Journal of Memory and Language **47,** 145–171 (2002)
doi:10.1006/jmla.2001.2835

Phonetic Biases in Voice Key Response Time Measurements

Brett Kessler and Rebecca Treiman

Wayne State University

and

John Mullennix

University of Pittsburgh at Johnstown

Voice response time (RT) measurements from 4 large-scale studies of oral reading of English monosyllables were analyzed for evidence that voice key measurements are biased by the leading phonemes of the response. Words with different initial phonemes did have significantly different RTs. This effect persisted after contributions of nine covariables, such as frequency, length, and spelling consistency, were factored out, as well as when variance associated with error rate was factored out. A breakdown by phoneme showed that voiceless, posterior, and obstruent consonants were detected later than others. The second phonemes of the words also had an effect on RT: Words with high or front vowels were detected later. Phoneme-based biases due to voice keys were large (range about 100 ms) and pervasive enough to cause concern in interpreting voice RT measurements. Techniques are discussed for minimizing the impact of these biases.

Key Words: voice key; response time measurement; naming task; oral reading; measurement bias.

Much research in psychology uses measures of vocal response latencies to make inferences about underlying processes. In some experiments, the time that it takes to initiate pronunciation of a printed word is used to study the processes involved in reading and word recognition (e.g., Forster & Chambers, 1973). Other studies use the time to start repeating a spoken word to shed light on auditory word recognition (e.g., Connine, Mullennix, Shernoff, & Yelen, 1990). In still other experiments, spoken responses are used to examine higher levels of language comprehension (e.g., Stanovich & West, 1983). By having participants name pictures instead of printed words, researchers can address issues about how speakers retrieve and produce words (e.g., Griffin & Bock, 1998). Vocal response times are used in studies of memory as well (e.g., Scarborough, Cortese, & Scarborough, 1977). The naturalness of spoken responses makes such tasks well suited to a variety of populations and a variety of issues.

By far the predominant technology used for determining voice response times, and the one used in the studies cited above, is the voice key. A voice key is a device that determines voice onset in real time. The voice key is connected to a microphone, which converts sound pressure (the physical correlate of the amplitude of the sound) into voltage. When a stimulus has been presented, a computer arms the voice key, which begins monitoring the microphone. When the sound pressure reaches a predefined target, the voice key is triggered. It notifies the computer, which stores the number of milliseconds that elapsed between arming and triggering the voice key. The typical voice key is triggered as soon as the sound pressure reaches a certain level, which the experimenter can predefine (e.g., Cedrus, 2000; Psychology Software Tools, 2000).

This research was supported by National Science Foundation Grant BCS-9807736. We thank the anonymous respondents to our Internet poll; Nancy Ciaparro, Kira Rodriguez, and Joe Inman for their assistance with the research; Howard Nusbaum, David Balota, Daniel Spieler, Mark Seidenberg, and Gloria Waters for sharing their data; and Christopher Kello, Stephen Lindsay, Ronald Peereman, and Kathleen Rastle for helpful comments on an earlier draft.

Address correspondence and reprint requests to Brett Kessler, Department of Psychology, Wayne State University, 71 W. Warren Avenue, Detroit, MI 48202. Fax: (313) 577-7636). E-mail: bkessler@brettkessler.com.

0749-596X/02 $35.00

Traditional voice keys are not the only ways of measuring voice response times. More sophisticated voice keys can be configured to be triggered by a lower level of sound pressure, provided that it is maintained for a certain length of time (e.g., Hutzler, 1999; Rastle & Davis, 2002). Digital signal processing techniques can compute voice response time algorithmically, or produce waveforms for visual inspection (e.g., Bachoud-Lévi, Dupoux, Cohen, & Mehler, 1998; Fushimi, Ijuin, Patterson, & Tatsumi, 1999; Kawamoto, Kello, Jones, & Bame, 1998; Morrison & Ellis, 1995). Although some of our discussion will be relevant to these more sophisticated methods, this paper concentrates on voice keys, which are the source of timing data for the large majority of published studies that use measurements of vocal response times.

The general goal of this paper is to increase our understanding of the potential inaccuracy of voice keys. In particular, we focus on phonetic bias. Psychologists often wish to compare reaction times across stimuli that provoke different word or word-like responses. They are rarely directly interested in response time differences that are due entirely to phonetic differences between responses. Unfortunately, there are strong a priori reasons for believing that systematic phonetic biases exist. A response beginning with /s/, for example, might systematically be detected later, or sooner, than a word beginning with /m/, yielding different measured response times for no other reason than the phonetics of the words involved.

One obvious reason that phonemes might be detected at different times is articulatory: Some sounds take more time for the human vocal apparatus to initiate. Sakuma, Fushimi, and Tatsumi (1997) reported, for example, that /s/ can be initiated especially quickly. The second major factor is acoustic. Even after phonemes are initiated, it may take the voice key different amounts of time to detect them. The culprit in this case is the sound pressure cutoff that voice keys use. Setting the voice key too low (making it too sensitive) results in an unacceptable number of suspiciously fast response times as the voice key is triggered by nonspeech noises such as the participant's breathing. However, any setting higher than zero means that part of the beginning of the vocal response will be missed, because all utterances have a nonzero rise time to target amplitude. This acoustic factor can result in phonetic bias, because different phonemes have different rise times. For example, fricatives such as /s/ have less sound pressure than vowels (Fry, 1979; Sacia & Beck, 1926). Because voice keys are triggered only after a certain sound pressure level is detected, we might expect that under some circumstances, at least, voice keys may take longer to register a fricative than to register a vowel, or, indeed, may not register the fricative at all (Sakuma, et al., 1997; Rastle & Davis, 2002).

Although we focus here on phonetic biases, it is useful to note that this acoustic factor can result in other types of problems with voice keys. If for any reason one utterance is spoken at a faster tempo than another, then the faster utterance will more quickly reach the point in the speech stream where the target amplitude occurs. For example, if initial /s/ is routinely missed, then the faster utterance will more quickly get to the following phoneme, and therefore have the quicker measured response time, even if the veridical times to initiate the responses are identical. That could introduce much variation into the data. It could even result in experimental bias if factors that influence response time also influence tempo, as we have good reason to expect (Kawamoto et al., 1998). Such effects would, of course, interact with phonetic biases. If a voice key has special trouble detecting initial /s/, then artifactual effects of slow speech will affect /s/-initial words disproportionately. In addition, differences in overall amplitude of the utterance can have effects similar to those caused by differences in tempo. In a word like *grow*, the /ɹ/ is louder than the /g/, and the /o/ is louder still. If the utterance as a whole is very loud, a voice key might trigger at the /g/. If the utterance is less loud, so that all of the phonemes are proportionately softer, perhaps only the /ɹ/ or even the /o/ will have enough acoustic energy to trigger the voice key. Of course the tempo of the utterance will largely

determine how long it takes to get to those later phonemes. Thus voice key measures can conflate veridical response time, tempo, and amplitude in a complex way that interacts with the differential acoustic natures of the phonemes in the response.

Our specific objectives here are threefold. First, we wish to determine what the magnitude of phonetic bias actually is in voice key studies. Is it a purely theoretical consideration of trifling import, perhaps amounting to errors of a few milliseconds, or is it a much larger problem that could seriously impact typical studies? Second, we seek to discover how pervasive the problem is. Could it be solved, for example, by treating /s/-initial words separately from other words, or does it pervade the entire phonemic inventory? Is the problem limited to the first phoneme, or might it extend more deeply into the word? Does phonetic bias affect different experiments equally strongly? Third, we try to separate true phonetic bias from associations that might be mediated by psycholinguistic factors. For example, if words beginning with /z/ consistently had slower measured response times than other words, we would want to rule out the possibility that that is due, perhaps, to the fact that /z/-initial words tend to be of lower frequency and familiarity. One potential objective we do not adopt is that of separating out the articulatory from the acoustic source of phonetic bias. That would be of little practical benefit to those who wish to use voice keys or evaluate earlier studies that used them. For those people, the two sources of phonetic bias form a bundled entity whose magnitude and variability needs to be dealt with or evaluated as a unit.

SURVEYS OF CURRENT PRACTICE

Awareness about possible bias in voice keys varies greatly. To gather information about researchers' concerns and beliefs, we conducted an informal Internet poll during the summer of 2000, to which 60 people responded. Of them, 55 indicated that they understood that different initial phonemes could differentially affect the triggering of voice keys. That understanding was unanimous among the 39 who had published research that used voice keys. In an open-ended question about what factors may be involved in such bias, several respondents mentioned that the intensity or, more precisely, the intensity rise time of the initial phoneme may affect how quickly the phoneme is detected, if at all. Many respondents stated some observations in articulatory terms. Few individual respondents offered more than a small part of the story, however. The most widely cited belief, that voiced phonemes are detected faster than voiceless ones, was mentioned by only about one third of the respondents.

While 92% of the respondents believed the first phoneme has an effect, only 45% believed that subsequent phonemes have an effect; that proportion crept up to 53% for researchers who had published voice key experiments. About one third of the respondents (19) suggested that the second phoneme may be important if the voice key is not triggered by the first phoneme at all. Almost as many (16) opined that the second phoneme could affect voice key responses by modifying the pronunciation of the first phoneme. However, few specifics were offered, and many volunteered that they considered this a theoretical possibility that they had never seen documented.

What about researchers' actual practices? To gather information about how experiments are conducted, we surveyed the 220 articles published from 1997 through 2000 in the *Journal of Memory and Language.* (All of these articles were accepted for publication before the second author of the present article, who edited many of the surveyed articles, was aware of most of the issues discussed here.) Our survey focused on experiments that compared vocal response time across different sets of stimuli. Some of the experiments used the same stimuli; we counted these groups of experiments only once in our summary statistics, giving 48 cases in total. The authors of a number of the articles were aware of some potential problems with voice key measurement and tried to deal with these problems in some way. These solutions will be discussed more fully later in this paper, but we introduce them here briefly. In 5 cases (all in one paper), the acoustic biases of voice keys were avoided by using a digital signal de-

tection algorithm to find the onset of the response. In 12 cases, a delayed naming task or similar control task was used in addition to a standard naming task. Delayed naming is an attempt to isolate acoustic influences on voice key measurements from psycholinguistic factors. Another tack, used in many of the 36 cases that did not use delayed naming, was to balance the stimuli by phonemic criteria. For example, if experimenters believed that the voice key is biased by the first phoneme of the word, they might ensure that each group of stimuli had an equal number of words beginning with each phoneme. We found a clear difference in how experimenters treated initial phonemes and following phonemes. Experimenters attempted to balance the sets of stimuli for the initial phoneme in 20 of the 36 cases. As a reason for this practice, the authors sometimes stated that different word-initial phonemes trigger the voice key at different times. Those authors who attempted to balance the stimuli for initial phonemes sometimes indicated that they were not able to do so in all cases. In such cases, they often attempted to match a phoneme by one that was similar, but they usually gave no details about what standard of similarity was used. In another 15 of the 36 cases, the authors did not report any attempt to equate the stimuli for initial phonemes. One additional case revealed an intermediate course of action: The authors equated the stimuli for the manner class of the initial phoneme. Two other authors also mentioned manner class as being important in determining when the voice key triggers, one of the two offering the specific hypothesis that plosives trigger the voice key more quickly than fricatives.

