Psychon Bull Rev (2013) 20:87–94
DOI 10.3758/s13423-012-0336-5

BRIEF REPORT

The unbearable articulatory nature of naming: on the reliability of word naming responses at the item level

Arnaud Rey · Pierre Courrieu · Sylvain Madec · Jonathan Grainger

Published online: 7 November 2012
© Psychonomic Society, Inc. 2012

Abstract Single-word naming is one of the most widely used experimental paradigms for studying how we read words. Following the seminal study by Spieler and Balota (Psychological Science 8:411–416, 1997), accounting for variance in item-level naming databases has become a major challenge for computational models of word reading. Using a new large-scale database of naming responses, we first provided a precise estimate of the amount of reproducible variance that models should try to account for with such databases. Second, by using an item-level measure of delayed naming, we showed that it captures not only the variance usually explained by onset phonetic properties, but also an additional part of the variance related to output processes. Finally, by comparing the item means from this new database with the ones reported in a previous study, we found that the two sets of item response times were highly reliable ($r = .94$) when the variance related to onset phonetic properties and voice-key sensitivity was factored out. Overall, the present results provide new guidelines for testing computational models of word naming with item-level databases.

Keywords Single-word naming · Item-level performance · Large-scale databases · Accounting for variance · Testing computational models

The precision of theories and the grain size of empirical evidence do not always follow the same developmental trajectories in science, and this statement is certainly also valid for psycholinguistic research. Indeed, three decades ago, with the development of the first computer-implemented models (e.g., McClelland & Rumelhart, 1981), the precision of theoretical predictions has grown dramatically with the possibility for these artificial, dynamic systems to generate item-level predictions. At the same time, the grain size of empirical evidence was still bounded by a set of benchmark effects (e.g., the word frequency effect and consistency or regularity effects) modulated, in the best cases, by interaction effects (e.g., the Regularity × Frequency interaction; Taraban & McClelland, 1987).

To compensate for the disequilibrium between the grain size of the empirical data and the grain size of theoretical predictions, Spieler and Balota (1997) collected a large-scale database (31 participants, reading aloud a set of 2,870 monosyllabic English words) and tested the descriptive adequacy of various computational models of word reading against these item-level measures of performance, initiating what we may call a "nanopsycholinguistic" approach. The outcome was quite disappointing for modelers, since the best model only accounted for around 10 % of the item variance. More surprisingly, a linear combination of three simple factors (Log Frequency, Neighborhood Density, and Word Length) accounted for 27.1 % of the variance, and when the onset phoneme characteristics were taken into account, this percentage even reached 43.1 %, which was much higher than the predictions of any of the models tested in that study. Instead of re-equilibrating the precision ratio between the data and models, the balance was therefore clearly tipping in the other direction.

In defense of computational models, Seidenberg and Plaut (1998) suggested that two important issues should be considered when testing models with item-level databases. First, Seidenberg and Plaut noted that the precision of large-scale databases might be contaminated by an undesirable source of error variance, and that one would need to

A. Rey (✉) · P. Courrieu · S. Madec · J. Grainger
Laboratoire de Psychologie Cognitive,
CNRS and Aix-Marseille University,
3, place Victor Hugo, Case D,
13331 Marseille Cedex 3, France
e-mail: arnaud.rey@univ-amu.fr

estimate this amount of experimental noise that models cannot or should not try to account for. Second, they argued that models should not be expected to account for all of the item variance, because some sources of variance are clearly outside their explanatory scope. For example, onset phoneme characteristics are responsible for a large part of the item variance in naming (about 30 % in Spieler & Balota's, 1997, study), and because these output articulatory factors are not implemented in current models, the models obviously cannot account for these sources of variance.

To address the first of these two critical issues, we recently proposed a method for estimating the amount of *reproducible variance* that is contained in any item-level database (Courrieu, Brand-d'Abrescia, Peereman, Spieler, & Rey, 2011; Courrieu & Rey, 2011; Rey & Courrieu, 2010; Rey, Courrieu, Schmidt-Weigand, & Jacobs, 2009). By reproducible variance, we mean the variance that remains when the error variance is subtracted from the total amount of item-related variance (i.e., the variance of the item means). This source of stable or reproducible variance corresponds precisely to the amount of variance that models should try to account for.

