A number of approaches have been used to decompose the accuracies in confusion matrices into more fundamental properties of stimuli and responses. Many of these have taken the form of the "similarity" or "biased" choice model (Luce, 1963; Shepard, 1957). This theory constitutes an axiomatic model of how biases and similarities, and especially the number of choice alternatives impacts detection accuracy. Other alternatives are possible as well, (cf. Townsend, 1971a; Smith, 1992), but the purpose of this paper is not to test the validity of these models or determine the best one. Instead, we attempt to build a fairly theory-agnostic statistical model of the joint impacts of perceivability, response bias, and similarity. However, as we show in the Appendix, the parameters of this model can be interpreted in terms of choice theory.[5]

[5] Although our model and choice theory are in some senses equivalent, we view our model as atheoretical in that it simply accounts for the statistical predictors of the data in our 2-AFC task. Choice theory makes a number of axiomatic predictions that go be-

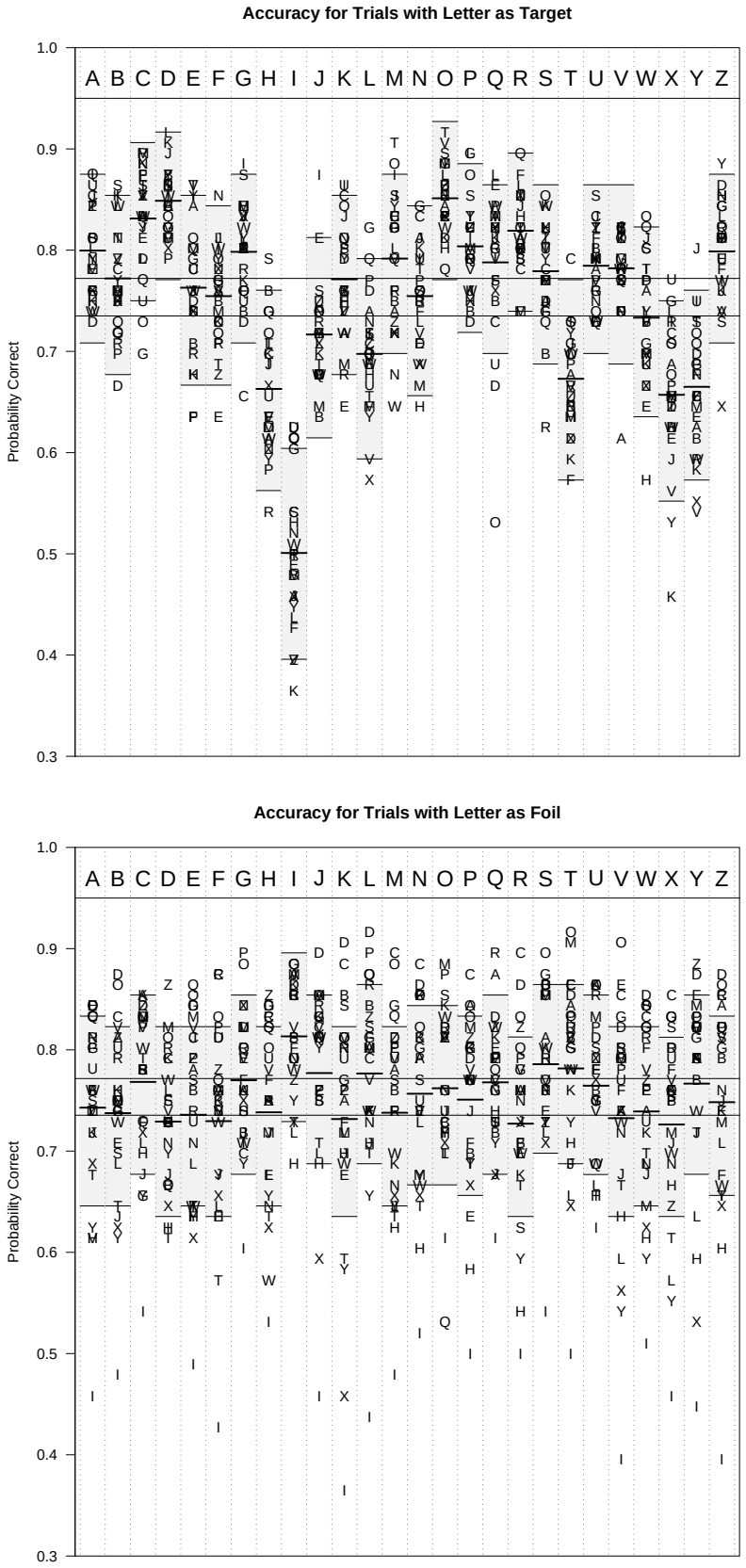


Figure 3. Accuracy for letter combinations in Experiment 2. Top panel shows accuracy for letter pairs sorted by target, and bottom panel shows accuracy for letter pairs sorted by foil. The two horizontal lines in each panel between .7 and .8 depict a 96% confidence range for the overall means of each target (which are shown by horizontal line segments in each column.) The grey boxes in each column depicts the 96% confidence range for each target-foil combination, assuming that all conditions in a single column came from the same mean. Conditions well outside these bounds correspond to conditions where the impact of similarity is strong. Exact values for every pair are available in Table 3.

A Statistical Model of Letter Detection

We developed the statistical model presented here specifically to account for the data from the two experiments described above. However, in principle it should be easy to extend it to apply to other similar paradigms. Our direct 2-AFC procedure measured each cell of the confusion matrix independently. For this direct comparison procedure, we assume that the perceptual processes produce an internal percept that differs somewhat from the target and this difference affects the accuracy for the target in general, regardless of the foil. We estimate the extent to which this difference affects accuracy with parameter λ_i, which describes the *perceivability* of stimulus s_i in log-odds units. If the value of λ_i were 0, (with no other contributors), this would produce a log-odds accuracy value of 0 for that stimulus, which is equivalent to an accuracy of 50%—chance guessing. As λ_i increases, baseline accuracy for that stimulus increase as well (all else being equal).