The situation was different for the second phoneme. There was an attempt to match the stimuli for the second phoneme in 8 of the 36 stimulus sets for which delayed naming was not used. In the remaining 28 cases, no such attempt was made.

Few authors reported the specific type of voice key used or the conditions of its use. The *Journal*'s guidelines stated that details of equipment need not be reported unless essential for replication, and most authors seemed to feel that the make of the voice key had no important effects.

Combining the results of our survey of experimenters' beliefs and our survey of publications, it appears that experimenters who publish work using voice keys are aware that different initial phonemes can differentially affect the triggering of voice keys. In practice, however, experimenters do not always equate the stimuli that they are comparing across conditions for initial phoneme. The gap between belief and practice may reflect the fact that such matching is often difficult to achieve given the other constraints on stimulus selection. In addition, some researchers appear to believe that the effects are relatively small and will wash out if there are a reasonable number of stimuli. Far fewer published studies equate stimuli for their second phonemes than for their initial phonemes. In addition, fewer researchers endorse the idea that the second phoneme is important. Researchers also seem to expect that different voice key setups will yield similar results.

PRIOR RESEARCH

Although none of the authors in the survey of articles from the *Journal of Memory and Language* cited published research to support their views about voice keys, there is some helpful research that contrasts the timings obtained by voice keys with more accurate timings inferred by visual inspection of the voice waveform. The impact of this research has been limited by the fact that several studies were written in languages other than English and investigated languages other than English. The first such study we know of was Pechmann, Reetz, and Zerbst (1989). They asked their participants to read the same word five times and measured the voice key error by comparing the response time as measured by the voice key to much more accurate measures determined by analyzing waveforms. They reported not only high errors, but also large differences between trials. For example, the word *Freude* was, on average, detected by the voice key 104 ms after it was visible in the waveform, and the average range of voice key errors for a given participant was 98 ms. Most words had smaller errors, but the

wide variation between words was perhaps more alarming than high but uniform error rates would have been. While a reasonable interpretation is that phonetic bias underlies the between-word variation, Pechmann et al. did not measure more than one word with the same initial phoneme, so we do not know conclusively to what extent the initial phoneme itself was the cause of the errors. Sakuma et al. (1997) ran experiments contrasting most of the possible word-initial phonemes in Japanese. They found that the voice key was always slower than the waveform measurement and that the difference was biased by manner of articulation: Vowels and nasals had a small difference from the waveform value, the liquid had more, plosives more yet, and voiceless fricatives had a very great difference. The range of the difference between the phoneme groups was about 95 ms.

Possible contributions of the second phoneme of the word to phonetic bias have been studied less. Rastle and Davis (2002) were the first to directly investigate the possible effects of voice key artifacts caused by the second phoneme of the word. Their experiments with speakers of British English contrasted words with simple (/s/ plus vowel) and complex (/s/ plus /t/ or /p/) onsets. When the data were studied by waveform, words with complex onsets were named faster than those with simple onsets. However, when a voice key of typical design was used, the reverse pattern was obtained. These results document how voice key artifacts may lead researchers to draw false conclusions when investigating questions such as the effect of onset complexity on the reading of printed words.

Voice waveforms are much more cumbersome to use than voice key timings, so previous researchers who directly compared the two measurement types have understandably restricted their experiments to a few stimulus types. Consequently, many questions remain. For example, does the second phoneme affect the measured naming latencies in words that do not begin with the /s/ plus plosive clusters studied by Rastle and Davis (2002), or are such (relatively infrequent) word-initial sequences of two voiceless phonemes an isolated special case? We also noted that languages differ in the way nominally equivalent phonemes are pronounced, and asked whether the results obtained for German, Japanese, and British English hold for North American English. Because of these outstanding issues, we perceived the need for a larger-scale study, using North American English. Because we investigated phonetic bias as a whole and did not seek to factor the articulatory from the acoustic components, we did not need to use voice waveforms. This freedom allowed us to economically run thousands of trials and also to compare our results with those of previous studies.

In the experiment reported here, we presented a large number of words—virtually all the familiar one-syllable monomorphemic words of English—to 20 native speakers in standard orthography. We recorded their response times using a typical voice key. Known or suspected covariables such as word frequency were accounted for statistically. Some (but not all) of the lexical covariables could have been reduced by asking participants to repeat nonsense syllables instead of reading words. However, we placed priority on replicating the conditions of speeded naming tasks, which are a more typical application for voice keys; we did not want to accidentally pass over any stage of the task which might introduce phonetic bias, whether it be articulatory or acoustic in origin. The task also allowed us to investigate variability between experiments by affording comparison with three other naming megastudies. These are the studies of Seidenberg and Waters (1989; henceforth SW) at McGill University; Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty (1995; henceforth TMBR), which studied consonant–vowel–consonant words at Wayne State University; and Spieler and Balota (1997; henceforth SB) at Washington University. We will refer to our own study as KTM. See Table 1 for the number of participants in each study, and the number of words that were tested. The KTM and TMBR studies used the voice key included in the response box supplied with the MEL software package (Schneider, 1988). In both of these studies, the voice key was left at the same fixed setting for all trials. SB used a Gerbrands

TABLE 1

Summary Statistics for Each Study

Measure	KTM	TMBR	SB	SW
Participants	20	27	31	30
Words, tested	3690	1327	2870	2897
Words, analyzed	2982	1234	2525	2536
RT (ms)	629	611	467	568
Error rate[a]	4.6	1.3	—	6.0
Bigram frequency[b]	1493	1816	1529	1531
Consistency of onset[c]	.976	.966	.975	.976
Consistency of rime[c]	.913	.906	.908	.907
Familiarity[d]	6.6	6.6	6.7	6.7
Frequency of spelling[e]	3172	2526	2559	2584
Homophones[f]	1.1	1.2	1.1	1.1
Length of pronunciation[g]	3.5	3.0	3.5	3.5
Length of spelling[h]	4.4	4.1	4.4	4.4
Neighborhood size[i]	6.8	9.4	7.1	7.1

Note. Averages across words. For RT and error rate, the per-word measures are themselves averages across trials.

[a] Percentage of mispronounced responses, excluding those with outlying response times or failures to respond. Data unavailable for SB.

[b] Average text frequency of the two-letter sequences in the spelling (Solso & Juel, 1980).

[c] Proportion (by type counts) of words in the list that have the same spelling that also share the pronunciation (Treiman et al., 1995).

[d] Nusbaun, Pisoni, and Davis (1984).

[e] Corpus frequency in Zeno, Ivenz, Millard, and Duvvuri (1995).

[f] Number of words in list with this pronunciation (minimum = 1).

[g] Number of phonemes.

[h] Number of letters.

[i] Number of words that differ by substituting one letter (Coltheart, Davelaar, Jonasson, & Besner, 1977).

Model G1341T voice-operated relay, and SW used a custom-made voice key.

METHOD

Participants

The participants were 20 undergraduate students from Wayne State University. Three participants dropped out from an original group of 23, leaving 6 men and 14 women. The students received course credit in exchange for participation. All were native speakers of English. None had a history of speech or hearing disorder, and none had uncorrected visual problems.

Materials

A set of 3690 words was taken from a computerized version of the *Merriam-Webster Pocket Dictionary* and *Webster's Seventh Collegiate Dictionary* that was developed for such studies as Nusbaum, Pisoni, and Davis (1984). Homographic heterophones (e.g., *wind*) and many unfamiliar words were excluded, but the list included many words that differ only in inflectional ending (e.g., both *blow* and *blown* were included) as well as many words unknown to the great majority of college undergraduates (e.g., *gneiss*). In order to avoid the statistical noise that would be introduced by redundancy and by the inclusion of unknown words, we performed our analyses only on the 2982 words that also occur in the word list used for Kessler and Treiman (2001). The latter list was carefully controlled to have only morphologically nonredundant words that are generally familiar to college undergraduates. The selection of this subset was made before we looked at any of the experimental data.

We also analyzed, separately, the data from the three previous megastudies. Table 1 shows

the number of words remaining in all four studies after intersecting their stimulus list with that of Kessler and Treiman (2001).

Procedure

Participants were instructed to read aloud words that appeared on a computer screen, as quickly and accurately as possible. On each trial, the prompt GET READY appeared for 2 s in the center of the screen. The screen then went blank for 1 s. The stimulus word was then presented in uppercase letters. It remained on the screen until the voice key picked up the spoken response. Then the screen went blank for 1 s until the next warning message. After every 100 trials, the participants received a 1-min break.

All of the 3690 monosyllabic English words were presented to each subject, in eight sessions that each tested from 461 to 463 words. Words were presented in a different, randomly selected, order for each subject. Participants were advised not to make any extraneous sounds that could trigger the voice key. They were asked to keep their lips about 4 inches from the microphone throughout the experiment. An Electrovoice RE16 cardioid microphone was used together with the aforementioned voice key (Schneider, 1988). Trials with response times quicker than 100 ms or slower than 2000 ms were rejected from the analysis. The remaining trials were hand-coded to indicate whether the pronunciation was in error. A subset of the responses (one randomly selected 461-word list from each of five subjects) was checked by two different judges to determine accuracy of this hand-coding; 96.7% of the judgments agreed on the correctness or incorrectness of a response. For the purpose of error analyses, mispronunciations were counted as errors even when the participants subsequently corrected themselves.

RESULTS AND DISCUSSION

We first present analyses dealing with voice key effects caused by the word-initial phoneme. We then discuss effects of the second phoneme.

Initial Phoneme

Overall analyses. First we grouped words by their initial phoneme and asked whether those groups differed significantly in their average response times. We chose a Monte Carlo test of significance (Good, 1995), as illustrated with a small subset of the data in Table 2. We directly determined the likelihood that our F ratio could be due to chance by randomly rearranging the data between phoneme groups 10,000 times, subject to the constraint that each group contains a constant number of words across rearrangements and seeing what proportion of those rearrangements had an F ratio greater than or equal to ours. This method makes fewer assumptions than are required for using the standard F distribution and also provides a straightforward way of factoring out the contribution of the second phoneme: Each time we randomly rearranged words across phoneme groups, we only swapped pairs of words that have the same second phoneme (i.e., words were not swapped across the horizontal lines in the table). In effect, the data were cross-classified in blocks based on the second phoneme, and in each rearrangement the number of words in each of those blocks did not change. This technique of blocking on variables to factor out the significance of their effect was used throughout our analyses. We also blocked simultaneously on the participant in each trial. Blocking allowed us to isolate the different sources of variance, yet still run all the data in one large test of significance.