One general way to tackle this issue is to consider that the amount of variance that models should try to account for corresponds to the amount of variance shared by two independent groups of participants. By definition, this shared or *reproducible* variance is immune to the inherent experimental noise that is present in any specific database. Correlating item performance averaged over participants from two independent groups therefore provides an index of this amount of reproducible variance. More formally, we have shown that the expected value of such correlations has the form of an intraclass correlation coefficient (ICC):

$$\rho = \frac{nq}{nq+1}, \tag{1}$$

where ρ is the expected correlation (or ICC) between two independent groups of n participants, and q is the ratio of the item-related variance (σ_β^2) to the noise variance (σ_δ^2) for the database under consideration (for more details, see Courrieu et al., 2011).[1] With this simple equation in hand, for any item-level database composed of n participants by m items, one can use various simple techniques (such as permutation resampling or analysis of variance) to estimate first q, and then the ICC—that is, the amount of variance that models should try to account for (e.g., Rey & Courrieu, 2010).

For example, in Courrieu et al. (2011), we reported two new item-level naming databases collected in English and French. The English database was composed of naming response times for a set of 770 disyllabic English words from 94 participants. Similarly, the French database consisted of 615 disyllabic French words for 100 participants. These databases produced ICCs of .94 and .96 for the English and French databases, respectively, meaning that the resulting item means contained 94 % and 96 % reproducible variance (and, conversely, 6 % and 4 % remaining error variance). Thus, if these estimations are correct, by running a new sample of 100 participants, we should obtain very strong correlations between the new item means and the old ones. The first objective of the present study was to test the reliability of this prediction simply by comparing the previous item means obtained in the French database of Courrieu et al. (2011) with those from a new database composed of a sample of the same items, but with a different group of participants.

The second objective was to better understand the contributions of output articulatory factors to word naming times. Since this source of variance is related to output mechanisms, it is important to quantitatively estimate its impact and to exclude this source of variance in order to keep only the variance related to central reading processes. As we noted above, determining the amount of variance that is inside versus outside the scope of models is a fundamental issue for providing a fair test of their descriptive adequacy. Once the amount of reproducible variance has been estimated, then the amount of variance explained by factors that are outside the scope of a model (such as output articulatory factors) can be factored out, and the resulting residual values can be compared to the performance of models in order to directly test their predictions.

Concerning the naming task, it has already been shown that a large amount of variance is explained by these output articulatory factors, which are clearly outside the scope of models. Spieler and Balota (1997) found that 30 % of the item variance was explained by the phonetic properties of a word's first phoneme (presumably because the acoustic properties of the first phoneme are critical in triggering the voice key; see Rastle, Croot, Harrington, & Coltheart, 2005; Rastle & Davis, 2002). In Courrieu et al.'s (2011) databases, these factors accounted for 35.6 % and 53.1 % of the variance in the English and French naming latencies, respectively. By testing a different group of French participants, we will also check whether these factors consistently account for such a large amount of variance (notably, in the French database). Similarly, as has been suggested by Kessler, Treiman, and Mullennix (2002), we will compare the role of a word's second phoneme in triggering the voice key. Indeed, these authors showed that, due to insufficient energy provided by the first phoneme, voice keys might be triggered for some words by the energy provided by the second phoneme.

Finally, in a recent study using a large database of letter-naming and delayed-naming response times, Madec, Rey, Dufau, Klein, and Grainger (2012) found that the delayed-

[1] Note that the ICC is on the order of a squared correlation, therefore providing a direct estimate of the amount of reproducible variance (see Courrieu et al., 2011).

naming times predicted 79 % of the naming time variance. This result suggests that output articulatory processes might be responsible for a much greater percentage of the total item variance than was initially estimated on the basis of the phonetic properties of the initial phoneme (which might, therefore, not capture all of the articulatory-related variance). Furthermore, delayed-naming measures might provide a better index of this source of variance by including all of the components involved in the production of the articulatory response. In the present new, large-scale database, participants were tested with a procedure combining an immediate-naming and a delayed-naming response, in order to potentially get a better index of the variance related to articulatory processes.

