Studies in Second Language Acquisition **40** (2018), 225–239
doi:10.1017/S0272263117000043

Research Report

RESPONSE TIME VARIABILITY SIGNATURES OF NOVEL WORD LEARNING

Katya Solovyeva*

University of Maryland, College Park

Robert DeKeyser

University of Maryland, College Park

Abstract

Response time variability and its changes over time have been interpreted as indicative of levels of knowledge automatization. Predominantly, only declines in variability have been examined over the course of practice and growing second language proficiency. We discuss possible scenarios that may involve increasing, rather than declining variability, such as the establishment of new memory representations. We present repeated-measures lexical decision data on novel words encountered incidentally in a lexical decision task, which show increased variability as familiarity increases. Similar patterns were observed in our reanalyses of data from other novel word learning experiments with task demands different from ours: auditory presentation and a posttest after seven days (Brown & Gaskell, 2014), as well as picture-matching and production tests (Bartolotti & Marian, 2014). Jointly, these results suggest that when no prior representation of a word exists, increased variability may index learning during its initial stages.

BACKGROUND AND MOTIVATION

Response time (RT) variability has been proposed as a marker of processing automaticity (Segalowitz & Segalowitz, 1993) in both first and second language performance. One operational definition of RT variability is the coefficient of variation (CV)—computed as standard deviation at the given level of mean response time (SD/*M*). It has

This work was supported in part by NSF-IGERT award DGE-0801465 to Colin Phillips. The authors thank Helen Brown for providing pertinent data for the calculation of the metrics used in this paper; Mike Long, Stephen O'Connell, and Ilina Kachinskse for comments on earlier drafts; and the University of Maryland Graduate Writing Center and Dr. Linda Macrí. We are grateful for the feedback from anonymous reviewers, whose comments resulted in considerable improvements to the paper.

*Correspondence concerning this article should be addressed to Katya Solovyeva, 2321 Symons Hall, University of Maryland, College Park, MD 20742. Email: kss@umd.edu

Copyright © Cambridge University Press 2017

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

been in use in neuropsychological studies for decades (e.g., Rastatter & Blair, 1984) and was reinterpreted by Segalowitz and Segalowitz (1993) as an operational definition of RT variability that could empirically distinguish between automatization and speed-up.

In the present paper, we review the evidence supporting the interpretation of the CV as an index of automaticity, as well as limitations that have become apparent in its application. We further argue that, in addition to the domain of practice and automatizing existing knowledge, for which the CV was initially proposed, an examination of learning situations falling outside of that domain may be desirable for its validation. One such extension is the development of new knowledge representations. We present data from an experiment involving the learning of novel surface forms of words and of novel word-meaning associations and discuss both RT and CV signatures of learning. While our present results are most comparable with the original studies, thanks to a high similarity of methods (lexical decision tasks [LDTs]), we augmented our analysis including two reanalyses of other novel word-learning studies. These were chosen to represent a variety of modalities (auditory—Brown & Gaskell, 2014) and tasks (picture matching, production—Bartolotti & Marian, 2014). We used the CV to complement the authors' original data; these show a pattern of increased variability over time that is similar to our own results.

According to the original proposal by Segalowitz and Segalowitz, the CV metric allows one to capture any changes in variability above and beyond those expected simply due to speed-up, when SDs decline in absolute terms at the same pace as mean RTs. Declining CVs, by contrast, result from reductions in variability that exceed the reduction of raw RTs—arguably due to the restructuring of knowledge representations during automatization—which characteristically involves the elimination of slower, less efficient component processes.

Declining CVs have been demonstrated with increasing proficiency in cross-sectional designs involving lexical (Hulstijn, Van Gelderen, & Schoonen, 2009; Segalowitz & Freed, 2004), morphological (Rodgers, 2011), and syntactic processing (Lim & Godfroid, 2014). Longitudinal within-subject studies have been rare, but their findings have tentatively supported the expected directions of change in variability. More finely grained measures of online psycholinguistic processing have also been shown to display the pattern of variability reduction with increased proficiency (Leung & Yoshikawa, 2015).