The first row of Table 3, "Raw RT," shows the results of this analysis. Words with different phonemes have different measured response times, at a significance of .000, i.e., $p < .001$. We ran the same analysis for the three other megastudies and obtained the same result. The only difference in those tests is that trial-level data were not available, and so we used as our basic level of analysis the average response time for each word. These results suggest that there is a voice key bias: Words with different initial phonemes have significantly different response time measures.

One might suspect, however, that the relationship between initial phoneme and response time is not direct, but is mediated through some third covariable such as spelling length. We therefore tagged each of our words to identify the levels

TABLE 2

Illustration of Monte Carlo Rearrangements

Blocking Variables		Original				One rearrangement			
Vowel	Speaker	/b/		/d/		/b/		/d/	
/æ/	1	*bad*	660	*dab*	748	*dab*	748	*bad*	660
		badge	806	*dad*	738	*badge*	806	*dad*	738
				dam	797			*dam*	797
/æ/	2	*bad*	840	*dab*	550	*dam*	490	*bad*	840
		badge	629	*dad*	521	*dab*	550	*badge*	629
				dam	490			*dad*	521
/ɪ/	1	*bib*	815	*did*	826	*bib*	815	*bid*	857
		bid	857	*dig*	518	*did*	826	*big*	612
		big	612	*dill*	473	*dill*	473	*dig*	518
/ɪ/	2	*bib*	508	*did*	565	*bid*	522	*bib*	508
		bid	522	*dig*	569	*big*	522	*dig*	569
		big	522	*dill*	737	*did*	565	*dill*	737
Sum			6,771		7,532		6,317		7,986
Metric[a]			9,312,229				9,305,132		

Note. Each rearrangement randomly shifts words and their response times across columns, subject to the following: (1) words are not shifted across blocks (demarcated by horizontal lines); (2) each column of each block keeps the same number of words.

[a] Sum of the mean squares of the column sums; this is the portion of the *F* ratio that varies between rearrangements. The fraction of rearrangements for which the metric is greater than or equal to that of the original is *p*.

of potential covariables. The mean values for these covariables are listed in Table 1. Table 3 shows what happened when we tested whether words with different initial phonemes differ significantly for the various covariables. The statistical techniques used for these tests were exactly like those used for response time, except that these tests were conducted for each word, instead of for each trial. In almost all cases, words with different initial phonemes have significantly different levels of each of the covariables. These tests leave open the possibility that there are in fact no phonetic voice key biases: The different phonemes could have different response time measurements because words that have different phonemes also happen to vary with respect to the true causal factors.

However, an additional round of statistical blocking can factor out the effect of individual covariables. For each covariable, we tested whether there was still a significant ($p < .05$) difference between the response times for words with different initial phonemes after we blocked on different levels of the covariable. That is, the setup was similar to that illustrated in Table 2, except that the second blocking variable was the covariable instead of the speaker, and each word was measured only once. In almost all cases, the association remained significant. For example, the footnotes (*a*) in Table 3 show that when we block on that covariable, words with different phonemes still have significantly different response times. (It should be noted that in each cell of the table, the number and the footnote are referring to two different tests. An entry like "Familiarity .090[a]" means that when words are grouped by initial phoneme, the groups do not differ significantly among themselves in familiarity, at $p = .090$; and they do differ significantly among themselves in raw response time, when blocked by familiarity.) These analyses with blocking show that the association between phoneme and response time is not ultimately due to spelling consistency, familiarity, frequency of spelling, number of homophones, word length as determined by pronunciation or spelling, or neighborhood size. Tests of bigram frequency were less readily interpretable because virtually every word has a unique bigram frequency, resulting in block sizes that were too

TABLE 3

Significance of Differences in Response Variables Caused by Initial Phonemes

Response	KTM	TMBR	SB	SW
Raw RT	.000	.000	.000	.000
Bigram frequency	.000[b]	.000[a]	.000[b]	.000[b]
Consistency of onset	.000[a]	.000[a]	.000[a]	.000[a]
Consistency of rime	.000[a]	.007[a]	.001[a]	.000[a]
Familiarity	.090[a]	.043[a]	.017[a]	.020[a]
Frequency of spelling	.000[a]	.000[a]	.000[a]	.000[a]
Homophones	.000[a]	.000[a]	.000[a]	.000[a]
Length of pronunciation	.000[a]	1.000[a]	.000[a]	.000[a]
Length of spelling	.000[a]	.000[a]	.000[a]	.000[a]
Neighborhood size	.023[a]	.000[a]	.000[a]	.000[a]
Error rate	.000[a]	.012[a]	—	.426[a]
Residual RT	.000	.000	.000	.000

[a] Raw RT remains significant ($p \leq .05$) when blocked by this covariable.
[b] Raw RT not significant when blocked by this covariable.

small to allow significance to be determined in any event.

Although those blocking tests show that individual covariables are not responsible for the association between initial phoneme and measured response time, one might suspect that the association is mediated by a combination of covariables. To test that possibility, our final statistical test on initial phoneme groups was a regression analysis, using response time as the response variable and all the covariables, as well as error rate, as predictor variables. For each covariable, we used the transformation that accounted for the most variance in the response time as averaged across all subjects, as computed by a single-factor linear regression. The predictor variables in the multiple regression accounted for 35% of the variance in the response times in the KTM study, 27% in TMBR, 28% in SB, and 23% in SW. For each word, we determined the average residual response time left after subtracting the predicted values accounted for by the linear regression. Then we tested whether there were significant differences in residual response time for words with different initial phonemes.

The results of this analysis are presented in the last row of Table 3. For all four studies, words with different initial phonemes differed significantly ($p < .001$) in residual response time. That is, even after factoring out the effect of all the psycholinguistic covariables by linear regression, there was still a significant difference in response times between words with different initial phonemes. This difference must be due to factors other than the psycholinguistic factors, namely, phonetic voice key measurement biases.

The converging evidence over four different studies leaves little doubt that the initial phoneme itself has a direct influence on the response time. This corroborates previous suggestions from analyses of the SB and TMBR data that were reported in Spieler and Balota (1997) and Treiman et al. (1995). However, the current analysis is more direct and convincing. Treiman et al., for example, included the phonetic features of the initial phoneme in a large regression test to predict the average response time of words from many predictors and found that some of those phonetic features were significant. However, there is no guarantee that a multiple regression will assign variation to the right predictor variable: The variance assigned to the phonetic features could actually belong to psycholinguistic covariables that are highly correlated with those features. Our current regression analysis is more conclusive because we use it to factor out all effects that can be attributed to nonphonetic factors and then do association

tests only on the residuals. Also, we factor out any contributions of the second phoneme, and we reinforce the results by blocking tests. Although the latter are limited to single variables, they reliably remove 100% of the effect of the covariable in consideration without making the statistical assumptions required for linear regression.

Analyses comparing pairs of phonemes. Even after being convinced that there is some voice key bias by initial phoneme, one could still imagine that the effect is quite limited and therefore, perhaps, easily controlled. One possibility would be that only particular phonemes are of concern. For example, Bates, Devescovi, Pizzamiglio, D'Amico, and Hernandez (1995), responding to reports that fricatives and affricates cause problems in voice key studies, added a flag in their regression analyses to indicate whether the word in question began with a fricative or affricate. To examine whether there is any possibility of addressing the voice key question in such a straightforward way, we ran separate tests contrasting each pair of initial phonemes. This may at first alarm readers who are alert to the fallacy of attempting to prove that variables are associated by running separate tests on each of their levels. However, we have already demonstrated that the initial phoneme is associated with the response time. Our purpose now is to explore what factors may account for the overall association.

The procedure for testing particular pairs was essentially the same as for the residual response time analysis across all phonemes, except that the presence of exactly two groups at a time allowed for a simpler metric abstracted from the standard *t* test. Tables 4 through 7 present the results. These tables are arranged by descending order of differential residual response times for the phoneme. These differentials show how much more slowly words with the indicated phoneme (in the first column) were named than words with different first phonemes but the same second phonemes. That is, for each phoneme, we computed first the average residual response time for words with that phoneme in second position (R_2). Similarly, for each two-phoneme sequence, we computed the average residual response time for words that begin with that sequence ($R_{1,2}$). Then, to obtain the differential residual response times for a phoneme, we averaged the difference $R_{1,2} - R_2$ for all the two-phoneme sequences that begin with that phoneme. This computation factors out any effect of the second phoneme. In each row in Table 4, we present first the initial phoneme in consideration, its residual response time offset, and the number of words that begin with that phoneme. Then we list the phonemes that have significantly faster response times. For example, the 512 words beginning with /s/ are measured as 38.5 ms slower than can be accounted for by the levels of the covariables in those words or by the effect of the following phoneme. Furthermore, /k/, /t/, /n/, and so forth have significantly smaller (faster) residual response time offsets. The word counts are important, because lack of significance between two phonemes can be due to the fact that there are only a few words that begin with those phonemes, diminishing the power of statistical tests.

The tables leave no doubt that significant differences in response time between words with different initial phonemes are pervasive. At our significance level of .05, we would expect each row to have only one or two significant cells, but almost all rows have many more. The few exceptions are vowels, about which firm conclusions cannot be drawn because very few monosyllabic words in English begin with vowels. They are included in the tables primarily in order to show their differential residual response times.

Analyses examining classes of phonemes. Tables 4–7 contain a wealth of information about differences between individual pairs of phonemes. Our next step is to ask whether any generalizations emerge from the data. We took several passes over the data, breaking them down by each of the three main classes of consonant articulatory features: voicing, place of articulation, and manner of articulation. Because significance testing has already been done, we apply only informal summary statistics in this phase of the analysis.