Several alternative psychological interpretations of λ_i are possible. For example, perceivability may have its impact during early perceptual stages, essentially affecting the probability that an accurate image is perceived. Or, it could be an aspect of the comparison process, assessing the similarity between an internal percept and a response option. We do not advocate a particular psychological interpretation, and our model does not distinguish between these accounts. Also, note that for experiments in which the distal stimulus is intentionally degraded with noise, λ_i would not distinguish between this external noise and any internal noise processes, although additional procedures could be used to assess these contributions independently (cf. Mueller & Weidemann, 2008b). If the mask character consistently introduces or erases features from the percept, this would likely result in a lower value for λ_i.

Next, we assume that response biases exist for each of the two alternatives. We denote bias with the symbol γ_i for response alternative i, and assume that it also impacts log-odds accuracy linearly. If γ_i is 0, the observer has no bias for a specific alternative. Positive values of γ indicate bias toward a response, and thus positive bias for a target and/or negative bias for a foil improves accuracy and, similarly, positive bias for foil and/or negative bias for a target harms accuracy.

Finally, we assume that the similarity between the proximal stimulus and the foil also impacts log-odds response accuracy linearly. In the model, we define a parameter corresponding to the *dissimilarity* between stimulus i and response j on log-odds accuracy called $\delta_{i,j}$. For δ, a value near 0 indicates that the accuracy can be well explained by the estimated perceivability and bias main effects alone. Positive δ values indicate greater dissimilarity, such that the letter pair is particularly easy to distinguish. Conversely, negative δ values indicate greater confusability, and produce lower accuracy than would be expected from the bias and perceivability parameters alone. Note that there might also be a number of psychological interpretation of $\delta_{i,j}$: typical accounts of similarity seem to explain it as the similarity between canonical representations of letter forms. It is unlikely to reflect perceptual biases in our experiment, because the response foil is not known before the stimulus flash, but it might be interpreted as the dissimilarity between a noisy proximal stimulus and the percept of the response alternative (which is consistent with choice theory's interpretation).

Estimating the values of these parameters is somewhat of a challenge because the set of predictors produces a linear model that is not orthogonal. In essence, one cannot uniquely identify perceivability *and* all 25 similarity parameters per target, because if perceivability changes, the similarity values can be adjusted to compensate. One approach to deal with this issue adopted by Shepard (1957) was to estimate perceivability as the mean similarity across targets, and bias as the mean similarity across foils. That approach is not consistent with our basic conceptual model in which the perceivability of a letter is a default value that is impacted by the effects of response alternatives. Furthermore, our prior examinations of the data (in Figures 2 and 3) suggested that most of the target-foil pairs fell within the range of what would be expected by chance based on perceivability alone, with a few particular pairs deviating from that level by significant amounts.

One way to estimate the parameters that is consistent with this approach is with statistical parameter selection methods. Using such a technique, we may be able to choose a small number of specific similarity values to estimate, and allow all other pairs to be accounted for only by their perceivability and bias. This approach also follows the basic logic of regression techniques, where the main effects (perceivability and bias) are allowed to account for variance initially, and interactions (such as similarity) are only used as necessary. However, we will also investigate a model in which all parameters are available for selection or omission.

For our forced-choice experiment, on trials with target i and foil j, the probability of choosing the target is denoted as $p_{i,j}$, and so the log-odds ratio of the correct response is $\ln\left(\frac{p_{i,j}}{1-p_{i,j}}\right)$. We assume that this value increases linearly with perceivability (λ_i), increases with bias for the target (γ_i), decreases with bias for the foil (γ_j), and increases as the dissimilarity between target and foil ($\delta_{i,j}$) increases as specified in Equation 1:

$$\ln\left(\frac{p_{i,j}}{1-p_{i,j}}\right) \propto \lambda_i + \gamma_i - \gamma_j + \delta_{i,j} \tag{1}$$

In this model, parameters are fairly easy to interpret. A log-odds value of 0.0 would indicate chance performance. Consequently, reasonable values of perceivability (λ) should be positive, because negative values would tend to push performance below chance. Accuracy in our experiments was around 75%, which corresponds to an odds ratio of 3:1. Thus, we expect the fitted λ parameters to have values around $\ln(3) = 1.1$. The neutral value for dissimilarity, δ, is 0, but some pairs might attain estimates reliably different from that default. These values can be considered the adjustment to

yond our model, such as its assumptions about how adding alternatives to the response set should impact accuracy.

the log-odds value stemming from the similarity between the percept and the foil. Negative values indicate more confusable letter pairs, and positive values indicate letter pairs that are especially easy to discriminate. Finally, the difference between the bias parameters (γ_i and γ_j) for a specific target-foil condition predicts the relative contribution the two biases have toward choice accuracy. Negative values indicate biases against a letter, which reduce accuracy when that letter is a target, but increase accuracy when that letter is a foil.

To apply the model to both experiments, we extend Equation 1 as follows:

$$\ln\left(\frac{p_{i,j}^{[x]}}{1-p_{i,j}^{[x]}}\right) \propto \lambda_i^{[x]} + \gamma_i^{[x]} - \gamma_j^{[x]} + \delta_{i,j} + \delta_{i,j}^{[X_2-X_1]}, \qquad (2)$$

where a raised $[x]$ indicates a code distinguishing between the two experiments and the raised $[X_2 - X_1]$, indicates that the associated parameter represents the differences in dissimilarities between the two experiments (this difference is added to the average dissimilarity across both experiments).

To fit the parameters of this model, we performed a linear regression accounting for log-odds accuracy with a linear combination of perceivability, bias, and dissimilarity as specified in Equation 2. For all model fits, we allowed the values of λ and γ to differ across the two experiments. The baseline model attempts to account for all data with just these 104 parameters (26 bias and 26 perceivability parameters per experiment). On the other extreme, we examined the complete model which included these parameters as well as a mean dissimilarity parameter for each letter pair (325 parameters), and differences in dissimilarities for each letter pair between the two experiments (325 more parameters). This model lacks identifiability, but is useful because it contains every parameter (and thus every intermediate model) we are interested in.