However, failures to replicate have been attested, as well; although not numerous, they have been helpful in sketching the limits of CV's sensitivity. First, Hulstijn et al. (2009) did not find evidence of a decline in CVs on three out of five tasks, lexical as well as syntactic, involving both comprehension and production. The data in question were obtained on three measurement occasions within 48 months from a sample of secondary school students in a second language learning context; one third of the participants were bilingual in Dutch and another language. In subsequent discussions of these results, the nonreplication has been attributed to insufficient time or exposure for participants to make significant progress (Lim & Godfroid, 2014). Crucially, follow-up work using the same tasks by Lim and Godfroid (2014) showed a very similar pattern of CV effects, depending on the nature of processing required by the tasks. The most robust effects have been obtained on tasks recruiting low-level processing skills that are particularly amenable to automatization, such as lexical decision and semantic classification; and

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

highly constrained sentence construction tasks, such as the maze task, which measures RTs at many "forks" throughout a sentence. By contrast, tasks that reflect the joint contributions of many processes and include reasoning components (such as sentence truth-value judgments) seem to be poor candidates for demonstrating CV effects (cf. Lim & Godfroid, 2014, p. 25). Performance on such cumulative tasks may be susceptible to different strategies in different subjects and to their differential reliance on the putative subprocesses. For example, truth-value judgments may be driven by associative semantic information and minimal syntactic analysis in some subjects, whereas in others, fine-grained parses of sentences may be computed. Any variability at this higher level would be in addition to the variability inherited from the lower levels of processing, such as lexical access; furthermore, it is not guaranteed that variability at these two levels would be correlated or change in the same direction with increasing proficiency. This suggests that the CV taps extremely localized changes in RTs at shorter time-scales and proficiency levels that still allow for growth, although determining these may depend on the skill in question.

Leaving aside the inconsistencies in empirical data pointed out in the preceding text, several conceptual problems have been raised with respect to the CV's interpretation. The main barrier to its widespread adoption is that it conflates the development of knowledge with improvements in processing (Hulstijn et al., 2009). This led researchers to propose restricting the computation of CVs to known stimuli (Hulstijn et al., 2009), arguably bolstering its validity as a measure of automaticity—conceived as the availability of knowledge already in place. Such a restriction would be consistent with both theories of skill acquisition that are foundational for the CV (Hulstijn et al., 2009, p. 556). Both rely on the prior existence of knowledge: On Anderson's account (Anderson, 1983), skill acquisition starts with the development of declarative knowledge, so something that is not represented in memory cannot be automatized. In Logan's model (Logan, 1988), the products of mental computations are stored in memory and later compete with the procedures that generated them.

Just like any other indicator of change in an unobservable underlying process, the CV is only as good as the theory behind the underlying process. Care should be taken to formulate hypotheses that can be falsified, counteracting the ease with which conflicting findings have been accommodated by post hoc reasoning, particularly relating to suboptimal amounts of time between measurements and the attendant levels of proficiency—too much, so the automatization period has passed (Akamatsu, 2008) or too little, so the automatization has not begun in earnest (Lim & Godfroid, 2014, p. 23, on Hulstijn et al.'s findings).

In our view, testing the logical extensions of Segalowitz and Segalowitz's proposal would be illuminating for the measure's validity even in the absence of conflicting data. In a perfect world where all reports agreed, testing the full range of the predictions generated by the hypothesis would be desirable as attempts to falsify the hypothesis (cf. Lim & Godfroid, 2014, p. 27). In the present situation, where discrepancies occur, testing additional contexts in which CVs do and do not change could potentially explain conflicting findings. There are at least two such extensions that the present study tests. First, at the level of single-word tasks, CVs have been reported for word targets only in such tasks as animacy classification and lexical decision. Although Segalowitz and Segalowitz do not make explicit claims about RTs to words compared to nonwords, it

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

follows from their argument that word targets should elicit less variable responses by being overlearned during one's life.

Second, as originally proposed, it is the patterns of change in variability, rather than its absolute values, that are consequential for inferences regarding automaticity. Specifically, declines in the CV were deemed indicative of automatization accompanying increasing proficiency. However, declines in CV are only meaningful to the extent that other patterns of change are attested or at least theoretically possible: If only decreases were observed across all studies, the CV as an indicator would be uninterpretable. Lack of change in variability, although reported in the literature, does not present the strongest possible evidence of the full scope of possible CV dynamics. To our knowledge, in the literature on second language learning there are no reports of increasing variability, nor a discussion of when such a pattern would be expected—in contrast to neuropsychological populations, where increases in variability have been taken to reflect breakdowns in processing (e.g., among others, Hultsch, MacDonald, Hunter, Levy-Bencheton, & Strauss, 2000).