The patterning most often mentioned by respondents to our Internet poll (35%) was that

TABLE 4

Differences in Response Time by Initial Phoneme Pairs (KTM Study)

Phone	RT[a]	*N*																												
z	40.6	9				f		p	dʒ	d			h		j		b		m	ɹ		w	l				θ			
i	39.9	6															b													
ʃ	38.6	76	k	t	n	ð	f	ɡ	p	dʒ	d		aɪ	h		j	v	b		m	ɹ		w	l	ɔ			θ		ɑ	
s	38.5	512	k	t	n	ð	f	ɡ	p	dʒ	d	ɚ		h	ɪ	j	v	b	ɛ	m	ɹ	aʊ	w	l	ɔ	e	ʌ	θ	o	ɑ	ɔɪ
tʃ	37.1	55		t	n	ð	f	ɡ	p	dʒ	d			h		j	v	b		m	ɹ		w	l				θ			
k	21.8	239		t	n	ð	f	ɡ	p	dʒ	d			h		j	v	b		m	ɹ		w	l	ɔ			θ		ɑ	ɔɪ
t	12.3	163					f	ɡ	p	dʒ	d		aɪ	h		j	v	b		m	ɹ		w	l				θ		ɑ	
n	9.8	78												h		j	v	b		m	ɹ		w	l				θ			
ð	7.8	13																					w	l				θ			
f	4.6	189									d			h	ɪ	j		b		m	ɹ		w	l				θ		ɑ	
ɡ	3.9	150									d			h				b		m	ɹ		w	l						ɑ	
p	2.9	201									d			h	ɪ		v	b		m	ɹ		w	l				θ		ɑ	
æ	0.2	16													ɪ							aʊ							o		
dʒ	−3.2	47												h		j		b		m	ɹ		w	l				θ			
d	−6.0	138												h	ɪ	j		b		m	ɹ		w	l							
ɚ	−6.3	7												b																	
aɪ	−7.1	6																b													
h	−9.1	134																					w	l				θ			
ɪ	−9.9	11																b							ɔ				o		
j	−13.2	27																					w	l				θ			
v	−13.3	36																					w	l							
b	−14.5	247																	ɛ			aʊ	w	l							
ɛ	−15.6	11																													
m	−17.4	124																					w	l							
ɹ	−17.6	131																					w	l							
aʊ	−21.7	6																													
w	−22.6	127																													
l	−26.7	136																													
ɔ	−27.8	11																													
e	−28.0	12																													
ʌ	−29.6	3																													
θ	−33.6	40																												ɑ	
o	−35.1	8																													
ɑ	−37.9	11																													
ɔɪ	−50.2	1																													
u	−79.0	1																													

Note. Entries tell which phonemes have voice key response times significantly faster than the phoneme in the first column, $p \leq .05$. Effect of second phoneme is blocked. See International Phonetic Association (1996, 1999) for explanation of phoneme symbols.

[a] Response time residuals (milliseconds) after effect of covariables is removed by regression.

voicing would affect the measured response time. All of those who mentioned the direction of the effect said that voiced phonemes are detected faster than voiceless phonemes. A quick glance at Table 4 shows that this cannot be an absolute rule: There are many voiced phonemes that are detected more slowly than voiceless ones, and the slowest of all phonemes are the voiced /z/ and /i/. However, if we look at tendencies rather than absolutes, the voiceless phonemes do seem to cluster in the top (slower) half of the table. One way to quantify that impression is to compare the number of voiced phonemes that are faster than voiceless ones, and vice versa. We counted 97 times where voiceless phonemes are significantly slower than voiced phonemes, but only 12 times where voiced phonemes are significantly slower than voiceless ones; the voiceless have the advantage in 89% of all the untied pairs. Applying the

TABLE 5

Differences in Response Time by Initial Phoneme Pairs (TMBR Study)

Phone	RT[a]	*N*																	
tʃ	29.1	37	t		f	p	g	dʒ	n	h	ɹ	b	v	m	d	l	w	θ	j	ð
ʃ	27.7	46					g	dʒ	n	h	ɹ	b	v	m	d	l	w	θ	j	ð
s	27.4	83	t	k	f	p	g	dʒ	n	h	ɹ	b	v	m	d	l	w	θ	j	ð
t	18.1	68						dʒ			ɹ	b		m	d	l	w	θ	j	ð
k	17.1	75					g	dʒ	n		ɹ	b	v		d	l	w	θ	j	ð
f	15.2	62						dʒ			ɹ	b			d	l	w	θ	j	ð
z	14.1	6																	j	
p	9.4	87										b			d	l	w	θ	j	ð
g	8.2	51													d	l	w	θ	j	ð
dʒ	1.0	30															w	θ		ð
ɚ	0.0	1																		
n	−1.7	59													d	l	w	θ	j	ð
h	−3.2	80													d	l	w	θ	j	ð
ɹ	−3.7	97													d		w	θ	j	ð
b	−9.8	93															w	θ	j	ð
v	−13.8	23																		ð
m	−14.8	79																	j	ð
d	−20.2	71																θ	j	ð
l	−25.2	88															w		j	ð
w	−25.9	59																	j	ð
θ	−48.7	13																		
j	−56.7	17																		ð
ð	−56.7	9																		

Note. See notes for Table 4.

same counts to TMBR (Table 5), we get a ratio of 75:6 pairs where voiceless phonemes are slower than voiced phonemes (93%). Thus, the rule for voiced phonemes having faster response times holds for both studies carried out using the same equipment in the same lab at Wayne State University. The picture is somewhat different for the other two megastudies, however. In SB (Table 6) the ratio is only 63:40 (61%); in SW (Table 7) the ratio is 80:29 (73%). However, it could be the case that voicing is only correlated with some other feature that is the true cause of the response time difference. To get around such problems, we next compared phonemes that differ only by the voice feature, but are otherwise identical. Comparing response times of such pairs in KTM, we found four pairs such as /tʃ/ and /dʒ/, where the voiceless phoneme is significantly slower than its voiced counterpart, but only one, /θ/ and /ð/, where the voiced phoneme is slower (4:1). These ratios are 4:0 for TMBR, 2:4 for SB, and 1:1 for SW.

What can account for these patterns? The general tendency to detect voiced phonemes more quickly is doubtless due to the fact that voiced phonemes tend to be louder than voiceless ones (Fry, 1979). If a voice key is set at a fairly high threshold, then it may skip over the softer, voiceless segments to a greater degree than it skips over voiced ones, resulting in shorter measured response times for the latter. The differences between studies could be due to differences in thresholds. The more sensitive the apparatus (e.g., the lower the voice key threshold is set), the less there should be a bias due to the difference in acoustic intensity between voiced and voiceless sounds. A less obvious source of bias could arise if the voice key trigger is so high that it tends to miss even voiced obstruents, which are less loud than vowels. In such an event, the voice key will trip earlier for words beginning with voiced phonemes, not because it detects those consonants, but because voiced phonemes are about

TABLE 6

Differences in Response Time by Initial Phoneme Pairs (SB Study)

Phone	RT[a]	*N*																										
z	33.5	8	ð	s	g	k		θ	p			j	b	d	h	t	dʒ		l	n	ɹ	w			m		tʃ	ʃ
f	18.1	161		s	g	k	v	θ	p			j	b	d	h	t	dʒ		l	n	ɹ	w	e		m		tʃ	ʃ
ð	12.3	13						θ	p			j	b	d	h	t	dʒ		l	n	ɹ	w			m		tʃ	ʃ
s	10.9	415			g				p			j	b	d	h	t	dʒ	ɛ	l	n	ɹ	w	e	ɔ	m	æ	tʃ	ʃ
g	10.5	136				k						j	b	d	h	t	dʒ		l	n	ɹ	w	e		m		tʃ	ʃ
k	10.4	203							p	ɑ		j	b	d	h	t	dʒ	ɛ	l	n	ɹ	w	e	ɔ	m		tʃ	ʃ
v	10.3	30										j	b	d	h	t	dʒ		l	n	ɹ	w			m		tʃ	ʃ
θ	6.9	36										j	b	d	h	t			l	n	ɹ	w			m		tʃ	ʃ
p	6.9	176									i	j	b	d	h	t	dʒ		l	n	ɹ	w			m		tʃ	ʃ
ɑ	4.5	6																										
i	4.3	6																										
aʊ	3.0	6																										
ɪ	2.2	6											b															
u	0.8	1																										
aɪ	0.4	4											b															
j	0.2	24																	l	n	ɹ	w			m		tʃ	ʃ
b	−1.4	219													h			ɛ	l	n	ɹ	w			m	æ	tʃ	ʃ
d	−2.3	112																ɛ	l		ɹ	w			m		tʃ	ʃ
h	−2.4	115														t					ɹ	w			m		tʃ	ʃ
t	−2.5	131																ɛ	l	n	ɹ	w			m		tʃ	ʃ
ɚ	−3.0	6																										
dʒ	−3.0	39																	l		ɹ	w			m		tʃ	ʃ
ɛ	−3.2	8																										
l	−5.7	119																				w			m		tʃ	ʃ
n	−5.9	68																							m		tʃ	ʃ
ɹ	−6.0	125																				w			m		tʃ	ʃ
w	−7.8	112																							m		tʃ	ʃ
o	−8.2	4																										
e	−10.3	10																										
ɔ	−10.3	6																										
m	−11.6	101																										ʃ
æ	−12.3	9																										
tʃ	−13.5	45																										ʃ
ɔɪ	−14.2	1																										
ʃ	−22.8	64																										

Note. See notes for Table 4.

40 ms shorter than voiceless phonemes (Klatt, 1976).

In our poll, five people (8%) mentioned place of articulation as a dimension along which initial phonemes could differentially affect voice key response time. Of those five, only two respondents offered a specific hypothesis. One stated that labials have shorter response times than other consonants. The other stated more generally that anterior consonants are detected faster than posterior ones. The data conform to a slight variation of the anteriority hypothesis: velar > coronal (alveolar, postalveolar, or palatal) > labiodental > dental > labial (bilabial or labiodental), where ">" means that the symbol to the left has the greater measured response time. (A strict interpretation of the anteriority hypothesis would place labiodental sounds after dental sounds, because the former involve the lips.) In the KTM study, if we look at consonant pairs that differ only by place of articulation, 10 pairs like /k/ and /t/ agree with the hypothesis; none contradict it (10:0). The ratios for the other studies are TMBR 5:0, SB 7:2, and

TABLE 7

Differences in Response Time by Initial Phoneme Pairs (SW Study)

Phone	RT[a]	*N*																												
z	55.6	8	tʃ		dʒ	k	f		t		j		d		b	h		ð	θ	n	p	l	w		m		ɹ			
aɪ	38.2	4		g		k	f						d		b						p									
s	37.9	418	tʃ	g	dʒ	k	f	v	t	ɪ	j	ɚ	d	ɛ	aʊ	b	h	e	ɔɪ	ð	θ	n	p	l	w	ɔ	m	o	ɹ	æ	
ʃ	34.7	64		g	dʒ	k	f	v	t		j	ɑ	d			b	h			ð	θ	n	p	l	w	ɔ	m		ɹ		
tʃ	25.0	46					f	v	t		j		d			b	h			ð	θ	n	p	l	w		m		ɹ		
g	24.1	137						v	i		j		d			b	h			ð	θ	n	p	l	w		m		ɹ		
dʒ	18.4	39									j		d			b	h			ð	θ	n	p	l	w		m		ɹ		
k	17.2	205							i		j		d			b	h			ð	θ	n	p	l	w		m		ɹ		
u	15.8	1																													
f	14.3	163							i				d			b	h			ð	θ	n	p	l	w		m		ɹ		
v	11.8	30																		ð	θ	n	p	l	w		m		ɹ		
i	10.3	6														b							p								
t	8.6	133											d			b	h			ð	θ	n	p	l	w	ɔ	m		ɹ		
ɪ	4.4	6														b							p								
j	3.9	24																				n	p	l	w		m		ɹ		
ɚ	0.6	6																					p								
ɑ	−0.9	6														b															
d	−1.9	111												ɛ	aʊ	b	h			ð		n	p	l	w		m		ɹ		
ɛ	−4.7	8														b							p								
aʊ	−4.7	6														b					θ		p								
b	−7.2	219																				n	p	l	w		m		ɹ		
h	−7.6	115																				n	p	l	w		m		ɹ		
e	−8.3	11																					p								æ
ɔɪ	−12.4	1																													
ð	−14.0	13																													
θ	−17.0	36																									m				
n	−17.7	68																													
p	−20.5	175																											ɹ		
l	−20.9	119																													
w	−20.9	112																													
ɔ	−21.0	6																													
m	−24.0	101																													
o	−25.4	4																													
ɹ	−28.8	125																													
æ	−40.0	10																													

Note. See notes for Table 4.