To summarize, in the present study we report a new item-level database in order to further address the issues of the reliability of item means in large-scale databases and of the role of output articulatory factors in single word naming. The new database is composed of naming response times for 400 words randomly extracted from the Courrieu et al. (2011) database, in order to directly compare item means between the old and new databases. This naming task is combined with the same conditional delayed-naming procedure that had been used in Madec et al. (2012). For half of the words (i.e., 200), after the presentation of a target signal (a green circle), participants had to repeat the word that they had just read aloud. For the other half, a no-go signal (red circle) was presented, and participants were supposed to stay silent. The use of both this conditional response (i.e., delayed naming was conditional on presentation of the green circle) and a randomized delay between the naming of the word and the presentation of the green circle are two methodological improvements over standard versions of the delayed-naming procedure (e.g., Balota & Chumbley, 1985). We expected these manipulations to reduce the occurrence of anticipatory responses and improve the quality of the resulting delayed-naming measure.

Method

Participants

A group of 100 undergraduate students from Aix-Marseille University participated in this experiment. All were native French speakers with normal or corrected-to-normal vision.

Material

A list of 400 French disyllabic words was randomly selected from the list of 615 words used in Courrieu et al. (2011). This word list was composed of four- to eight-letter words and excluded verbs and plural forms. Half of the 400 words were randomly selected and tested with the naming and delayed-naming procedures. The other half were associated with the naming procedure only, and not with delayed naming. A PC and the E-Prime software (Psychology Software Tools, Pittsburgh, PA) were used to control the experiment. The responses were recorded using a microphone connected to a PST Serial Response Box (Psychology Software Tools) interfaced with the computer.

Procedure

Each trial started with a fixation point that was presented for 300 ms on the computer screen. This was followed by an empty screen for 300 ms and by the target word, which appeared in the middle of the computer screen in Courier 24-point font in white color on a black background. The word remained on the screen until the participant's response. Participants were instructed to read aloud the target word as quickly and accurately as possible. The *naming time* corresponded to the delay between the presentation of the word and the beginning of the vocal response detected by the voice key. The word then disappeared immediately after the response and was followed by an empty screen for a randomized delay, varying from 1 to 2 s. On 50 % of the trials, a green circle appeared at the center of the screen, and participants had to name the previously presented word again, as quickly as possible. The *delayed-naming time* corresponded to the delay between the presentation of the green circle and the vocal response triggered by the voice key. On the remaining 50 % of the trials, a red circle appeared for 1 s and the participants had to remain silent. The trial ended with the presentation of an empty screen (i.e., the intertrial interval) for 750 ms. Among the 400 words and for all participants, the same 200 words were systematically followed by the green circle (i.e., a delayed-naming response was recorded for this set of words, as well as an immediate naming response), and the remaining 200 words were systematically followed by the red circle (i.e., no delayed naming was recorded for that set of words).

The experimenter sat behind the participant and recorded errors and voice-key failures. The experiment started with a training session composed of 20 trials (ten involving a naming and a delayed-naming response, and ten requiring only a naming response). The experiment then started, with test words presented in a randomized order for each participant. The experiment was divided into three blocks of 140, 140, and 120 trials, respectively, each block starting with a set of five word fillers.

Results

The resulting database was a set of word naming latencies for 400 items by 100 participants, with 7.29 % missing data

(including outliers, error responses, and technical failures). Outliers were defined using a two-step procedure: First, response times less than 250 ms and greater than 1,200 ms were excluded; second, response times that were 2.5 *SD*s below or above a participant's mean were also excluded from the database. Among the 400 tested words, both a naming and a delayed-naming response time were available for half of the words, while only the naming response times were recorded for the other half.

The overall reproducible proportion of item variance in the database was computed using Courrieu et al.'s (2011) method. The overall ICC for the naming measure on the 400 words was equal to .909 (it was .958 in the old database), with a 99.9 % confidence interval of [.886, .929] ([.949, .965] in the old database). The ICC for the naming measure on the restricted set of 200 words was equal to .897, with a 99.9 % confidence interval of [.860, .928]. These analyses indicated that the new database has a slightly greater amount of undesirable error variance (i.e., 9.1 %), as compared to the old database (4.2 %). Still, according to Courrieu et al.'s method, the two databases are highly reliable (with both ICCs > .9), and we should obtain a strong correlation between their item means. However, quite surprisingly, we found that the correlation between the mean naming times for the 400 words from the old and new databases was r = .2675. This extremely low correlation was clearly problematic (relative to the high ICCs), but the next set of analyses, on the role of output articulatory factors, provided an explanation for what was at first a puzzling result.