The most useful model lies somewhere between these two extremes. To identify a set of similarity parameters that reliably accounts for deviations from the baseline model, we used a stepwise regression procedure available in the stepAIC function of the MASS package (Venables & Ripley, 2002) of the R statistical computing environment (R Development Core Team, 2008), adopting a Bayesian Information Criterion (BIC) to determine which parameters should be included in the model. The BIC statistic combines maximum likelihood goodness of fit with a penalty factor for model complexity ($log_2(N)$ for N parameters), so that a parameter is only retained in the model if its goodness of fit improves more than the complexity penalty term. This approach began with the baseline model, then fit all models with one additional parameter that were subsets of the complete model, on each step choosing the model that had the smallest BIC score. This procedure continued, on each following iteration fitting all models that differed from the current model by one parameter (either by including a new parameter or excluding parameter that had previously been used). This stepwise procedure is fairly robust, and attained the same final model from various different starting models.

Table 4 shows the fits for the full model and the baseline model as well as those of two other models that we identified with the stepwise BIC procedure: Model 2 in Table 4 is based on the baseline model and includes an additional 24 global similarity parameters as well as 6 similarity differences between experiments. The parameters for the fit of this model to both experiments are shown in Table 5.

Model 4 in Table 4 was also selected by the stepwise procedure, but in this case we allowed all parameters—including those for perceivability and bias—to be removed. Model 4 had a lower number of parameters than any of the other models we considered, because it eliminated bias parameters that were not reliably different from 0 (it retained 6 and 11 bias parameters from the fits to the two experiments respectively) while providing roughly the same goodness of fit—a fact reflected in the relatively low BIC value for that model. Nevertheless, for ease of interpretation and comparison to earlier models, we focus on Model 2. This choice has little real consequence, because the parameters that were removed in Model 4 take on default values of zero, instead of their fitted values which were all close to zero.

The fitted parameters in Table 5 can be used to compute predicted accuracies for any condition in the experiments. For example, to determine the predicted accuracy for "A" with the "B" foil in Experiment 1, one adds together the perceivability for "A" (.666), the bias for "A" (.131), the dissimilarity between "A" and "B" (0), and subtracts the bias for "B" (.247). This estimates the log-odds accuracy for that condition to be $.666 + .131 - .247 = 0.55$, which corresponds to a probability of .634 (the actual accuracy for this condition was .636).

The model itself did not test for whether perceivability or bias differed reliably across experiments, but we can assess this with post-hoc tests. The standard error for each of these parameter estimates is roughly .072, and a critical difference indicating a reliable change across experiments is roughly 3.3 standard error units (for a two-tailed t test with $p < .05$, using the Bonferroni correction for 52 simultaneous contrasts.) According to this criterion, estimates for bias differed reliably across experiments for seven letters ("C", "G", "O", "Q", "S", "U", and "Y"), and estimates for perceivability differed reliably for 20 letters (all but "J", "K", "P", "R", "S", and "Z"). Bias was most impacted for those letters that were similar in shape to the "@" mask. In sum, this analysis suggests that masks have their greatest impact on the perceivability of letters, and to a lesser extent on the response biases for letters. In addition, only six out of 325 similarity parameters changed across experiments.

Many of these overall effects can be seen in Figure 4, which jointly displays estimates of bias and similarity across the two experiments. These effects are interesting because they reveal aspects of the perceptual decision system for letter stimuli, and possibly constrain models of perceptual processes. For example, consider the relative stability of similarity estimates across experiments, in contrast to perceivability. Theories of masking may hypothesize a number of processes of how perceivability is impacted, but whatever is assumed, they need to impact perceivability without impacting inter-

Table 4
Summary of four models predicting log-odds accuracy based on various predictor sets.

Model name	BIC	R^2	Adjusted R^2	RSE	F Statistic
1. Full model	4162	.9823	.9616	.252	$F(700,600) = 47.5$
2. Baseline + BIC-selected parameters	888	.9666	.9629	.2484	$F(132,1168) = 256.4$
3. Baseline model (γ and λ)	1046	.9555	.9517	.2833	$F(102,1198) = 252.3$
4. Smallest model	708	.965	.962	.2423	$F(96,1204) = 341$

Note: RSE = residual standard error (error sum of squares divided by the residual degrees of freedom)

Table 5
Reliable parameters from log-linear model using the BIC model selection technique. Smaller and more negative values indicate that a letter was less perceivable, biased against, or that a letter combination was discriminated less well than was expected by the perceivability and bias alone. The rightmost column shows difference in the similarity space between experiments, with negative values indicate the letters were more similar in Experiment 1 than Experiment 2.

	Experiment 1 ("@")		Experiment 2 ("#")							
Letter	Perceivability (λ)	Bias (γ)	Perceivability (λ)	Bias (γ)	Consistent Dissimilarity (δ)			Differing Similarity (δ)		
A*	0.666	0.131	1.282	0.061	B	G	-0.417	A	Q	-0.332
B*	1.391	0.247	1.083	0.096	B	K	0.366	C	Y	0.345
C*†	1.019	0.156	1.941	-0.226	B	Z	0.375	I	Z	0.378
D*	1.190	0.188	1.780	0.018	C	D	0.437	L	V	0.340
E*	1.385	-0.002	1.117	0.058	C	G	-0.779	R	Y	0.352
F*	1.286	0.027	1.012	0.097	C	O	-0.610	U	X	-0.349
G*†	1.192	0.156	1.673	-0.206	C	Q	-0.556			
H*	1.278	0.061	0.575	0.079	C	U	-0.637			
I*	1.320	-0.218	0.335	-0.337	D	G	-0.402			
J	1.045	-0.004	1.036	-0.136	D	O	-0.627			
K	1.216	0.182	1.080	0.096	D	Q	-0.426			
L*	1.261	-0.079	1.009	-0.174	D	U	-0.581			
M*	1.571	0.040	1.349	0.021	E	F	-0.394			
N*	1.497	-0.107	1.235	-0.086	G	Q	-0.442			
O*†	1.085	0.250	1.975	-0.164	G	U	-0.450			
P	1.205	0.154	1.384	0.003	H	N	-0.390			
Q*†	1.122	0.118	1.635	-0.202	I	L	-0.469			
R	1.434	0.170	1.420	0.087	M	N	-0.469			
S†	1.594	0.057	1.467	-0.210	M	W	-0.423			
T*	1.506	-0.115	0.855	-0.168	N	W	-1.132			
U*†	1.125	0.175	1.444	-0.123	O	Q	-0.671			
V*	1.691	0.093	1.186	0.068	Q	U	-0.409			
W*	1.323	0.024	0.978	0.040	R	S	-0.383			
X*	1.103	0.011	0.446	0.162	X	Y	-0.333			
Y*†	1.497	0.081	0.819	-0.122						
Z	1.326	0.000	1.357	0.000						