SW 12:0. Thus there are only two contradictions to the hypothesis in all four megastudies. This consistency suggests a real and pervasive bias that is not very sensitive to differences in experimental setups. For the plosives, this patterning correlates with the length of burst and aspiration noise: Velar plosives have the greatest burst and aspiration components; labial plosives have the least (Klatt, 1975). If voice keys tend to skip either or both of these components, that could explain why they tend to register velars most slowly and labials most quickly. For the fricatives, the relative ordering corresponds to the average duration of the phoneme; Kent and Read (1992) suggested that /ʃ/ and /s/ are longer than /f/, which is longer than /θ/. One may hypothesize that some voice keys have a threshold setting so high that they often skip over initial fricatives entirely. This is corroborated by the findings of Sakuma et al. (1997) and Rastle and Davis (2002), who found that typical voice keys, on average, were tripped more than 100 ms after the true beginning of an /s/-initial word; that length of time corresponds to the entire duration

of the phoneme. A prominent exception in our megastudies is in SB, where /ʃ/ is detected faster than all other phonemes. That is explainable by the fact that /ʃ/ is the loudest of all obstruents, and that fricatives, as mentioned before, take the shortest time for speakers to initiate (Sakuma et al., 1997). The differences in the treatment of /ʃ/ between the studies could be due to the voice key being set to trip a little lower (or the microphone being a little more sensitive, or the participants sitting a little closer to the microphone or talking a little louder) in the SB study than in the other three megastudies.

Manner of articulation was suggested as a source of voice key bias by about half of the respondents in our poll, and it was also considered by some researchers in our analysis of published experiments. Twelve of the poll respondents (20%) suggested that plosives are detected faster and more reliably than other sounds. However, 4 people offered the opposite idea. A second hypothesis, offered by 16 respondents (27%), was that fricatives are detected later, and less reliably, than other sounds. The strictest way to investigate whether manner is a causal factor is to contrast pairs of phonemes that differ only in manner of articulation. In all four megastudies, /s/ is detected more slowly than /t/, and in three studies (KTM, SB, SW) we see a significant effect for /z/ being slower than /d/. That is, fricatives are detected more slowly than plosives, at least for the coronal place of articulation. These results are fairly meager because there are relatively few minimally contrasting pairs of phonemes one can check in English. A broader pattern can be discerned if we relax our criterion a bit and simply ask whether phonemes with different manners of articulation, as groups, differ in response times. For example, for affricates, we counted all pairs where an affricate phoneme was significantly slower than a nonaffricate phoneme, and divided that by the number of all pairs where an affricate is significantly slower or faster than a nonaffricate phoneme. Table 8 summarizes that computation across seven manners of articulation, and for all four studies. A fairly strong pattern here is that obstruents (plosives, affricates, and fricatives) are detected more slowly than sonorants (nasals, approximants, and vowels). The only exception is that in SB, affricates are detected quickly, like sonorants. These ratios corroborate the hypothesis that fricatives are detected very slowly, but contradict the idea that plosives are detected especially quickly.

TABLE 8

Effect of Manner of Articulation of Initial Phoneme on Residual Response Time

Manner	KTM	TMBR	SB	SW
Affricate	.82	.79	.19	.84
Fricative	.73	.56	.82	.78
Plosive	.65	.64	.67	.56
Nasal	.40	.50	.09	.00
Vowel	.13	—	.12	.32
Glide	.07	.03	.26	.20
Liquid	.03	.33	.22	.00

Note. Number of phoneme pairs where the phoneme with the indicated feature is significantly slower than the phoneme lacking the feature, divided by the number of phoneme pairs where exactly one of the phonemes has the indicated feature and there is a significant difference between their response times.

Physical causes for these patterns have, to a large extent, already been mentioned. Some types of sounds begin with a period of silence corresponding to complete air blockage in the vocal tract (plosives, affricates, and often vowels), which naturally slows their detection relative to other sounds; at the other extreme, fricatives can be produced exceptionally quickly. Another consideration is that sound types differ in their loudness. Sonorants (especially vowels), are the loudest of all sounds; obstruents are less loud and so may be skipped over entirely.

The inconsistent status of affricates can be attributed to the intermediate status of postalveolar consonants in general (/ʃ/ as well as /tʃ/ and /dʒ/), which tend to be the loudest of the obstruents (Fry, 1979). It would appear that in SB alone the voice key was sensitive enough to detect them readily, and in the case of /ʃ/ this combined with the articulatory advantage to make it the most quickly detected of all phonemes. In the other studies, the apparatus was apparently not sensitive enough to detect even postalveolar obstruents, and so words beginning with them were detected slowly.

Summary of results for initial phonemes. The results clearly show a response time measurement bias. The statistical tests showing that response time varies across phonemes even when covariables are factored out (Table 3) are buttressed by finer-grained evidence at the level of phoneme pairs (Table 4–7). The differences between phonemes align along the well-known dimensions of voicing, place, and manner of articulation; it is not the case that every phoneme behaves in an idiosyncratic fashion. The patterns are explicable in terms of our prior knowledge of the articulation and acoustics of the phonemes. However, the details can vary widely across different phonemes and different studies. Thus, while the fricatives /s/ and /ʃ/ are slow to detect in three of the megastudies, /ʃ/ but not /s/ is extremely quickly detected in one of the four studies (SB). It also appears that there may be substantial variation beyond that found in our four megastudies, in that we found no evidence of a response time advantage for plosives, yet that was mentioned by several poll respondents and reported by Connine et al. (1990). There is also the issue that several different factors influence response time measurements simultaneously. One cannot assume that by treating, say, fricatives, differently, one can readily correct the bias problem. Rather, all three dimensions of articulatory features have an effect, so that, in this example, fricatives will differ among themselves depending on their voicing and place of articulation.

Second Phonemes

For the most part, it has been assumed that effects of phonemes on voice key response times are limited to the initial phoneme of the word. As we have mentioned, less than a quarter of the experiments we surveyed included any effort to control for the second phoneme. However, some recent research shows that second-position phonemes may need to be taken into consideration. Rastle and Davis (2002) reported that typical voice keys trigger about 10 ms later on words that begin with /s/ plus obstruent than on words that begin with /s/ plus vowel. The analysis presented here asks whether this is an isolated problem or whether it is the tip of an iceberg.

Overall analyses. We performed the same analyses as with initial phonemes, but this time used the second phoneme of the word as the grouping variable and blocked by the first phoneme. That is, we asked whether words with different second phonemes had significantly different response time or covariables, after factoring out the effect of the first phoneme.

The results in Table 9 are similar to those for initial phonemes (Table 3), though a few details vary. A test of whether words with different word-second phonemes have different raw (unadjusted) response times is significant in all four studies at $p < .001$. The covariables, by and large, are less plausible as indirect causes than they were for initial variables. As the superscripts indicate, even those covariables that varied significantly with the second phoneme cannot individually be responsible for the differences in response time. When we conducted a multiple regression to factor out the contribution of all these covariables from the response time, we found that words with different second phonemes differed in their residual response times. This difference just reaches significance in SW, but is highly significant ($p < .001$) in the other three studies. Overall, we have strong evidence of a voice key bias caused by the second phoneme of the word.

Analyses comparing pairs of phonemes and their features. Tables 10 through 13 show figures for the individual phonemes and pairs of phonemes. Recall that the response time measurements are the average amount by which the residual response time exceeds the average residual response time of other second-position phonemes when they follow the same initial consonant. We ran these tests only for words that begin with a consonant. We omitted vowel-initial words from these analyses because relatively few monosyllabic words begin with vowels, leaving very little data with which to conduct significance tests. On phonetic principles, we would expect voice keys to be triggered near the beginning of vowel-initial words in any case, and therefore to be little affected by following phonemes.

Rastle and Davis (2002) found that, with a typical voice key, words that begin with /s/ fol-

TABLE 9

Significance of Differences in Response Variables Caused by Second Phoneme of Word

Response	KTM	TMBR	SB	SW
Raw RT	.000	.000	.000	.000
Bigram frequency	.029[b]	.001[b]	.115[a]	.105[b]
Consistency of onset	.000[a]	.800[a]	.000[a]	.000[a]
Consistency of rime	.000[a]	.000[a]	.000[a]	.000[a]
Familiarity	.413[a]	.714[a]	.652[a]	.551[a]
Frequency of spelling	.236[b]	.184[b]	.709[a]	.701[a]
Homophones	.000[a]	.000[a]	.000[a]	.000[a]
Length of pronunciation	.000[a]	1.000[a]	.000[a]	.000[a]
Length of spelling	.000[a]	.000[a]	.000[a]	.000[a]
Neighborhood size	.000[a]	.000[a]	.000[a]	.000[a]
Error rate	.000[a]	.003[a]	—	.193[a]
Residual RT	.000	.000	.000	.050

[a] Raw RT remains significant ($p \leq .05$) when blocked by this covariable.
[b] Raw RT not significant when blocked by this covariable.

lowed by /t/ or /p/ were detected slower than words that begin with /s/ followed by a vowel. Can their finding be broadened to other consonant clusters? A quick look at Tables 10 through 13 reveals that vowels are interleaved with consonants. Consonants are not clustered at the top of the tables, as they would be if consonant clusters were consistently measured as beginning later than single-consonant onsets. When we count, for example, the number of times consonants are significantly slower than vowels, no striking patterns emerge. Even the positions of particular consonants are rarely consistent across the studies. The semivowel /j/ makes for very slow response time measurements in all studies, but its spelling after an initial consonant is so unusual in English (typical spellings of /ju/ encode the consonant and vowel simultaneously) that we must reserve the possibility that cognitive factors are at play. The other approximants appear to speed up response time measurements in comparison to vowels; but the effect does not hold for SB. Of the remaining consonants, which appear only after /s/, only /k/ appears to consistently make for slower measurements than for words with a vowel after the /s/.