We first looked at the respective contributions of the first and second phonemes in the old and new databases. Following Treiman, Mullennix, Bijeljac-Babic, and Richmond-Welty (1995), the phonetic properties of these phonemes (i.e., voice, manner, and place of articulation) were coded for each initial and second phoneme within each word as binary variables. One variable coded for voiced versus voiceless phonemes. For coding manner of articulation, five variables were used: fricative versus other, nasal versus other, liquid versus other, high vowel versus other, and low vowel versus other. For coding place of articulation, six variables were used: bilabial versus other, labiodental versus other, alveolar versus other, velar versus other, anterior vowel versus other, and posterior vowel versus other. A hierarchical multiple regression analysis was then conducted on the 400 mean naming times by entering, in a first step, the phonetic properties of the first phoneme, and in a second step, the phonetic properties of the second phoneme (e.g., Yap & Balota, 2009).

The results of this regression analysis are reported in Table 1. We can first see from this analysis that the contribution of the initial phoneme to the naming times was two times greater in the old database (52.96 %) than in the new one (25.25 %). Second, in both databases, we observed a significant and independent contribution of the second phoneme that accounted for an additional part of the item variance (10.24 % and 14.24 % in the old and new databases, respectively). Third, the total amount of variance explained by these two factors altogether was higher in the old database (63.21 %) than in the new one (39.48 %). Clearly, these two databases differed with respect to the contributions of the first two phonemes to the variance of item naming times.

Table. 1 Hierarchical multiple regression analysis testing for effects of the phonetic properties of the first and second phonemes on the 400 word naming times from the old (i.e., Courrieu et al., 2011) and new (present study) databases

		R^2	ΔR^2	F	df	p
Old RT	Phon. 1	.5296	–	–	–	–
	Phon. 2	.6321	.1024	8.70	12, 375	.0001
New RT	Phon. 1	.2525	–	–	–	–
	Phon. 2	.3948	.1424	7.35	12, 375	.0001

To further explore the contributions of articulatory output mechanisms, we conducted another hierarchical multiple regression analysis with the restricted set of 200 words for which we had recorded both naming and delayed-naming times. In that analysis, the contributions of the first and second phonemes were tested first, followed by delayed naming, log frequency (computed on the basis of the Brulex database; i.e., Content, Mousty, & Radeau, 1990), and the number of letters. The results are displayed in Table 2, and apart from being consistent with the previous regression analysis (concerning the contributions of the first two phonemes), they revealed that delayed-naming times accounted for an additional and significant part of the variance of the item naming times (7.82 %). When log frequency and number of letters were added to the regression analysis, these factors also explained independent and additional parts of the remaining variance (9.95 % and 5.01 %, respectively). Note that, in a different analysis, when delayed naming was entered first in accounting for naming

Table 2 Hierarchical multiple regression analysis testing for effects of the phonetic properties of the first and second phonemes, delayed-naming times, log frequency, and number of letters on the 200 word naming times from the database of the present study, for which both immediate-naming and delayed-naming times were recorded

	R^2	ΔR^2	F	df	p	β
Phon. 1	.2207	–	–	–	–	–
Phon. 2	.4242	.2036	5.16	12, 175	.0001	–
Delayed RT	.5024	.0782	27.33	1, 174	.0001	.4507*
Log frequency	.6019	.0995	43.25	1, 173	.0001	–.3206*
Num. of letters	.6520	.0501	24.78	1, 172	.0001	.2655*

* $p < .0001$.

response times, the contribution of the first phoneme was not significant anymore, suggesting that the delayed-naming measure captures all of the variance related to onset properties.

Given that the magnitude of the contributions of articulatory factors to naming times varied greatly from the old database to the new one, we computed the residual values obtained from these two databases when the contributions of the first and second phonemes were factored out from the item means. We then calculated the correlation between these two sets of residual item means, and obtained a result that was clearly more consistent with our expectations, with $r = .9352$, which is very close to the average of the ICCs of the two original databases (.9335).