If reductions in the CV result from the elimination of (inefficient) component processes or their restructuring, increases in CV should signal the opposite—the addition of component processes or new representations to the set that jointly determines response latencies. In situations involving the development of new knowledge, the lack of separation between the improvements in knowledge versus its processing may, in fact, be instrumental and make it possible to use the CV as an indicator of learning. The idea that increased variability may accompany the development of knowledge is by no means new and has been present in various discussions of morphological development, at least regarding nonlinear patterns in accuracy (e.g., Meisel, Clahsen, & Pienemann, 1981). Common to such proposals is the notion of disruption in the knowledge system following the refinement or addition of new knowledge.

Building on this broad argument, we hypothesized that the automatization of existing knowledge, supported by established representations, may follow a different time course and have a different mean RT/CV signature than learning in situations requiring the establishment of new representations—whether of word forms or of new form-meaning mappings. Whereas all existing research employing the CV falls into the former category, automatization of existing knowledge, the present study focuses on two situations involving the learning of novel stimuli.

First, we consider the learning of surface attributes of novel word forms from mere exposure; this kind of learning represents the lower bounds of learning. Second, we examine the learning of new form-meaning pairings, using associations between known words in the first language and their novel translation equivalents during intentional learning in recall practice. Our analyses follow the concomitant changes in RT and RT variability in both learning scenarios. In the first case, if our extended reading of Segalowitz and Segalowitz's proposal stands, we should observe an increase in variability (CVs) over time, even in the presence of overall faster RTs, which would result from mere familiarity with task and response demands. In the second case, if novel form-meaning pairings have been learned, one would see lower RTs to the stimuli presented in trained pairings (compared to before training) and higher RTs to targets presented in mismatched pairings—compared to the matched stimuli but not necessarily before the training. CVs to word targets should increase.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

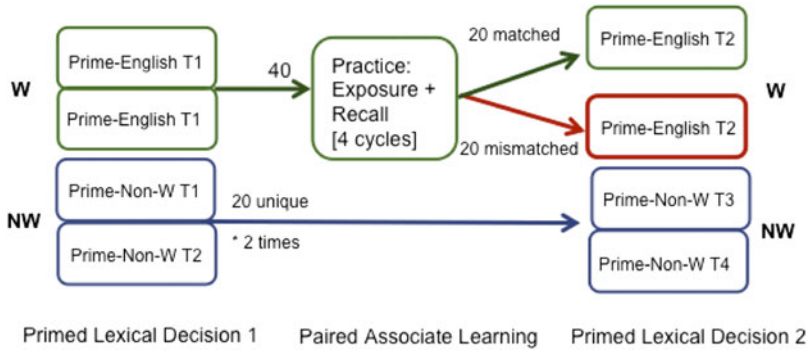


FIGURE 1. Study design and progression of tasks.

Finally, addressing the comparison between words and nonwords, according to the original proposal of Segalowitz and Segalowitz, one would predict faster RTs to words than nonwords coupled with lower variability of response latencies to word targets.

METHODS

The experiment included a lexical decision task (LDT1) with English word targets primed by Swahili translation equivalents and nonword targets primed by Swahili stimuli; a paired-associate learning task (PAL) with only the Swahili-English word pairs; and a second lexical decision task (LDT2) with the same stimuli as the first, except for half of the word stimuli being mismatched and presented with wrong Swahili translations (Figure 1). Therefore, the nonword targets and their primes were only encountered during lexical decision and thus learned incidentally, while the word targets and their Swahili translation equivalents were encountered both in LDTs and in a recall task forcing their deliberate practice. Tracking mean RT and CV changes for word versus nonword stimuli allowed us to contrast the effects of practice on already known words and the effects of repeated exposure to novel stimuli that are yet to be represented.

The inclusion of the intervening paired-associate practice task makes our design similar both to those of Akamatsu (2008), Fukkink, Hulstijn, and Simis (2005), and Segalowitz, Watson, and Segalowitz (1995), whose experiments had built-in practice, as well as, indirectly, comparable to Segalowitz, Segalowitz, and Wood (1998) and Segalowitz and Freed (2004), whose repeated administrations of the experimental tasks were spaced widely enough to capture effects of concurrent L2 instruction external to the experiments.