What can account for these patterns? The obvious a priori expectation is that if words that begin with /s/ typically do not trigger voice keys at all, then words that begin with /s/ plus a following voiceless plosive (/p/, /t/, /k/) should fare even worse, because those consonants have even less amplitude than /s/ (Kawamoto & Kello, 1999; Rastle & Davis, 2002). However, there are compensating factors. Research has shown that both of the initial consonants in clusters are typically quite a bit shorter in American English than they are when followed by vowels (Klatt, 1974; Umeda, 1977). After both elements in a cluster shorten, the /s/-plosive cluster can be almost as short as a single /s/ before an average vowel. Umeda also noted a large difference in her measurements of the length of /s/ before /p/ and those of Schwartz (1970). It is conceivable that this variation has some bearing on the fact that /sp/ behaves differently in SW than in our other studies. Other factors that can affect the outcome of consonant cluster response time measurements include a partial devoicing of sonorant consonants after voiceless consonants, making the vowel much less loud; and a retraction of /t/ and /d/ to the postalveolar region before /ɹ/, making them louder. As for initial consonants, it is easy to see how minor differences in acoustic thresholds between studies and minor differences in speech amplitude and phoneme timings between subjects (both idiosyncratic and dialectal) could account for major differences in how particular consonants behave in different studies. In addition, evi-

TABLE 10

Differences in Response Time by Second Phoneme Pairs (KTM Study)

Phone	RT[a]	*N*																					
j	41.4	21	i	u	ɪ	aʊ		aɪ	e		o	ɑ	ɔɪ	æ	ɔ	ɛ		ɚ	ʌ	ɹ	ʊ	l		
i	16.9	142			ɪ	aʊ		aɪ	e		o	ɑ	ɔɪ	æ	ɔ	ɛ		ɚ	ʌ	ɹ	ʊ	l		
u	16.9	94						aɪ	e		o	ɑ		æ	ɔ	ɛ		ɚ	ʌ	ɹ		l		
p	14.1	70				aʊ		aɪ			o			æ	ɔ	ɛ	t					l	m	
f	12.7	2																						
ɪ	10.8	225									o	ɑ		æ	ɔ	ɛ		ɚ	ʌ	ɹ	ʊ	l		
aʊ	7.3	42					k			n					ɔ				ʌ	ɹ				
k	6.4	74						aɪ			o				ɔ	ɛ	t					l	m	
aɪ	4.0	127								n		ɑ			ɔ		t		ʌ	ɹ	ʊ	l		
e	3.6	160										ɑ							ʌ	ɹ		l		
n	1.0	25																						
o	0.1	132																		ɹ		l		
ɑ	0.1	140												æ						ɹ				
ɔɪ	−0.6	21																				l		
æ	−2.2	186																	ʌ			l		
ɔ	−2.4	149																						w
ɛ	−3.2	182																		ɹ				
t	−3.9	107																						
ɚ	−4.5	77																						
ʌ	−5.3	180																		ɹ				
ɹ	−6.8	336																						
ʊ	−7.2	32																						
l	−12.1	253																						
m	−14.8	18																						
w	−23.0	77																						

Note. Entries tell which phonemes have voice key response times significantly faster than the phoneme in the second column, $p \leq .05$. Effect of first phoneme is blocked.

[a] Response time residuals (milliseconds) after effect of covariables is removed by regression.

dence has recently been adduced that words with initial clusters are uttered sooner than words with no clusters, apparently for cognitive reasons (Kawamoto & Kello, 1999; Rastle & Davis, 2002).

Perhaps the most intriguing pattern to emerge from Tables 10–13 concerns the vowels. Little research using voice response latencies has mentioned the possibility that response time of words can be biased by the identity of an interior vowel. The published research on voice key biases has not looked at interior vowels. While a minority of our poll respondents were willing, in principle, to entertain the possibility that vowels could have an effect, they offered no concrete hypotheses about which vowel factors could be involved.

It is obvious that there are significant differences between the vowels. In SW, there are 7 different vowels that are associated with response times that are significantly different from some other vowel or vowels; in SB, there are 8; in KTM and TMBR, 9. The differences in residual response time between the fastest and slowest vowel range from 12 ms in SB to 67 ms in TMBR. Thus the systematic differences can be large, depending on factors that vary between experiments. One way to explore the differences between vowels is to break the vowels down by their major articulatory dimensions, as was done earlier for the initial consonants. One important dimension is front versus back. If we look only at vowels that differ only in frontness, in no case do we find back vowels with slower measured response times than front vowels, whereas all the studies have at least one pair where the front vowel is slower. Counting all pairs that differ in frontness and significantly

TABLE 11

Differences in Response Time by Second Phoneme Pairs (TMBR Study)

Phone	RT[a]	*N*													
æ	33.0	128	i	aɪ	ɛ	ɑ	e	ɪ	u	ʌ	ɚ	o	ɔ	ɔɪ	ʊ
aʊ	30.6	23				ɑ	e	ɪ	u	ʌ	ɚ	o	ɔ	ɔɪ	ʊ
i	20.0	110						ɪ	u	ʌ	ɚ	o	ɔ	ɔɪ	ʊ
aɪ	10.7	91								ʌ		o	ɔ	ɔɪ	
ɛ	5.3	95						ɪ		ʌ	ɚ	o	ɔ	ɔɪ	ʊ
ɑ	4.5	76									ɚ	o	ɔ	ɔɪ	ʊ
e	1.1	123						ɪ		ʌ	ɚ	o	ɔ	ɔɪ	ʊ
b	0.0	1													
ɪ	−6.5	140												ɔɪ	ʊ
u	−7.5	66													ʊ
ʌ	−13.0	117													
ɚ	−15.7	61													
o	−19.0	88													
ɔ	−19.6	75													
ɔɪ	−31.9	11													
ʊ	−34.4	29													

Note. See notes for Table 10.

differ in response time, we find in KTM 14 vowel pairs that show the front vowel as slower than the back vowel; only 4 times are back vowels slower (14:4, front vowel is slower in 78%). In TMBR, the ratio is 24:0 (100%); in SB, 12:6 (67%); in SW, 16:0 (100%). There is a clear effect whereby front vowels are detected more slowly, although the effect is a bit weaker in SB.

The other important dimension in vowels is their height. Among contrasting vowel pairs, in KTM we find 24 cases where a high vowel is associated with a slower response time than a nonhigh vowel and only one case where the reverse is true (24:1, high vowel is slower in 96%). In TMBR, the quantities are equal, 6:6 (50%). In SB, the numbers are 27:0 (100%); in SW, 16:1 (94%). If we look at the four pairs that clearly contrast minimally (/i/-/e/, /ɪ/-/ɛ/, /u/-/o/, /ʊ/-/ɔ/), SB has all four pairs with the high vowel significantly slower; KTM and SW have three of the four in that direction; TMBR, however, has none in that direction, but it does have the contrary /ɛ/ > /ɪ/. With the rather mysterious exception of TMBR, therefore, there is a strong effect whereby words with a high vowel after an initial consonant have measured response times slower than analogous words with a nonhigh vowel.

Thus we find pervasive evidence that there is a response time bias due to the second phoneme, such that front vowels are associated with greater response times than back vowels, and high vowels have greater response times than lower vowels. To the above-stated evidence we might mention that /i/ is slower than /ɪ/ (significantly so in three of the studies); while both are front and high, /i/ is even more front and high than /ɪ/. Recall also that the semivowel /j/ as second element slows the response time considerably. Although we cannot discount the possibility that its odd spelling is a factor, it is noteworthy that /j/ is the only consonant in English that is front and high and is in fact virtually identical in articulation to /i/.

Are there articulatory or acoustic facts that could account for the differences among the vowels? Articulatory biases seem unlikely in light of the preliminary report of Rastle and Davis (2000) that words with internal vowels differing in frontness did not have different response times when measured by waveform. The effect may be due in part to a general tendency to lengthen consonants before high vowels (Klatt, 1975; Schwartz, 1970). If the onset consonant is so low in amplitude that the voice key sometimes misses it altogether, or if the time it

TABLE 12

Differences in Response Time by Second Phoneme Pairs (SB Study)

Phone	RT[a]	*N*																						
j	17.6	18	i		ʊ	u	aɪ		ɪ	ɔ	e	l	ɹ				aʊ	ɛ	ɑ		o	æ	ɚ	ʌ
f	12.2	2																			o			
p	8.3	53		w				k	ɪ	ɔ		l		m	n			ɛ		t	o		ɚ	ʌ
i	7.5	123				u			ɪ	ɔ	e		ɹ				aʊ	ɛ	ɑ		o	æ	ɚ	ʌ
w	6.9	58						k		ɔ		l	ɹ								o			ʌ
ʊ	6.0	29								ɔ	e		ɹ				aʊ	ɛ	ɑ		o	æ	ɚ	ʌ
u	4.2	76																ɛ	ɑ		o	æ	ɚ	ʌ
aɪ	4.0	114																ɛ	ɑ		o	æ	ɚ	ʌ
k	2.5	56								ɔ										t	o			
ɪ	2.3	193															aʊ	ɛ	ɑ		o	æ	ɚ	ʌ
ɔ	1.2	130																	ɑ			æ		ʌ
e	1.2	137																ɛ	ɑ			æ		ʌ
l	1.1	199											ɹ				aʊ		ɑ		o	æ		ʌ
ɹ	0.9	285																	ɑ			æ		ʌ
m	0.8	13																			o			
n	−0.3	16																			o			
aʊ	−0.3	39																						
ɔɪ	−0.8	19																						
ɛ	−2.1	164																						
ɑ	−2.9	117																						
t	−2.9	96																						
o	−3.3	118																						ʌ
æ	−3.7	167																						
ɚ	−3.8	68																						
ʌ	−4.4	162																						

Note. See notes for Table 10.

takes the consonant to reach sufficient amplitude to trigger the voice key is proportional to its length, then the longer the consonant, the greater the measured response time. Another possible explanation involves the amplitude of the vowel itself. Front vowels have less amplitude than back vowels, and high vowels have less amplitude than low vowels (Fry, 1974). If the voice key misses the initial consonant entirely and has to contend with the vowel, it may take longer to detect weak vowels like /i/ than louder vowels like /ɑ/. It is also possible that through coarticulation, the initial consonants themselves may in part anticipate the articulatory gestures of the vowels that make for dampened amplitude.

Summary and implications of results for second phonemes. The phonemes that follow word-initial consonants have an effect on measured response time. General statistical tests across all phonemes (Table 9) indicated that words with different second phonemes had different response times, even when blocked by individual covariables. They also had different residual response times after all covariables were factored out. When we looked at the contrasts between individual pairs of phonemes occurring in word-second position (Tables 10–13), we found significant differences between many of the pairs. Those differences are aligned along articulatory dimensions that provide coherent explanations for most of the effects. The fact that our data make sense in terms of prior knowledge of articulatory and acoustic phonetics makes the statistics especially convincing.