Finally, we computed the ICC on the delayed-naming times for the restricted set of 200 words (by 100 participants) and found an ICC value of .781. This lower value, relative to the one obtained for naming times (for which we had an ICC of .897), suggests that the delayed-naming times were slightly less reliable than the naming responses. In order to evaluate the contributions of onset properties and of other standard variables (i.e., log frequency and number of letters) to the delayed-naming times, a hierarchical regression analysis was conducted on these 200 item mean delayed responses. As is shown in Table 3, 40.73 % of the variance was explained by the first phoneme, and an additional 16.18 % of the variance was explained by the second phoneme. Overall, the phonetic properties of onset phonemes accounted for 56.9 % of the total variance, and for 72.9 %[2] of the reproducible variance in delayed naming, indicating that a large part of the delayed-naming variance is captured by the properties of the first and second phonemes. Note that when log frequency and the number of letters were entered in the next steps, they did not account for any significant part of the remaining variance (contrary to the result reported by Balota & Chumbley,[3] 1985). These results therefore show that most of the variance in delayed naming is related to output articulatory processes and, more specifically, to the variance related to the phonetic properties of the first and second phonemes.

Discussion

Using a new large-scale database composed of the naming performance of 100 participants on a set of 400 words (together with the delayed-naming performance on a subset of 200 words), we found that, contrary to recent estimates that have been proposed to quantify the amount of reproducible variance in item databases (e.g., Courrieu et al., 2011), the correlation between the item means from the old and new databases was extremely low (i.e., $r = .2675$). By running a series of regression analyses, we found that the phonetic properties of the first and second phonemes contributed significantly and independently to the item variance. The measure of delayed naming was also found to explain an additional and significant part of the item variance. Finally, after factoring out the contributions of the first and second phonemes to item naming times in the old and new databases, the correlation between the resulting sets of residual values reached a value consistent with the expected estimates (i.e., $r = .9352$).

By comparing the performance of two independent groups of 100 participants who performed a naming task on the same set of words and with exactly the same experimental procedure, we found important differences in the amounts of variance explained by the phonetic properties of the first phoneme. This factor accounted for 52.96 % of the item variance in the old database, but only 25.25 % in the new one. In the English database reported by Courrieu et al. (2011), it explained 35.6 % of the item variance, and Spieler and Balota (1997) also obtained similar estimates for their database (i.e., 30 %). Recently, Yap and Balota (2009) found that the amount of variance explained by onset phonemes varied as a function of syllable length. Using the English Lexicon Project (Balota et al., 2007), they reported that for monosyllabic words, onset phoneme properties accounted for 28 % of the variance in naming, while for polysyllabic words, the same properties accounted for only 4.3 % of the variance. This surprising result suggests that the amount of variance explained by onset phonemes is not fixed among all categories of words. This result could be explained by assuming that the amount of variance devoted to visual identification processes is not fixed, either, among word categories. Indeed, if these central processes are fast and efficient (as for high-frequency monosyllabic words), then the total amount of the variance in naming could be influenced less by these processes, and the contribution of articulatory factors might therefore be stronger (for a similar result in single-letter naming, see Madec et al., 2012). Conversely, if central processes are slower and more time consuming, then they would account for a larger amount of the total variance, and as a consequence, the amount of variance captured by onset phonemes would decrease. In fact, that hypothesis is corroborated by the smaller contribution of standard lexical variables for monosyllabic words (23 %) than for polysyllabic words (40.1 %; see Yap & Balota, 2009, Table 3). This explanation can also account for the small amount of variance explained by onset properties in Chateau and Jared's (2003) study conducted with disyllabic words (see Perry et al., 2010, Table 3).

However, the fact that onset properties accounted for larger amounts of variance in disyllabic word naming in both the old

[2] This percentage was obtained by dividing the contribution of the onset phonemes to the total variance (i.e., 56.9 %) by the amount of reproducible variance (i.e., the ICC, 78.1 %).

[3] In fact, the precise role of onset properties was not known at that time. This factor was therefore not controlled in that series of experiments, which could explain the reported frequency effect on delayed-naming times that could, in fact, be confounded with an effect of onset phoneme properties.