PARTICIPANTS

Seventy-three undergraduate and graduate students (26% male) were recruited to participate in the study in exchange for course credit; 54 of them completed all tasks (Figure 1). All of them were native speakers of American English (16% left-handed) with no more than three semesters of foreign language instruction (excluding high school), screened to control for confounding effects of concurrent foreign language instruction. Their ages ranged from 18 to 26.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

PROCEDURE

After reading and signing the consent form, participants completed an LDT, which included 40 word targets preceded by Swahili primes (requiring a yes response); 20 nonword targets (requiring a no response) preceded by Swahili primes, both phonotactically legitimate in English, and presented twice—for a total of 40 nonword trials but only 20 unique items. After that, the 40 Swahili-English word pairs were practiced in a PAL task that involved four alternating study and recall phases; the 20 pairs of Swahili primes and nonword targets were not practiced in PAL.

Finally, the LDT was administered a second time: It included the 20 Swahili-nonword pairs (not practiced during PAL) presented twice again, and the 40 Swahili-English pairs, which had been practiced during PAL. The Swahili-English pairs were randomly assigned to be presented either in their original pair assignments or mismatched, paired with novel translations from the same pool of pairs (see "Materials" section).

Thus, for nonword targets, the two LDTs created a sequence of four RT measurements: times 1 and 2 before PAL (in which they were not included), and times 3 and 4 after PAL. For word targets, there were two measurements, one before and one after PAL training. At the end of the experiment, subjects were debriefed and thanked for their participation.

During both LDTs, the presentation and timing of stimuli were controlled using DMDX software (Forster & Forster, 2003). Each trial started with a fixation cross presented for 330 ms, followed by a prime presented in lowercase letters for 200 ms and 100 ms of blank screen. After that, a target was presented for 200 ms in uppercase, yielding a stimulus onset asynchrony (SOA) between the primes and the targets of 300 ms. This SOA value represented a balance between giving participants sufficient time to process completely novel stimuli (primes) while still minimizing strategic expectancy effects that could contaminate potential learning-induced priming. According to Altarriba and Basnight-Brown (2007), 300 ms is at the high end of recommended SOAs for capturing priming, but it is worth emphasizing that their recommendations were developed for studies where learners already had some knowledge of the language, which was not the case in our experiment.

MATERIALS

Both LDTs (before and after PAL) included the same 40 word targets (English) and 20 unique nonword targets (presented twice within each LDT). The word targets were high-frequency words of English of lengths from three to eight characters ($M = 5.42$, $SD = 1.19$). The nonword targets were generated using the MCWord database (Medler & Binder, 2005) to match the word targets on the following variables: word length in letters, number of orthographic neighbors (Coltheart's N), average frequency per million of the orthographic neighbors, average frequency of all bigrams in a word, and average frequency of all trigrams in a word (see Online Supplementary Material for actual items and Table 1 for item characteristics).

Two-tailed t-tests of mean difference were run and failed to reject the null hypothesis of lack of difference in mean scores between words and nonwords on all attributes of interest.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

TABLE 1. Stimuli characteristics

	M (range), SD		t(df), *p*,2-tailed
	Nonwords	Words	
Length	5.8 (4–8), 1.24	5.42 (3–8), 1.19	1.12 (37), *0.27*
Neighborhood size	4 (0–21), 5.75	3.87 (0–18), 4.2	0.05 (29), *0.96*
Neighborhood frequency	42 (0–312), 80.63	42 (0–473), 90.71	0.005 (42), *0.99*
Bigram frequency	1176 (287–4242), 1006	1117 (43–5651), 1018	0.21 (38), *0.83*
Trigram frequency	349 (51–1689), 446	223 (15–1432), 281	1.16 (27), *0.26*

Note: Means are followed by ranges in parentheses; standard deviations are presented after a comma.

For both word and nonword targets, the primes were always Swahili words. For word targets, this was done to preclude participants from forming conflicting associations between English words that would later reappear during training as translations of Swahili items and other irrelevant words. Given this restriction, the only option was to present as priming pairs the pairings that would be learned later. Primes for nonwords could not be different from the primes for words because this would encourage use of strategies and render the targets unnecessary for making the lexical decision.