Could the second-phoneme effects we have uncovered here have an impact on current experimental paradigms? Consider a naming study reported by Peereman, Content, and Bonin (1998; Experiment 1b). They asked whether printed words containing sounds that can be spelled in many different ways (feedback incon-

TABLE 13

Differences in Response Time by Second Phoneme Pairs (SW Study)

Phone	RT[a]	*N*																			
j	27.5	17	i	ɪ	aʊ	e	ɑ	u		o	ɛ	æ	ʌ	ɔ	l	ɚ	ɔɪ	ɹ		w	
f	24.8	2																		w	
i	14.8	123				e	ɑ	u		o	ɛ	æ	ʌ	ɔ	l	ɚ	ɔɪ	ɹ		w	
k	14.2	56	aɪ			e	ɑ		n		ɛ	æ		ɔ	l				p	w	m
aɪ	9.6	114			t		ɑ	u	n	o	ɛ	æ	ʌ	ɔ	l	ɚ		ɹ	p		
ɪ	9.5	194								o	ɛ	æ	ʌ	ɔ		ɚ	ɔɪ	ɹ			
ʊ	8.8	29											ʌ	ɔ	l			ɹ		w	
t	7.3	96					ɑ				ɛ	æ		ɔ					p	w	m
aʊ	5.8	39														ɔ		ɹ		w	
e	5.4	138									ɛ		ʌ	ɔ			ɔɪ	ɹ			
ɑ	4.6	116																ɹ			
u	2.0	76																ɹ			
n	−0.4	16																		w	
o	−0.5	121																ɹ			
ɛ	−0.6	163																ɹ			
æ	−2.7	168														ɔ		ɹ			
ʌ	−3.8	162																ɹ			
ɔ	−3.9	131																			
l	−4.3	199																ɹ			
ɚ	−4.8	69																			
ɔɪ	−6.5	19																			
ɹ	−7.4	287																			
p	−8.7	55																		w	
w	−15.5	58																			
m	−19.0	13																			

Note. See notes for Table 10.

sistent words) cause more difficulty in an oral reading task than words that contain sounds that are usually or always spelled the same way (feedback consistent words). The study was carried out in French. By way of example, the word *gland* /glɑ̃/ 'acorn' was considered inconsistent because /ɑ̃/ has at least 13 different spellings in French, whereas *piste* /pist/ 'track' was considered consistent because the constituent sounds are rarely spelled otherwise than as in *piste* itself. If there are feedback effects, one would expect the inconsistent words to be read more slowly and with more errors. Peereman et al. found that errors were marginally greater for the inconsistent words, but that the 6-ms increase in response time was not significant. Their conclusion from this and other experiments was that feedback consistency cannot be shown to have a reliable effect on performance. However, many more of the consistent words than the inconsistent words contained high front vowels and glides in word-second position (39% versus 11%). If high and front vowels and glides slow measured response times in French, as in English, then that imbalance would counter any feedback consistency advantage. Perhaps if the stimuli had been equated for second phoneme, the difference between the two conditions would have been of greater magnitude and significance, supporting the conclusion that feedback inconsistency does play a role in word naming. We mention this experiment because the authors were extremely careful to match their stimuli across many variables, and yet there is an imbalance in word-second phonemes that could lead to a bias against the experimental hypothesis. This is a new issue that future work in this paradigm will need to correct for.

Possible Solutions and Recommendations

We have documented large phonetic biases that come into play when voice keys are used to

measure vocal response times. These biases align along several phonological dimensions, extend at least two phonemes deep into the word, and are very sensitive to differences in experimental procedures. How can these problems be dealt with? There are several alternatives. We will discuss the solutions roughly in increasing order of usefulness, but all have their pros and cons.

First let us briefly consider the possibility of getting around the issue by not using voice response times at all. For example, lexical decision tasks measure how fast it takes a participant to decide whether a stimulus is a word and press the appropriate key. However, the lexical decision task is often viewed as a less "clean" task than naming, because in addition to a lexical access component, it includes a decision component, which may add bias to response time measurements (Balota & Chumbley, 1984; Morrison & Ellis, 1995). Lexical decision performance also varies with such things as the nature of the nonword foils (e.g., Lewellen, Goldinger, Pisoni, & Greene, 1993). Another way to avoid or supplement voice response times is to focus on error rates or error types (e.g., Andrews & Scarratt, 1998; Laxon, Masterson, & Moran, 1994). However, when studying adults' reading of familiar words, it may be difficult to elicit enough errors to make for statistically significant differences, unless special measures are taken (Kello & Plaut, 2000). Because alternatives to voice response time measurement have limitations of their own, many researchers choose to use a combination of these approaches and look for converging results (e.g., Connine et al., 1990; Wurm & Ross, 2001).

Another lateral solution is to adopt a protocol whereby voice response times are compared only against different utterances of the same word. In many types of research, the crucial difference is not in the stimuli themselves, but in some other experimental factor such as the prime that precedes the stimulus. If one compares the response times to multiple instances of exactly the same word, there is less room for phonetic bias to enter. Even here, however, one may need to beware of possibilities such as a prime's causing a response to be spoken more or less loudly, or more or less fast. That could affect the amplitude and duration of initial consonants that may fall below detection thresholds for some voice keys.

If one cannot get around the need to determine vocal response times for different word types, one optimistic idea would be to eliminate uncertainty by strictly controlling the setup in voice key experiments. Lowering the voice key trigger until one cannot tolerate the number of false starts caused by nonspeech noise would clearly reduce much phonetic variance. So would encouraging participants to speak at a constant loud amplitude, at a fixed distance from the microphone, and to make absolutely no nonverbal noise. It would also be advantageous to encourage participants to pronounce their whole response as quickly as possible, thus minimizing the duration of utterance-initial sound components that fall below the voice key cutoffs. It is difficult to infer on the basis of four studies just how far one can go with such measures. We might note, however, that the smallest range in phonetic bias between phonemes was 41 ms, in SB (for initial phonemes, ignoring uncommon ones such as /z/), which is still a good chunk of time. Also, there is a limit to how far one can go with forcing all participants to behave ideally. Even under the best of conditions, with very loud speakers and very low voice key thresholds, some phonemes will reach that target amplitude faster than others.

Another idea, attractive in its simplicity, would be to compensate numerically for the bias introduced by voice keys. After measuring the response time to a word that begins with /s/, the experimenter would look in Table 4, note that words in /s/ are detected 39 ms later than the average word, and so subtract 39 ms from the measured response time. This is one of the techniques considered by Kondo and Wydell (1997), using the Japanese data of Sakuma et al. (1997). Of course one would offset that number by a value corresponding to the effect of the second phoneme as well. Such an approach might be better than doing nothing, but it is not good enough. The numbers that obtain in one study do not necessarily apply to a different study; the differences range from the trivial to the alarm-

ing. Numerical compensation could conceivably work if the computation for the correction included factors for voice key threshold, amplitude of the beginning of the utterance, and so forth. However, as soon as one goes to the effort of measuring amplitude, one might as well compute the entire waveform, which obviates the need for a voice key in the first place.

The delayed naming paradigm is a popular way of addressing phonetic bias. This paradigm was introduced by Eriksen, Pollack, and Montague (1970). In addition to asking participants to read a word as quickly as possible immediately after it was presented (immediate, or standard, condition), in later sessions participants pronounced the same word after a variable delay of from 1 to 2 s (delayed naming condition). Presumably, 1 s should be enough time for even the most difficult of words to be read. Therefore, in the delayed condition, no response time differences should be attributable to lexical retrieval processes in the broad sense. Those differences should show up only in the immediate condition. In contrast, acoustic voice key biases should contribute equally to the measured response time under both conditions. Therefore the response times in the delayed condition can be treated statistically as equivalent to the noisy component of the response times from the immediate condition. Eriksen et al. found that syllable count (their variable of interest) had a significant effect on measured response time and that there was a significant interaction such that syllable count was more important under the immediate naming condition. Under the assumptions just outlined, that analysis convincingly indicated that syllable count affects naming time in a way that cannot be explained by phonetic biases alone. Another statistical approach, which does not require factorial analysis and which may be just as convincing, is to subtract the delayed response time for each participant from the immediate response time for the same participant and perform all computations on the difference (e.g., Fushimi et al., 1999; Lange & Content, 1999). Delayed naming is better than the idea of subtracting standardized offsets that was mentioned in the previous paragraph, because the compensating numbers are taken from the experiment itself and so will not be polluted by possible systematic differences between voice keys or between participants.

There are, unfortunately, several things that can go wrong with the delayed naming procedure. Researchers sometimes simplify the procedure to the extent that it is no longer statistically convincing. Often they simply show that the variable of interest significantly affects response time in the immediate but not in the delayed condition, but neglect to show that the difference between the two conditions is significant. Also, many researchers may not be aware of the great variance in voice key readings even for the same speaker producing the same utterance (Pechmann et al., 1989). To our knowledge, all researchers take only a single delayed naming trial for each word instead of averaging over several trials. In some instances the delayed naming data are taken from different participants from those who produced the immediate naming data (e.g., Forster & Chambers, 1973).

A much more important problem with the delayed naming corrective is that its core assumption may be wrong. The technique assumes that all word-specific biases that show up in the delayed condition are the same as those in the immediate condition. If that provision is not fulfilled, then, while the technique is removing one source of bias with the right hand, it is introducing a new source of bias with the left hand. Possible sources of such a problem are easy to imagine. For example, the delayed naming data may be gathered under conditions that inspire the participants to speak more loudly or softly than for the immediate naming trials. Participants who must hold a word for a second or more before uttering it may well get their articulators in position to say the word before they are asked to utter it. If so, delayed naming would not erase articulatory biases, and may introduce a new source of problems, through hyperarticulation. A more insidious case is if the delayed naming task itself is biased with respect to some psycholinguistic variable, perhaps the very one that is being studied. Goldinger, Azuma, Abramson, and Jain (1997) showed that low word frequency inhibits delayed naming. Using

frequency-biased measurements as a correction to one's primary data would introduce those biases into the data. Admittedly, the frequency effect in delayed naming is weak and may not, in practice, do much harm. However, until we are sure that the delayed naming task does not introduce any measurement bias of its own, the use of the delayed naming procedure in a test of statistical significance may be questioned.

Another solution available for many experiments, especially those with a small number of trials or a large amount of funding, is to record spoken responses and analyze them using modern digital processing techniques such as waveform inspection. There is no question but that such techniques permit a much more accurate determination of response time. The main drawback of the voice key is that the experimenter, in order to eliminate false readings caused by nonspeech noise, must set the key to ignore a certain level of sound. Transient noise is not as big a problem when using a digital recording: The experimenter can generally differentiate speech from noise by visual inspection of the waveform or by playing back parts of the recording that are in doubt. The use of visual waveforms greatly reduces the variance and therefore greatly increases the power of studies. Some researchers have abandoned voice keys, especially in recent years (e.g., Fushimi et al., 1999; Kawamoto & Kello, 1999), although a large majority of researchers still uses voice keys. A dividend of waveforms and digital signal processing is that these techniques permit one to study variables other than time to initiate response, such as the duration of the response and of its individual parts.