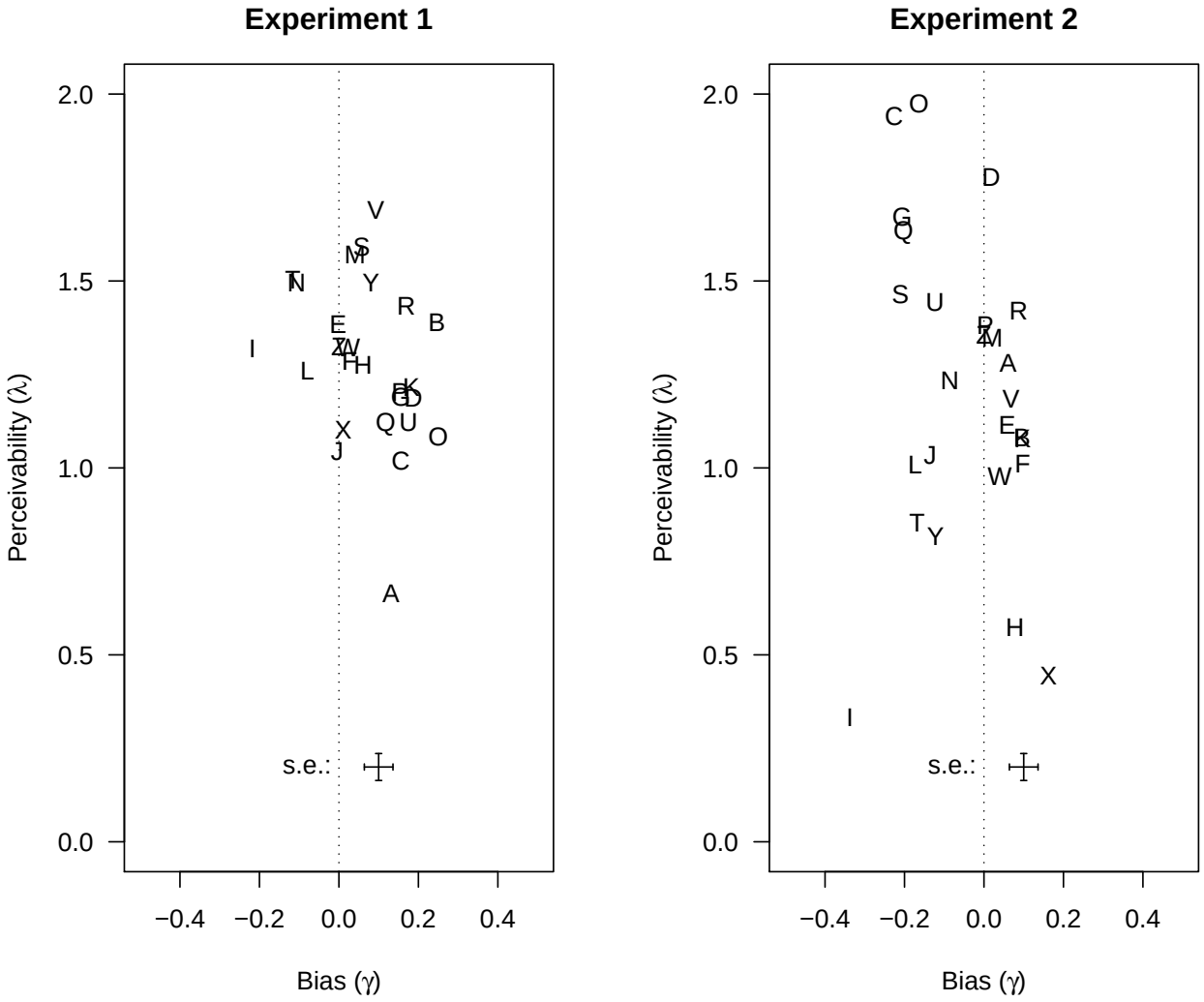
Note: * denotes letter for which estimates of perceivability differed reliably across experiments, and † denotes letters for which estimates of bias differed reliably across experiments.

letter similarity.

Relation of our model to the Biased Choice Model. A prominent model used for analyzing alphabetic similarity data is the called the "similarity" or "biased" choice model. This model was developed in various forms by Shepard (1957) and Luce (1963), and constitutes a simple model to predict recognition accuracy based on the number of alternatives, biases for making responses, and similarities between stimuli. It has often been used to decompose confusion matrices into similarity and response bias terms. Throughout its history, researchers have made various efforts at incorporating aspects equivalent to what we call perceivability. For example, Shepard (1957) described both stimulus and response weights to account for asymmetries in the response processes; and following him Nosofsky (1991) proposed stimulus and response biases (as well as both stimulus and response similarities). And more recently, (Rouder, 2004) has framed the *variable similarity choice model*, which accounts for variability in the strength of perceivability across experimental manipulations.