RESULTS

Following standard practice in psycholinguistics, only correct responses were included in all analyses. Reaction times shorter than 200 ms or longer than 2,000 ms were treated as missing values. For each participant, we calculated mean reaction times for word and nonword stimuli. For nonwords, there were four measurement occasions, two before and two after PAL. For words, there were two measurement occasions, one before and one after PAL; after PAL, half of the word targets were presented with wrong Swahili translations, that is, ones that differed from the trained pairings. The coefficient of variability was calculated for each participant separately for each stimulus group, mirroring the calculations of mean reaction time; thus, there is a separate CV for nonwords at times one, two, three, and four, and for words before and after learning, subdivided into matched and mismatched groups. All descriptive statistics for RTs and CVs are presented in Table 2.

TABLE 2. Descriptive statistics for reaction time and coefficient of variation

		Time 1	Time 2	Time 3	Time 4
Nonwords	RT, ms	691.53 (16.94)	657.47 (10.17)	559.64 (14.52)	501.38 (15.37)
	CV	0.17 (0.01)	0.18 (0.01)	0.22 (0.01)	0.22 (0.02)
Words	RT: mismatched	569.97 (10.53)		489.14 (14.35)	
	RT: matched	564.41 (10.61)		464.1 (12.97)	
	CV: mismatched	0.21 (0.01)		0.28 (0.02)	
	CV: matched	0.21 (0.01)		0.26 (0.01)	

Note: Standard errors are presented in parentheses.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

DEVELOPMENT OF NEW KNOWLEDGE REPRESENTATIONS

Learning of Novel Word Forms

The first domain of interest in forming new knowledge representations was the learning of new surface forms of words alone, as reflected by four repeated measures of response latencies to nonword stimuli: two within the first LDT (before PAL) and two during the second LDT (after PAL). The nonword stimuli were not practiced during the PAL task intervening between the two LDTs. The learning that we assume for these stimuli would be superficial, which can be construed as the lowest bound of new knowledge formation—the creation of novel representations in memory. Even without practice, reaction times were expected to decline from presentation to presentation due to increased familiarity with the sensory aspects of the nonwords, as well as task demands. With respect to CVs, we hypothesized that, according to the extension of the CV's original proposed meaning, an increase should be expected following learning. However, in the context of the response becoming more practiced—potentially leading to decreased variability—we were cautious about predicting whether the magnitude of such a learning-induced increase would result in a *net* increase in variability.

Changes in Mean Reaction Times

Reaction times to nonword targets declined throughout the four presentations (Figure 2): A one-way repeated-measures ANOVA revealed a significant effect of Time, F (2.44, 119.62) = 107.27, $p < .001$, partial $\eta^2 = .69$. Degrees of freedom were corrected using the Greenhouse-Geisser value of sphericity, $\varepsilon = .81$, because Mauchly's test indicated a violation of the sphericity assumption: χ^2 (5) = 17.76, $p = .003$.

Changes in Variability

Subject-level variability of responses to nonwords *increased* over the measurements (Figure 2). These changes were statistically significant, as indicated by a repeated-measures ANOVA, $F(2.01, 98.66) = 9.78$, $p < .001$, partial $\eta^2 = .17$, with degrees of freedom corrected using the Greenhouse-Geisser method for nonsphericity, $\varepsilon = .67$, χ^2 (5) = 38.73, $p < .001$.

Learning of New Form-Meaning Pairings

The second area of interest was the emergence of associative links between known words and their translation equivalents in a new language, instantiated by Swahili-English parings. These pairings were also practiced during four iterations of exposure followed by recall, which were administered between the two LDTs. During the second LDT, half of the word targets were presented in pairings with primes different from the first LDT and the recall practice tasks; stimuli were assigned to either condition randomly, and all critical sublexical characteristics were balanced for the two conditions (Table 1). Therefore, we compared both the CV and mean RTs for the matched and mismatched halves of the stimuli before and after practice.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