Table 3 Hierarchical multiple regression analysis testing for effects of the phonetic properties of the first and second phonemes, log frequency, and number of letters on the 200 word delayed-naming times from the database of the present study

Delayed RT	R^2	ΔR^2	F	df	p	β
Phon. 1	.4073	–	–	–	–	–
Phon. 2	.5690	.1618	5.47	12, 175	.0001	–
Log frequency	.5724	.0034	1.38	1, 174	n.s.	.0648
Num. of letters	.5731	.0007	0.26	1, 173	n.s.	.0292

(52.96 %) and the new (25.25 %) databases in French (as compared to English) could be explained by two different factors. First, onsets and their properties may vary from one language to the other, leading to differences in the amounts of variance accounted for by these onset properties (e.g., the onset phoneme of the word "three" does not exist in French, and the reverse is true in English for the onset of the French word *utile* "useful"). Second, a more technical reason may also account for these differences. Indeed, the sensitivity of most recoding devices (such as the Serial Response Box provided with E-Prime) can be adjusted manually before the beginning of an experiment, which can substantially modify the contribution of onset properties to the total amount of variance in naming responses. Figure 1 illustrates this point by displaying the acoustic waveforms generated by the production of the French word *bateau* "boat" and by showing that, depending on the threshold chosen for the voice key, important variations in response times can be obtained (see also Duyck et al., 2008, Fig. 1). Following that hypothesis, the voice-key sensitivity could have been higher in the experimental setup for the old French database, leading to the detection of finer differences between initial phonemes and, as a consequence, increasing the total amount of variance explained by this factor. Conversely, with a less sensitive voice key, the triggering of the response probably depended on a greater amount of energy in the signal, and this decrease in sensitivity was certainly associated with a decrease in the amount of variance that was accounted for by the initial phoneme. These technical differences in voice-key sensitivity seem to be responsible for the low correlation that we initially obtained between the item means of the two databases. Indeed, when the variations related to the first and second phonemes of a word were factored out, we found a high correlation between the two databases, which was predicted by our analysis estimating the amounts of reproducible variance.

Note that a more drastic solution would be to record the entire naming production of participants[4] and to recalculate manually (e.g., Rastle & Davis, 2002) or automatically (e.g., Duyck et al., 2008) the onset of the acoustic response. Such a method would certainly provide better measures of naming response times, and it would also overcome the problem of having to estimate voice-key sensitivity. However, even though the variance explained by onset properties would likely decrease by using this method, onset properties might still account for a significant part of the variance, because the time dedicated to response execution is likely to be phoneme dependent.

Consistent with Kessler et al.'s (2002) study, we also found that the phonetic properties of the second phoneme accounted for a significant and independent part of the naming variance. This effect can be explained by the weak energy signal that can be associated with some initial phonemes, and the voice key would therefore be triggered by the energy provided by the second phoneme. If we follow the above-mentioned hypothesis about the impact of voice-key sensitivity, then with a more sensitive device, more responses would be triggered by the initial phoneme, and the contribution of the second phoneme would be lowered. Conversely, the role of the second phoneme could be increased with a less sensitive voice key. The reported set of hierarchical multiple regressions are consistent with this assumption, since the amount of variance explained by the second phoneme was lower (i.e., 10.24 % vs. 14.24 %) when the amount of variance explained by the first phoneme was higher (i.e., 52.96 % vs. 25.25 %).

The use of a delayed-naming measure associated with half of the words in the new database allowed us to test whether additional sources of output-related variance could be detected by this behavioral index. We found that this measure indeed captured a significant source of variance that seemed independent of the variance explained by the phonetic properties of the first and second phonemes. However, the precise nature of this additional source of variance still remains a matter of debate. One possibility is related to the binary coding of the phonetic features characterizing each phoneme. With such a simple coding, the precise influence of each phoneme on the variability of naming times (and on the triggering of the voice key) was certainly underspecified. The delayed-naming measure would therefore be a better index of this source of variance. Another possibility is that delayed naming better captures some kind of interaction between the effects of the first and second phonemes (see Rastle et al., 2005). Finally, the fact that the delayed-naming measure accounts for less variance in word naming (i.e., 38.7 %)[5] than in letter naming (79 %; Madec et al., 2012) is likely due to the greater amount of visuo-phonological processing in word naming (including both lexical and sublexical processing), which is responsible for a greater amount of the total variance.

More generally, if the present version of the delayed-naming task accounts for more variance in naming than

[4] Although, a decade ago, this procedure was limited by the memory capacity of most computers, this is no longer an issue for the most recent generation of desktop computers.