However, the analyses of the confusion matrices we reviewed in Table 1 typically decomposed accuracy into just similarity and response bias. Those experiments usually produced indirect measures of similarity (such as a naming accuracy), and so the number of alternatives theoretically in-

Figure 4. Bias and perceivability for the characters of the alphabet in Experiments 1 and 2. The average standard error estimate was 0.072. Letters had a larger range for perceivability than bias within experiments, and also differed more in perceivability than bias between experiments.



cludes the entire alphabet. According to the choice model, the probability of giving response j for a given stimulus i is:

$$p_{i,j} = \frac{\beta_j \eta_{i,j}}{\sum_k \beta_k \eta_{i,k}}, \tag{3}$$

where β is bias and η is similarity. The corresponding probability correct for a two-alternative response is:

$$p_{i,j} = \frac{\beta_i \eta_{i,i}}{\beta_i \eta_{i,i} + \beta_j \eta_{i,j}}, \tag{4}$$

In the Appendix, we show that the parameters estimated with our statistical model (λ, δ, and γ) are equivalent with transformation to corresponding parameters in choice theory. Yet in contrast to our statistical model, the choice model makes several theoretical assumptions about the nature of these factors, and especially about how number of options effects accuracy. For example, according to choice theory, accuracy depends on the number of options being considered, and although it assumes that the relative proportion of any two responses is unaffected by the introduction of additional alternatives, the probability of making any particular response will usually be reduced by adding alternatives. Despite the fact that the choice model performs surprisingly well (Nosofsky, 1991; Smith, 1992), these assumptions have also sometimes failed to be supported by appropriate empirical tests (e.g., Rouder, 2001, 2004; Ashby & Perrin, 1988). Thus, although our parameter estimates can be interpreted in terms of choice theory, they do not rely on the same psychological assumptions embodied by choice theory.

In sum, our statistical model can be interpreted in terms of choice theory, and our model demonstrates a method for es-

timating the theory's parameters that allow bias, perceivability, and similarity to be estimated, while selecting only those similarity values that account for variance. By applying it to the data from our experiments, we were able to identify a number of effects regarding the similarity space of the alphabet, and how perceivability is impacted by visual masks.

Discussion

In this paper, we identified more than 70 previously published studies of alphabetic letter similarity, dating back to research conducted by Helmholtz's proteges in the 19th century. Many of these studies were perceptual detection tasks, and across experiments three primary factors have been used to account for performance: perceivability, bias, and similarity. We conducted two experiments that enabled us to estimate these three factors simultaneously. In general, these three factors appear to account fairly well for our data, and exhibited the following primary effects across our experiments, two of which can be seen in Figure 4.

- Similarity effects were relatively stable across experiments;
- Perceivability effects differed substantially across experiments;
- Bias effects differed less strongly across experiments.

We also identified a number of secondary results, many of which can also be seen in Figure 4:

- Letters for whom bias differed reliably across experiments (mostly round letters) had *positive* bias for the experiment in which they were similar to the mask.
- In Experiment 2 (with the "#" mask), round letters had high perceivability and square letters had low perceivability; in Experiment 1 (with the "@" mask), this pattern was reversed, but less strong.
- Some letters ("I", "L", "T", and "N") were associated with *negative* biases for the experiment in which they were similar to the mask.
- Several letters appeared to be extreme outliers ("A" and "I" in Exp. 1; "I", "H", and "X" in Exp. 2.)

Our basic finding that bias effects were relatively small is probably a consequence of our empirical paradigm more so than an intrinsic lack of response bias induced by a masks. In fact, for the small subset of letters with reliable differences in bias across experiments (i.e., the round letters), biases were positive when these letters were similar to the mask, and negative when they were different from the mask. This indicates that characters similar to the mask can become attractive foils, perhaps because on trials in which no percept is seen, observers mistake the roundness features for features of the unseen target.

However, this bias effect is not universal: the letters "I", "L", "T", and "N" had negative biases in Experiment 2, even though they were arguably similar to the "#" mask. Interestingly, these were also the four most negatively biased letters in Experiment 1. What about these letters might lead to such a universal negative bias? Perhaps because "I", "L", and "T" are fairly simplistic stimuli, they become unattractive foils–maybe because a masked stimulus is typically very complex. Thus, if one used the perceived complexity of the stimuli to help make a decision, this would bias responses against these relatively simple stimuli.

There were a number of individual characters for which performance lies outside the bounds of typical stimuli. These include the "A" and "I" in Experiment 1, and the "X", "H", and "I" in Experiment 2. As just discussed, the "I" has universally negative bias, but has much lower perceivability in Experiment 2, where it was highly similar to the mask character. While it is not surprising that the perceivability of the "I", "X", and "H" are impacted by the "#" mask, it is perhaps surprising how poor performance actually becomes. These letters are particularly harmed by their similarity to the mask.

The "A" in Experiment 1 has a similar data pattern. However, this result is quite unexpected, because unlike the "I", "X", and "H" in Experiment 2, "A" does not have strong visual similarity to "@". Yet the "A" still does have a strong similarity to the mask, in that the "@" mask embeds a small letter "a" within it. This suggests that just as physical similarity between the mask and stimulus can reduce perceivability, so can more abstract similarity. One mechanism that could produce this effect is evidence discounting (cf. Huber et al., 2001; Weidemann et al., 2005 and Weidemann et al., 2008). According to this explanation, the ubiquitous presence of the "A" letter code in the mask means that evidence for it is nondiagnostic of the presence of an "A" target. Consequently, the perceptual system discounts evidence for that letter, leading to a lower probability of choosing it, even when it is the correct target. This hypothesis is somewhat speculative, and is difficult to separate from explanations such as high-level detector fatigue, or possibly even high-level masking. These accounts may require careful experimental controls to distinguish, which are beyond the scope of this paper.