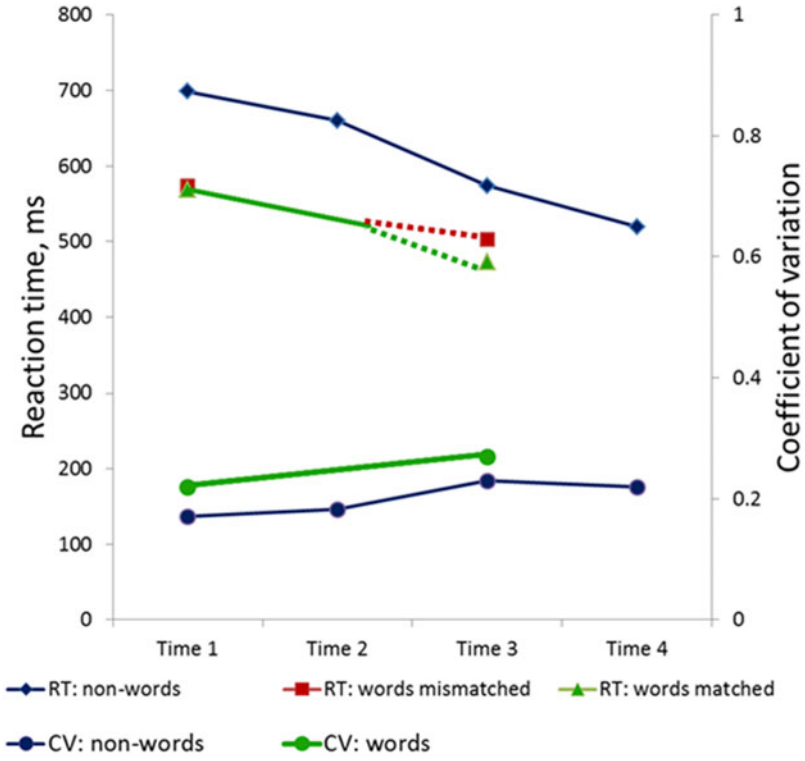


FIGURE 2. Declines in mean RT to words and nonwords over two (words) and four (nonwords) presentations. Increased variability to both nonwords and words.

We expected an interaction between the variables Time and Status (matched, mismatched) at least for the RT analysis, such that mismatched pairs would elicit slower responses, but only after the practice. Before any practice, the matched/mismatched distinction was meaningless and no difference in mean RT was expected due to random assignment to the matched/mismatched condition. For the CV comparison, we predicted a main effect of time only, but not necessarily an interaction with matched status. Because mismatched pairings were created by rearranging previously encountered ones, if variability were sensitive to purely sensory attributes of the stimuli no changes would be expected—regardless of matching, all primes and targets would be equally well practiced. If, however, the CV changed in response to the association strength between primes and targets, CV changes could be observed with changes in pairings.

Changes in Mean Reaction Times

A repeated-measures ANOVA showed an interaction of Time and Status [matched, mismatched] in determining response latencies to word targets, $F(1, 49) = 8.19$, $p < .01$, partial $\eta^2 = .14$, indicating that RTs to matched pairings declined significantly more ($M_1 = 564.41$ ms, $M_2 = 464.1$ ms) than did RTs to mismatched pairings ($M_1 = 569.97$ ms, $M_2 = 489.14$ ms) over time. Time showed a significant main effect, $F(1, 49) = 60.2$,

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

$p < .001$, partial $\eta^2 = .55$. There was a main effect of Status, $F(1, 49) = 11.24$, $p < .01$, partial $\eta^2 = .19$, driven by slower RTs to mismatched than matched pairs after, $t(72) = 4.29$, $p < .001$, but not before, practice, $t(53) = 1.18$, $p = .24$, as predicted.

Changes in Variability

Similar to the mean RT analysis for words, we compared subject-level CVs for matched and mismatched words before and after practice and included both Time and Status as predictor variables, as well as their interaction, acknowledging it as possible but not necessarily expected. Indeed, the interaction between Time and Status [matched, mismatched] was not significant, $F(1, 49) = .50$, $p = .48$, suggesting that, in the absence of a main effect of Status, $F(1, 49) = .39$, $p = .54$, both matched and mismatched stimuli displayed similar increases in RT variability from prepractice to postpractice, as indicated by the main effect of Time, $F(1, 49) = 18.7$, $p < .001$, partial $\eta^2 = .28$.