Despite their advantages, digital techniques do not solve all problems. For one thing, differentiating noise from speech is not always as easy as it would seem in theory. The beginnings of utterances tend to be very soft: If there is any background noise at all, it can be challenging to pinpoint exactly where speech begins. Also, digital techniques do not address articulatory biases at all. Fricatives are produced faster than sonorants, for example (Sakuma et al., 1997). Plosives and affricates (typically including even voiced ones in English) begin with periods of silence, of varying durations, as the vocal tract builds up pressure for a noisy release. There is no way that digital techniques can detect the beginning of a silent articulatory gesture, unless of course the default state of the speaker is to be noisy, as in the postvocalic naming technique of Kawamoto et al. (1998), where participants drone a schwa sound until they utter a plosive-initial response; but such a technique could not be used uniformly with all phonemes. For Japanese, at least, Sakuma et al. report an articulatory bias range of about 40 ms, comparable to the voice key phonetic bias in SB. From a qualitative standpoint, we are simply trading one set of biases for another: It is not much of an improvement to have fricatives go from being detected later than other phonemes to being detected earlier than other phonemes. Quantitatively, it must be admitted that digital techniques are superior: Voice key biases are a combination of the articulatory biases that also affect digital techniques, plus another set of acoustic biases unique to voice keys. Thus the variance is greater, and the magnitude of the bias is potentially (and, in most studies, actually) greater. However, it should not be assumed that digital techniques would allow the experimenter to ignore the effect that different phonemes may have on measured response time.

Even if one uses the most accurate digital signal processing techniques, measurement biases will still exist, for which one must apply strict statistical controls. The most reliable of such controls is to block the data by the leading phonemes. For example, in an experiment on the effect of spelling-to-sound consistency on naming latency, one would want to measure differences only between words that have the same beginnings. This is already best practice, though our survey of the literature showed that only a slim majority of the studies (56%) matched as much as the initial phoneme, even in the absence of controls like the delayed naming paradigm. We have confirmed here the wisdom of controlling for the initial phoneme by showing how pervasive the acoustic differences are, how great their magnitude can be, and how unpredictable they can be across experiments. In addition, we have demonstrated that the second

phoneme can have a significant effect. In the special case when the first two phonemes are both obstruents (/s/ followed by a plosive), this echoes and corroborates the recent finding of Rastle and Davis (2002). However, our findings greatly broaden our awareness of the scope of the problem. All sorts of second-position consonants can cause differential voice key effects, and, surprisingly, even the identity of a second-position vowel makes a difference. Therefore it is necessary to equate words at least through their first vowel. This is currently done in only a small minority of studies. It is not sufficient to match by broad class of phonemes, because we have seen that virtually all the main features of phonemes (manner, place, and voicing of consonants; height and frontness of vowels) have an effect; balancing words depending on whether or not they were fricatives, for example, would still leave open the possibility of biases caused by voicing or place of articulation. Besides, some of the details vary from study to study in unpredictable ways, so that phonemes that behave the same in one study may behave completely differently in another.

Our exhortation to "match" the entire head of stimuli has been intentionally vague, because there are many ways to do so. In a small study where the stimuli are carefully selected for balance, one will want to select words in advance so that sets vary on a property of interest, but have the same beginnings, and match on other confounding variables as well. In an exhaustive megastudy, the matching of items would occur after data gathering, in the analysis phase. One approach then would be to compute significance of the variable of interest by Monte Carlo tests, blocking by the head of the stimuli along with other covariables. Another approach, already employed by some researchers for the initial phoneme (e.g., Connine et al., 1990; Spieler & Balota, 1997; Treiman et al., 1995), is to analyze the data by regression analyses, first factoring out variance that can be accounted for by the head of the stimuli.

We do not wish to seem blithe about our recommendations. Frankly, blocking stimuli so that entire heads match can be difficult, especially given the need to match stimuli on other variables as well. Some 20 years ago, Cutler (1981) lamented that it was extremely difficult to design experiments with well-matched stimuli; the addition of a new variable on which to equate them makes the task even harder. The onset and vowel carry a disproportionate amount of the difference between words (Kessler & Treiman, 1997); there are often very few words that share both. Finding suitable words can be frustrating at best or impossible at worst when one must also manipulate the psycholinguistic variable being studied, not to mention controlling for a host of potential covariables. Nevertheless, our study of megastudies has shown that voice key bias is so large, pervasive, systematic, and variable that it cannot be ignored. The good news is that there are several correctives available to cautious researchers.

REFERENCES

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? *Journal of Experimental Psychology: Human Perception and Performance*, **24,** 1052–1086.

Bachoud-Lévi, A.-C., Dupoux, E., Cohen, L., & Mehler, J. (1998). Where is the length effect? A cross-linguistic study of speech production. *Journal of Memory and Language*, **39,** 331–346.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10,** 340–357.

Bates, E., Devescovi, A., Pizzamiglio, L., D'Amico, S., & Hernandez, A. (1995). Gender and lexical access in Italian. *Perception & Psychophysics*, **57,** 847–862.

Cedrus Corporation (2000). SuperLab Pro Input Options. [On-line]. Available: http://www.superlab.com/pro/input-options.htm.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16,** 1084–1096.

Cutler, A. (1981). Making up materials is a confounded nuisance; or, Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, **10,** 65–70.

Eriksen, C. W, Pollack, M. D., & Montague, W. E. (1970). Implicit speech: Mechanism in perceptual encoding. *Journal of Experimental Psychology*, **84,** 502–507.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, **12,** 627–635.

Fry, D. B. (1979). *The physics of speech*. Cambridge: Cambridge University Press.

Fushimi, T., Ijuin, M., Patterson, K., & Tatsumi, I. F. (1999). Consistency, frequency, and lexicality effects in naming Japanese kanji. *Journal of Experimental Psychology: Human Perception and Performance*, **25,** 382–407.

Goldinger, S. D., Azuma, T., Abramson, M., & Jain, P. (1997). Open wide and say "Blah!" Attentional dynamics of delayed naming. *Journal of Memory and Language*, **137,** 190–216.

Good, P. (1995). *Permutation tests*, 2nd corrected printing. New York: Springer.

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, **38,** 313–338.

Hutzler, F. (1999). Voice-key [On-line]. Available: http://www.sbg.ac.at/psy/people/hutzler/voicekey.htm.

International Phonetic Association (1996). *Reproduction of the international phonetic alphabet* [On-line]. Available: http//www2.arts.gla.ac.uk/IPA/ipachart.html.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, England, Cambridge University Press.

Kawamoto, A. H. & Kello, C. T. (1999). Effect of onset cluster complexity in speeded naming: A test of rule-based approaches. *Journal of Experimental Psychology: Human Perception and Performance*, **25,** 361–375.

Kawamoto, A. H., Kello, C. T., Jones, R., & Bame, K. (1998). Initial phoneme versus whole-word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **24,** 862–885.

Kello, C. T., & Plaut, David C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **26,** 719–750.

Kent, R. D., & Read, C. (1992). *The acoustic analysis of speech*. San Diego, CA: Singular.

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language*, **37,** 295–311. Available on-line: http://www.idealibrary.com/links/doi/10.1006/jmla.1997.2522.

Kessler, B., & Treiman, R. (2001). Relationships between sounds and letters in English monosyllables. *Journal of Memory and Language*, **44,** 592–617. Available on-line: http://www.idealibrary.com/doi/10.1006/jmla.2000.2745.

Klatt, D. H. (1974). The duration of [s] in English words. *Journal of Speech and Hearing Research*, **17,** 51–63.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, **18,** 686–706.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59,** 1208–1221.

Kondo, T., & Wydell, T. N. (1997, December). *Nature of naming latency for Japanese kanji words*. Paper presented at the International Conference on Cognitive Processing of Asian Languages & Symposium on Brain, Cognition, and Communication (ICCPAL97), Nagoya, Japan. Available on-line: http//www.brl.ntt.co.jp/people/kondo/wydell/index.html.

Lange, M., & Content, A. (1999, June). *The grapho-phonological system of written French: Statistical analysis and empirical validation*. Paper presented at the meeting of the Association for Computational Linguistics, University of Maryland. Available on-line: http//homepages.ulb.ac.be/~mlange/acl/LangeContentACL99Submitted.html.

Laxon, V., Masterson, J., & Moran, R. (1994). Are children's representations of words distributed? Effects of orthographic neighbourhood size, consistency and regularity of naming. *Language and Cognitive Processes*, **9,** 1–27.

Lewellen, M. J., Goldinger, S. D., Pisoni, D. B., & Greene, B. G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General*, **122,** 316–330.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, Cognition*, **21,** 116–133.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words*, Research on Speech Perception Progress Report 10. Bloomington, Indiana University.

Pechmann, T., Reetz, H., & Zerbst, D. (1989). Kritik einer Meßmethode: Zur Ungenauigkeit von Voice-key Messungen. [Critique of a method of measurement: On the unreliability of voice-key measurements.] *Sprache & Kognition*, **8,** 65–71.

Peereman, R., Content, A., & Bonin, P. (1998). Is perception a two-way street? The case of feedback consistency in visual word recognition. *Journal of Memory and Language*, **39,** 151–174.

Psychology Software Tools (2000). *PST serial response box* [On-line]. Available: http://www.pstnet.com/srbox/srb.htm.

Rastle, K., & Davis, M. H. (2000, November). *On the complexities of naming: What do voice keys measure?* Paper presented at the meeting of the Psychonomic Society, New Orleans.

Rastle, K., & Davis, M. H. (2002). On the complexities of measuring naming. *Journal of Experimental Child Psychology: Human Perception and Performance*, **28,** 307–314.

Sacia, C. F., & Beck, C. J. (1926). The power of fundamental speech sounds. *Bell System Technical Journal*, **5,** 393–403.

Sakuma, N., Fushimi, T., & Tatsumi, I. (1997). Onseiha no shisatsu ni yoru kana no ondokusenji no sokutei: Ondokusenji wa gotō on no chōonhō ni yori ōkiku kotonaru. [Naming latency measurements of kana based on inspection of voice waveforms: Naming latency varies greatly depending on the manner of articulation of the initial phoneme.] *Shinkeishinrigaku*, **13,** 126–136.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, **3,** 1–17.

Schneider, W. (1988). Micro Experimental Laboratory: An integrated system for IBM PC compatibles. *Behavior Research Methods, Instruments, and Computers*, **20,** 206–217.

Schwartz, M. F. (1970). Duration of /s/ in /s/-plosive blends. *Journal of the Acoustical Society of America*, **47,** 1143–1144.

Seidenberg, M. S., & Waters, G. S. (1989, November). *Reading words aloud: A mega study*. Paper presented at the meeting of the Psychonomic Society, Atlanta, GA.

Solso, R. L., & Juel, C. L. (1980). Positional frequency and versatility of bigrams for two- through nine-letter English words. *Behavior Research Methods & Instrumentation*, **12,** 297–343.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, **8,** 411–416.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, **112,** 1–36.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, **124,** 107–136.

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America*, **61,** 846–858.

Wurm, L. H., & Ross, S. E. (2001). Conditional root uniqueness points: Psychological validity and perceptual consequences. *Journal of Memory and Language*, **45,** 39–57.

Zeno, S. M., Ivenz, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.

(Received February 8, 2001)
(Revision Received May 18, 2001)