Finally, we note that with proper experimental and statistical methods, the joint effects of perceivability, similarity, and bias can be estimated, and the obtained parameters can be interpreted in terms of choice theory; the most commonly-adopted model used in the past to interpret such data. Past approaches (especially for naming paradigms) have estimated only similarity and bias. Interestingly, we found that bias was relatively unimportant in our experiments, in contrast to the role of perceivability. Although this at least partly stemmed from our use of the 2-AFC procedure, it is nevertheless interesting that substantial variance can be attributed to a factor that has frequently been ignored. Our results suggest that perceivability should be the first place researchers look, rather than the last.

This suggestion is actually quite consistent with the approach taken by many of the early studies we reviewed in Table 1, dating back to experiments from 19th and early 20th century. Those early experiments often focused on ranking letters with respect to "legibility", but they would also call out a small number of highly-confusable letter pairs. Similarly, we found that perceivability figured prominently, whereas a relatively small subset (just 24 out of 325) of similarity parameters were necessary to account for the data. This result is also similar to many of the past results obtained using confusion matrices, in which most off-diagonal cells

were empty or had just a few confusions, whereas the observed confusions were concentrated in just a few pairs. This result is so consistent that it really appears to be a fundamental aspect of letter similarity.

References

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

Banister, H. (1927). Block capital letters as tests of visual acuity. *British Journal of Ophthalmology*, *11*, 49–61.

Bell, G. L. (1967). Effects of symbol frequency in legibility testing. *Human Factors*, *9*(5), 471–478.

Blommaert, F. J. (1988). Early-visual factors in letter confusions. *Spatial Vision*, *3*(3), 199–224.

Blough, D. S. (1985). Discrimination of letters and random dot patterns by pigeons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(2), 261–280.

Boles, D. B., & Clifford, J. E. (1989). An upper- and lowercase alphabetic similarity matrix, with derived generation similarity values. *Behavior Research Methods,Instruments, & Computers*, *21*, 579–586.

Bouma, H. (1971). Visual recognition of isolated lower-case letters. *Vision Research*, *11*, 459–474.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparison. *Biometrika*, *39*, 324–345.

Briggs, R., & Hocevar, D. J. (1975). A new distinctive feature theory for upper case letters. *Journal of General Psychology*, *93*(1), 87–93.

Cattell, J. M. (1886). Über die Tägheit der Netzhaut und des Sehcentrums. *Philosophische Studien*, *3*, 94–127.

Coffin, S. (1978). Spatial frequency analysis of block letters does not predict experimental confusions. *Perception & Psychophysics*, *23*, 69–74.

Cohen, A. L., & Sautner, M. (2001). *ExpLib (version 1.0.1 [beta])* [Computer Programming Library for Visual C++]. Retrieved from http://people.umass.edu/alc/explib/index.htm.

Courrieu, P., & de Falco, S. (1989). Segmental vs. dynamic analysis of letter shape by preschool children. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *9*(2), 189–198.

Craig, J. C. (1979). A confusion matrix for tactually presented letters. *Perception & Psychophysics*, *26*(5), 409–411.

Dockeray, F. C., & Pillsbury, W. B. (1910). The span of vision in reading and the legibility of letters. *Journal of Educational Psychology*, *1*, 123–131.

Dunn-Rankin, P. (1968). The similarity of lowercase letters of the English alphabet. *Journal of Verbal Learning and Verbal Behavior*, *7*(6), 990–995.

Dunn-Rankin, P., Leton, D. A., & Shelton, V. F. (1968). Congruency factors related to visual confusion of English letters. *Perceptual & Motor Skills*, *26*, 659–666.

Engel, G. R., Dougherty, W. C., & Jones, G. B. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, *27*(3), 317–326.

Eriksen, C. W. (1980). The use of a visual mask may seriously confound your experiment. *Perception & Psychophysics*, *28*(1), 89–92.

Fisher, D. F., Monty, R. A., & Glucksberg, S. (1969). Visual confusion matrices: Fact or artifact? *The Journal of Psychology*, *71*, 111–125.

Gervais, M. J., Harvey, L. O., & Roberts, J. O. (1984). Identification confusions among letters of the alphabet. *Perception & Psychophysics*, *10*, 655–666.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*, 487–490.

Geyer, L. H., & DeWald, C. G. (1973). Feature lists and confusion matrices. *Perception & Psychophysics*, *14*, 471–482.

Gibson, E. J. (1969). *Principles of learning and development*. New York: Meredith.

Gibson, E. J., Osser, H., Schiff, W., & Smith, J. (1963). An analysis of critical features of letters tested by a confusions matrix. In *A Basic Research Program on Reading: Cooperative Research Project No. 639* (p. 1-22). Ithaca, NY: Cornell University.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, *25*, 425–431.

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

Gupta, S. M., Geyer, L. H., & Maalouf, J. A. (1983). Effect of font and medium on recognition/confusion. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 144-149.

Hartridge, H., & Owen, H. B. (1922). Test types. *British Journal of Ophthalmology*, *6*, 543–549.

Heiser, W. J. (1988). Selecting a stimulus set with prescribed structure from empirical confusion frequencies. *British Journal of Mathematical & Statistical Psychology*, *41*(1), 37–51.

Hodge, D. C. (1962). Legibility of a uniform-strokewidth alphabet: I. Relative legibility of upper and lower case letters. *Journal of Engineering Psychology*, *1*, 23–46.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Quach, R. (2002). Mechanisms of source confusion and discounting in short-term priming 2: Effects of prime similarity and target duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1120–1136.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, *108*, 149–182.

Huber, D. E., Shiffrin, R. M., Quach, R., & Lyle, K. B. (2002). Mechanisms of source confusion and discounting in short-term priming: 1. Effects of prime duration and prime recognition. *Memory & Cognition*, *30*, 745–757.

Jacobs, A. M., Nazir, T. A., & Heller, O. (1989). Perception of lowercase letters in peripheral vision: A discrimination matrix based on saccade latencies. *Behavioral and Brain Sciences*, *46*(1), 95–102.

Javal, M. (1881). L'évolution de la typographie considérée dans ses rapports avec l'hygiène de la vue. *Revue Scientifique*, *XXVII*, 802–813.