WORDS VERSUS NONWORDS

As commonly attested in the psycholinguistics literature, word targets were expected to elicit faster responses than nonwords at all measurement points. We expected higher variability of responses to word versus nonword targets. First, the automaticity account would predict that words should elicit less variable responses by being highly practiced or overlearned. Second, the automaticity literature explicitly acknowledges frequency effects and predicts less variability for highly frequent items and more variability for less frequent items; on this view, nonwords are just an extreme case of low frequency. For this comparison, we focus on the first LDT before practice (PAL), specifically, on responses to word targets compared to nonword targets at the first presentation only; these reflected normal lexical access processes. By contrast, responses *following* PAL reflected learning, from mere repeated exposure for nonwords and from deliberate practice for words.

Mean Reaction Times

Nonwords at the first presentation were responded to slower than either matched, $t(53) = 10.43$, $p < .001$, Cohen's $d = 0.37$, or mismatched, $t(53) = 9.62$, $p < .001$, Cohen's $d = 0.31$, word targets.

Variability

RT variability was significantly higher among word targets than nonword targets: matched $t(53) = -3.99$, $p < .001$, Cohen's $d = -0.66$; mismatched $t(53) = -3.37$, $p = .001$, Cohen's $d = -0.59$. Even though the matched-mismatched distinction is meaningless at the first presentation, we report separate t-tests for the two groups of stimuli to facilitate comparison with post-PAL data.

DISCUSSION

As Figure 2 shows, variability and mean RTs changed in opposite directions following both the establishment of new form-meaning pairs during deliberate study (words) and

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

mere exposure to novel word forms (nonwords). Moreover, variability was higher in words than nonwords, a finding not directly mentioned in, nor following from, Segalowitz and Segalowitz's argumentation. On a narrow reading of their proposal, which guided our predictions, one would expect words as a class to elicit less variable responses than nonwords because words would arguably have been practiced throughout one's entire life. However, this literature also typically incorporates frequency effects and thus posits some heterogeneity in automatization in the lexicon, leaving room for inferring that *novel* words—or "nonwords" in the context of the present study—would as a class be less variable because nonword status would act as an equalizing force, as long as their sublexical properties are closely matched to those of the word targets. It is this latter extrapolation that was borne out by our data.

There are two points of uncertainty with the present design, however, which, if addressed, could make our conclusions from the results more compelling. Our reanalyses of other novel word learning data in the next section address these in turn. First, our experiment may have induced a conflict between the participants' growing familiarity with the nonword stimuli and the requirement to keep treating them as nonwords for the purposes of the LDTs. Second, shorter lexical decision latencies on the nonwords are nonspecific indicators of learning broadly conceived, and are not indicative of its depth. In particular, the status of newly represented word forms in the mental lexicon cannot be determined based on RTs alone. Two indirect ways for probing just what this status might be are exemplified in the studies selected for the reanalysis section. First, examining lexical competition effects exerted by the novel words on existing lexical entries can be highly illuminating, as Brown and Gaskell (2014) have shown. Second, varying the response demands of tasks may help mitigate threats to validity stemming from stimulus-response binding effects that are an inherent part of priming paradigms (Henson, Eckstein, Waszak, Frings, & Horner, 2014) by tapping lexical processes in an almost latent-trait fashion, extracting shared aspects across lexical-decision and production tasks, as in Bartolotti and Marian (2014).

OTHER STUDIES OF NOVEL WORD LEARNING

The studies by Brown and Gaskell (2014) and Bartolotti and Marian (2014) provide an indirect comparison to our results and tentative support for our conclusions. In contrast to the results reported in the preceding text, where individual-level RTs and response variability were analyzed, these reanalyses of necessity involve data *aggregated* across subjects, admittedly qualifying our interpretation, especially in the absence of significance testing. However, interpreted in the context of our own findings, these observations may nevertheless add value to the exploration of the different RT signatures of learning and automaticity.

Brown and Gaskell (2014)

Brown and Gaskell (2014) report a series of experiments on the incidental learning of novel words during a phoneme monitoring task with immediate, next day, and next week posttests. The learning was similar to nonwords in our experiment, in that the stimuli were encountered in the absence of paired meanings. Unlike the present experiment,

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

Brown and Gaskell's study gauged the status of novel words through the amount of lexical competition experienced by known words already established in participants' lexicons and sharing the same onset with the novel competitors. Thus, the learning of novel words would be reflected in *slower* RTs to those words' neighbors, compared to a group of control words. For example, if *biscal* is the novel word being learned, participants would encounter *biscuit* in the LDT; if *biscuit* is responded to slower than a closely matched control word (e.g., *blossom*) whose neighbor had not been encountered in the learning task, the slowdown would be attributable to the incorporation of *biscal* into the participant's mental lexicon. For the purposes of the present paper, the fact that Brown and Gaskell measure the variability of RTs to words rather than the novel (non)words is a fortunate design feature because it eliminates the conflict between making a lexical decision response to nonwords and their growing familiarity.