[5] This percentage is the result of the linear regression between naming times and delayed-naming times for the restricted set of 200 words.

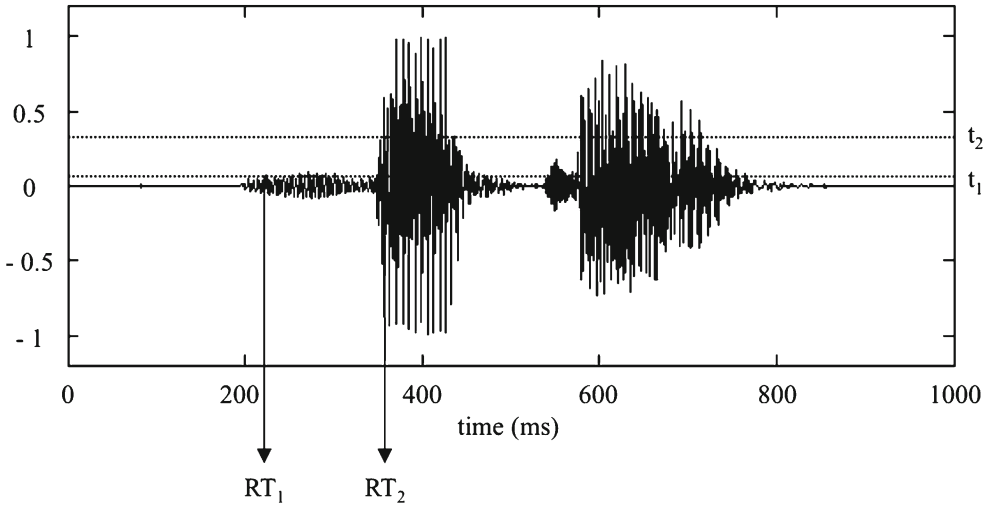

Fig. 1 Sample acoustic waveform for the production of *bateau* "boat." Two very different response times (RT_1 and RT_2) are generated, depending on the sensitivity of the voice key, which is given by two different thresholds (t_1 and t_2)

does a linear combination of the phonetic properties of the first and second phonemes, it may still underestimate the amount of variance related to output articulatory processes. Indeed, as was shown by Kawamoto, Liu, Mura, and Sanchez (2008), delayed naming is traditionally viewed as a good measure of the variability related to response execution processes. On the basis of the motor production model proposed by Sternberg, Monsell, Knoll, and Wright (1978; see also Sternberg, Knoll, Monsell, & Wright, 1988), delayed naming is considered as reflecting the duration of response execution processes. While naming can be decomposed into two sets of processes (i.e., a first set including perceptual and response selection processes, and a second set gathering response execution processes), in delayed naming, it is usually considered that a response has been selected and is maintained in a response buffer until the presentation of a signal that triggers response execution processes. Since a constant signal is always used as a trigger, it is assumed that variance in delayed naming only reflects variability in response execution.[6] However, contrary to this idealistic view, Kawamoto et al. found that, in delayed naming, participants frequently activate part of the motor response before the presentation of the response signal. For example, for plosives, participants can already prepare the motor response by building up the lip pressure, and therefore accelerating the generation of the acoustic signal. Thus, part of the response execution variance is lost in this case, due to articulatory preparation. By controlling for this kind of response preparation, future research may therefore reveal that response execution processes account for slightly more variance than we have observed with the present version of the delayed-naming task.

[6] As was noted by a reviewer, the conditional procedure that we used in the present version of the delayed-naming task also requires a decision component. However, in the same way that detection of the green circle should be item independent, the variance related to that decision component can also be considered as being constant across items.

A final caution concerns the type of coding that we used for characterizing the phonetic features of the first and second phonemes (i.e., the one proposed by Treiman et al., 1995). Although our results suggest that this coding allowed us to capture most of the variance related to the phonetic properties of the initial phonemes (due to the high correlation obtained between the residual values of the old and new databases when the variance explained by these codes had been factored out; i.e., $r = .94$), future work may reveal that slightly more variance can be accounted for either by using different coding schemes or by adding interaction terms in the hierarchical regression (e.g., between manner and voicing, given that the effect of voicing may vary for, e.g., fricatives and stops).