Johnson, J. R., & Phillips, K. O. (1981). Tactile spatial resolution. i. two-point discrimination, gap detection, grating resolution, and letter recognition. *Journal of Neurophysiology*, *46*, 1177–1192.

Keren, G., & Baggen, S. (1981). Recognition models of alphanumeric characters. *Perception & Psychophysics*, *29*, 234–246.

Kikuchi, T., Yamashita, Y., Sagawa, K., & Wake, T. (1979). An analysis of tactile letter confusions. *Perception & Psychophysics*, *26*(4), 295-301.

Korte, W. (1923). Über die Gestaltauffassung im indirekten Sehen. *Zeitschrift für Psychologie*, *93*, 17–82.

Kuennapas, T., & Janson, A. (1969). Multidimensional similarity of letters. *Perceptual & Motor Skills*, *28*, 3–12.

Laughery, K. R. (1969). Computer simulation of short-term memory: A component-decay model. In G. H. B. . J. T. Spence (Ed.), *The psychology of learning & motivation: Advances in resarch and theory* (Vol. 3, pp. 135–200). New York: Academic Press.

Liu, L., & Arditi, A. (2001). How crowding affects letter confusion. *Optometry and Vision Science*, *78*, 50–55.

Loomis, J. M. (1974). Tactile letter recognition under different modes of stimulus presentation. *Perception & Psychophysics*, *16*, 401–408.

Loomis, J. M. (1982). Analysis of tactile and visual confusion matrices. *Perception & Psychophysics*, *31*, 41–52.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Luce, R. D. (1963). Handbook of mathematical psychology, vol 1. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), (chap. Detection and recognition). New York: Wiley.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" measures. *Psychological Bulletin*, *107*, 401–413.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide, 2nd edition*. New York: Lawrence Erlbaum Associates.

Mayzner, M. S. (1975). Information processing and cognition: The loyola symposium. In R. L. Solso (Ed.), (chap. Studies of visual information processing in man). Hillsdale, New Jersey: L. Erlbaum Associates.

McGraw, G., Rehling, J., & Goldstone, R. (1994). Letter perception: Toward a conceptual approach. In *Proceedings of the sixteenth annual conference of the cognitive science society* (pp. 613–618). Atlanta, GA: The Cognitive Science Society.

Miozzo, M., & De Bastiani, P. (2002). The organization of letter-form representations in written spelling: Evidence from acquired dysgraphia. *Brain and Language*, *80*, 366–392.

Mueller, S. T., & Weidemann, C. T. (2008a). *Alphabetic letter similarity matrices: Effects of bias, perceivability and similarity.* Manuscript submitted for publication.

Mueller, S. T., & Weidemann, C. T. (2008b). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin and Review*, *15*, 465–494.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*, 94–140.

Ovink, G. W. (1938). *Legibility, atmosphere-value, and forms of printing types*. Leiden, Holland: A. W. Sijthoff's Uitgeversmij N. V.

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation-verification model for letter and word recognition: The word-superiority effect. *Psychological Review*, *89*, 573–594.

Phillips, J. R., Johnson, K. O., & Browne, H. M. (1983). A comparison of visual and two modes of tactual letter resolution. *Perception & Psychophysics*, *34*(3), 243–249.

Podgorny, P., & Garner, W. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics*, *26*, 37–52.

Popp, H. M. (1964). Visual discrimination of alphabet letters. *The Reading Teacher*, *17*, 221–226.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria. (ISBN 3-900051-07-0)

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, *104*, 319–343.

Ratcliff, R., McKoon, G., & Verwoerd, M. (1989). A bias interpretation of facilitation in perceptual identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 378–387.

Reich, L. N., & Bedell, H. E. (2000). Relative legibility and confusions of letter acuity in the peripheral and central retina. *Optometry and Vision Science*, *77*(5), 270–275.

Roethlein, B. E. (1912). The relative legibility of different faces of printing types. *American Journal of Psychology*, *23*, 1–36.

Rouder, J. N. (2001). Absolute identification with simple and complex stimuli. *Psychological Science*, *12*, 318–322.

Rouder, J. N. (2004). Modeling the effects of choice-set size on the processing of letters and words. *Psychological Review*, *111*, 80–93.

Sanford, E. C. (1888). The relative legibility of the small letters. *American Journal of Psychology*, *1*, 402–435.

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

Smith, J. E. K. (1992). Alternative biased choice models. *Mathematial Social Sciences*, *23*, 199–219.

Thorson, G. (1976). An alternative for judging confusability of visual letters. *Perceptual & Motor Skills*, *42*(1), 116–118.

Tinker, M. A. (1928). The relative legibility of the letters, the digits, and of certain mathematical signs. *Journal of General Psychology*, *1*, 472–496.

Townsend, J. T. (1971a). Theoretical analyses of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*, 40–50.

Townsend, J. T. (1971b). Alphabetic confusion: A test of models for individuals. *Perception & Psychophysics*, *9*, 449–454.

Uttal, W. R. (1969). Masking of alphabetic character recognition by dynamic visual noise (DVN). *Perception & Psychophysics*, *6*, 121–128.

van der Heijden, A. H. C., Malhas, M. S., & van den Roovart, B. P. (1984). An empirical interletter confusion matrix for continuous-line capitals. *Perception & Psychophysics*, *35*, 85–88.

van Nes, F. L. (1983). New characters for Teletext with improved legibility. *IPO Annual Progress Report*, *18*, 108–113.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer. (ISBN 0-387-98825-4)

Watson, A. B., & Fitzhugh, A. E. (1989). Modelling character legibility. *Society for Information Display Digest of Technical Papers*, *20*, 360–363.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (2005). Confusion and compensation in visual perception: Effects of spatiotemporal proximity and selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 40–61.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (2008). Prime diagnosticity in short-term repetition priming: Is primed evidence discounted, even when it reliably indicates the correct answer? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 257–281.

Appendix
Correspondence of Logistic Model to Luce's (1963) Choice Theory.