Turning back to the project of accounting for the item-level variance in large-scale databases, what have we learned from the present set of results? First, we already knew that a large amount of the item variance in naming is related to output articulatory constraints (e.g., Spieler & Balota, 1997), but the current practice for testing the predictions of computational models is still to consider that this source of variance is mainly explained by the phonetic properties of the first phoneme (e.g., Perry et al., 2010). Following Kessler et al. (2002), we have shown that both the first and second phonemes play a role, and therefore, when testing models, we should factor out these undesirable sources of variance that are outside the explanatory scope of models. Second, even if most of the output-related variance seems to be captured by the phonetic properties of the first and second phonemes, using a delayed-naming measure provided a better estimate for this source of variance. Third, the comparison of the old and new databases revealed that variations in the sensitivity of the recording device (i.e., the voice key) might dramatically change the pattern of item-level naming response times. However, we have shown that by factoring out the contributions of the first and second phonemes, this bias can be satisfactorily cancelled. We also have shown that if delayed-naming times are not available,

factoring out the variance related to the first and second phonemes seems to provide a reasonable solution for excluding most of the variance related to output articulatory factors. Finally, the present data set provides a positive empirical test of the method for estimating the amount of reproducible variance for any item-level database proposed by Courrieu et al. (2011), provided that all of the components of the experimental procedure have been carefully controlled and specified (e.g., voice-key sensitivity, in the case of the naming task).

Author note The authors thank Max Coltheart, Michael Cortese, and one anonymous reviewer for their helpful comments and suggestions. We are also indebted to Bob McMurray for his invaluable editorial work. Part of this study was funded by ERC Advanced Research Grant No. 230313.

References

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language, 24,* 89–106.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods, 39,* 445–459. doi:10.3758/BF03193014

Chateau, D., & Jared, D. (2003). Spelling–sound consistency effects in disyllabic word naming. *Journal of Memory and Language, 48,* 255–280.

Content, A., Mousty, P., & Radeau, M. (1990). Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique, 90,* 551–566.

Courrieu, P., Brand-d'Abrescia, M., Peereman, R., Spieler, D., & Rey, A. (2011). Validated intraclass correlation statistics to test item performance models. *Behavior Research Methods, 43,* 37–55. doi:10.3758/s13428-010-0020-5

Courrieu, P., & Rey, A. (2011). Missing data imputation and corrected statistics for large-scale behavioral databases. *Behavior Research Methods, 43,* 310–330. doi:10.3758/s13428-011-0071-2

Duyck, W., Anseel, F., Szmalec, A., Mestdagh, P., Tavernier, A., & Hartsuiker, R. J. (2008). Improving accuracy in detecting acoustic onsets. *Journal of Experimental Psychology. Human Perception and Performance, 34,* 1317–1326.

Kawamoto, A. H., Liu, Q., Mura, K., & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language, 58,* 347–365.

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic biases in voice key response time measurements. *Journal of Memory and Language, 47,* 145–171.

Madec, S., Rey, A., Dufau, S., Klein, M., & Grainger, J. (2012). The time course of visual letter perception. *Journal of Cognitive Neuroscience, 24,* 1645–1655.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407. doi:10.1037/0033-295X.88.5.375

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the connectionist dual process (CDP++) model. *Cognitive Psychology, 61,* 106–151. doi:10.1016/j.cogpsych.2010.04.001

Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology. Human Perception and Performance, 31,* 1083–1095.

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

Rey, A., & Courrieu, P. (2010). Accounting for item variance in large-scale databases. *Frontiers in Psychology, 1,* 200. doi:10.3389/fpsyg.2010.00200

Rey, A., Courrieu, P., Schmidt-Weigand, F., & Jacobs, A. M. (2009). Item performance in visual word recognition. *Psychonomic Bulletin and Review, 16,* 600–608. doi:10.3758/PBR.16.3.600

Seidenberg, M., & Plaut, D. C. (1998). Evaluating word reading models at the item level: Matching the grain of theory and data. *Psychological Science, 9,* 234–237.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8,* 411–416. doi:10.1111/j.1467-9280.1997.tb00453.x

Sternberg, S., Knoll, R. L., Monsell, S., & Wright, C. E. (1988). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica, 45,* 175–197.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117–152). New York: Academic Press.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word recognition. *Journal of Memory and Language, 26,* 608–631.

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.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60,* 502–529. doi:10.1016/j.jml.2009.02.001