For the correspondence between choice theory and our statistical model to make sense, we must make several assumptions about the meaning of different parameters. First, we must distinguish between a number of corresponding stimulus classes: distal stimuli (s_i), the perceived target (s_i'), long-term perceptual memory for each letter ($\hat{s_i}$), and the perceived response alternative with little noise and strong bottom-up support ($\bar{s_i}$). Although applications of choice theory have typically not distinguished between these, they are in principle distinct. If λ estimates the similarity between the perceived stimulus and the correct response option (s_i' and $\bar{s_i}$), this corresponds (with appropriate transformation) to the usual interpretation of $\eta_{i,i}$, which is typically assumed to be 1.0. Note that this interpretation of our theory places the role of perceivability at the comparison process between proximal stimulus and response standard. Our bias parameters match fairly directly (with proper transformation) to their corresponding notions in choice theory, and our dissimilarity parameters are interpreted as the degree to which the percept and the response alternative match, and have a 1:1 correspondence (with transformation) to $\eta_{i,j}$.

Luce (1963) defined choice theory to model response probabilities for a two-alternative case:

$$p_{i,j} = \frac{\beta_i \eta_{i,i}}{\beta_i \eta_{i,i} + \beta_j \eta_{i,j}} \tag{5}$$

where η is a measure of similarity between two stimuli, and β as a response bias. The theory (cf. Luce, 1963) makes three primary assumptions:

- A1. For all s_i, $s_j \in S$, $\eta_{s_i,s_j} = \eta_{s_j,s_i}$.
- A2. For all $s \in S$, $\eta_{s,s} = 1$.
- A3. For all s_i, s_j, and $s_k \in S$, $\eta_{s_i,s_k} \geq \eta_{s_i,s_j}\eta_{s_j,s_k}$.

For an indirect measurement procedure such as naming, one compares the noisy percept to the perceptual memory for all letter options, and so one technically estimates $\eta_{i',\hat{j}}$, where i' represents the noisy percept, and $\hat{j}$ represents the memory for a letter. Researchers have often assumed that $\eta_{i',\hat{i}} = 1$ (e.g., Townsend, 1971a, 1971b), which is not the same as assumption A2 of choice theory, which assumes that $\eta_{x,x} = 1$ for whatever x represents, and making this assumption essentially defines the unit of η (Luce, 1963). In contrast, the assumption that $\eta_{i',\hat{i}} = 1$ constrains the similarity between the percept and the perceptual memory to be 1.0, which essentially assumes perfect perceivability.

Our statistical model is applied to a 2-AFC task, where the noisy percept (i') is compared to two stimuli with strong perceptual support (i and j). We suppose that the log-odds of a correct response $p_{i,j}$ is proportional to the influence of three factors:

$$log(\frac{p_{i,j}}{1-p_{i,j}}) \propto \lambda_i + \gamma_i - \gamma_j + \delta_{i,j}, \tag{6}$$

where γ is an measure of bias, λ is a measure of perceivability, and δ is a measure of dissimilarity.

By making the following substitutions into Equation 6:

$$\begin{aligned} \gamma_i &= log(\beta_i) \\ \lambda_i &= log(\eta_{i',\bar{i}}) \\ \delta_{i,j} &= -log(\eta_{i',\bar{j}}) \end{aligned}$$

one obtains the equation

$$log(\frac{p_{i,j}}{1-p_{i,j}}) = log(\eta_{i',\bar{i}}) + log(\beta_i) - log(\beta_j) - log(\eta_{i',\bar{j}}). \tag{7}$$

Equation 7 can be solved for $p_{i,j}$ with the following intermediate steps:

$$\begin{aligned} \frac{p_{i,j}}{1-p_{i,j}} &= \frac{\beta_i \eta_{i',\bar{i}}}{\beta_j \eta_{i',\bar{j}}}, \\ \frac{1-p_{i,j}}{p_{i,j}} &= \frac{\beta_j \eta_{i',\bar{j}}}{\beta_i \eta_{i',\bar{i}}}, \\ \frac{1}{p_{i,j}} - 1 &= \frac{\beta_j \eta_{i',\bar{j}}}{\beta_i \eta_{i',\bar{i}}}, \\ \frac{1}{p_{i,j}} &= 1 + \frac{\beta_j \eta_{i',\bar{j}}}{\beta_i \eta_{i',\bar{i}}}, \\ \frac{1}{p_{i,j}} &= \frac{\beta_i \eta_{i',\bar{i}} + \beta_j \eta_{i',\bar{j}}}{\beta_i \eta_{i',\bar{i}}}, \\ p_{i,j} &= \frac{\beta_i \eta_{i',\bar{i}}}{\beta_i \eta_{i',\bar{i}} + \beta_j \eta_{i',\bar{j}}}, \end{aligned} \tag{8}$$

After applying these algebraic steps, Equation 8 is identical to Equation 5. This reiterates how $\eta_{i',\bar{i}}$ is explicitly related to perceivability, and shows that the common assumption that $\eta_{i',\hat{i}} = 1$ is a technical assumption and not one based on theoretical assumptions of choice theory. In our model, one still assumes that $\eta_{i,i} = 1$, and so it does not violate assumption A2 of choice theory.

Interpretation of our obtained parameters in terms of choice theory is straight-forward, although there is one point of caution. The obtained value δ corresponds roughly to traditional distance measures used in choice theory (cf. Luce, 1963), but our obtained measures are mostly negative, indicating high confusability. This corresponds to values of η that are greater than 1.0, but η is often constrained to be between 0 and 1, with 1.0 corresponding to "identical". This creates a problem because if response biases are ignored, one could produce situations where response tendencies still place accuracy below 0.5 for a forced-choice task. To avoid this situation, $\eta_{i,j}$ must be smaller than $\eta_{i,i}$ for all i and j, which corresponds to $-\delta_{i,j} < min(\lambda_i, \lambda_j)$. Violations of this would indicate that choice theory provides an inadequate account of our data. This was never the case for the values of similarity and perceivability estimated in our experiments.