The three experiments reported by the authors differed in talker variability between the incidental exposure to novel words and the testing phase. Although this manipulation is irrelevant for our purposes, we report the data respecting the original design, without collapsing the two test conditions into one. Thus, the two test conditions included the learned nonwords' real word competitors, which had not been encountered during the training; the competitors were either presented in the same voice at testing as they had been during training or in a different voice. The control condition included words balanced on all relevant metrics with the words from the two test conditions; the control words did not have any neighbors that had been encountered incidentally during training and, therefore, no source of lexical competition pressure. As Figure 3 shows, variability of responses in both test conditions was higher than that in the control condition throughout the three experiments and on all test days, including the delayed posttests one week after initial exposure.

In the authors' analysis of mean RTs, in the first experiment the difference between test and control items was only marginally significant. By contrast, the difference in variability, as reflected in the CVs, emerges early and persists until the follow-up posttest one week after learning (left panel of Figure 3). In the second experiment, the RT analysis showed that test items were responded to slower than control items only on day 2, whereas no difference was detected on days 1 and 7. A CV analysis, however, showed that the two conditions did differ in variability on all three days. Variability was higher in

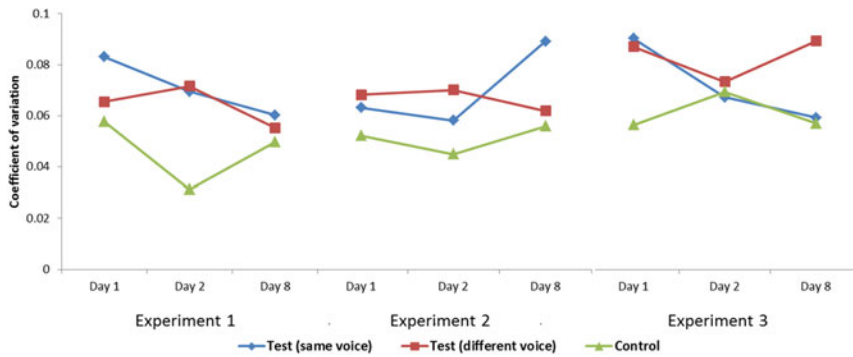


FIGURE 3. Changes in RT variability in three word-learning experiments by Brown & Gaskell (2014).

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

the test condition (i.e., stimuli that have newly learned competitors) on all three days (middle panel of Figure 3). In the third experiment (rightmost panel of Figure 3), no differences in mean RTs were observed between test and control items, but the CV differences were present at two out of three measurement points, that is, days 1 and 7 (rightmost panel of Figure 3). Together, these findings suggest that mean RTs alone may not always be sensitive to emergent learning and that CV increases can signal shifts in the status of newly learned stimuli words in the mental lexicon.

Bartolotti and Marian (2014)

The study by Bartolotti and Marian (2014) involved deliberate learning of novel nonwords that differed in their "wordlikeness" and were paired with pictures. The testing involved timed recognition, operationalized as matching the novel words to pictures, and production. Declining RTs across five blocks of repeated testing are accompanied by an increase in variability (Figure 4). RTs on the production task, however, decline and become less variable over time (Figure 4). The contrast between the two tasks suggests that changes in variability over time are at least skill specific (comprehension, production), and possibly task specific (matching, naming).

CONCLUSION

Previous work on response variability only analyzed variability for known words already represented in subjects' lexicons. Ours is a first attempt to use the CV when subjects start with zero familiarity and progress to establishing a mental representation. We examined the emergence of new representations at the levels of surface word forms and form-meaning mappings. In both cases, increased variability accompanied declines in mean RTs. These results complement the findings on response variability declines during automatization of existing knowledge, validating the CV as a measure sensitive enough to capture movement along the continuum from first forming a representation to accessing it automatically after practice.

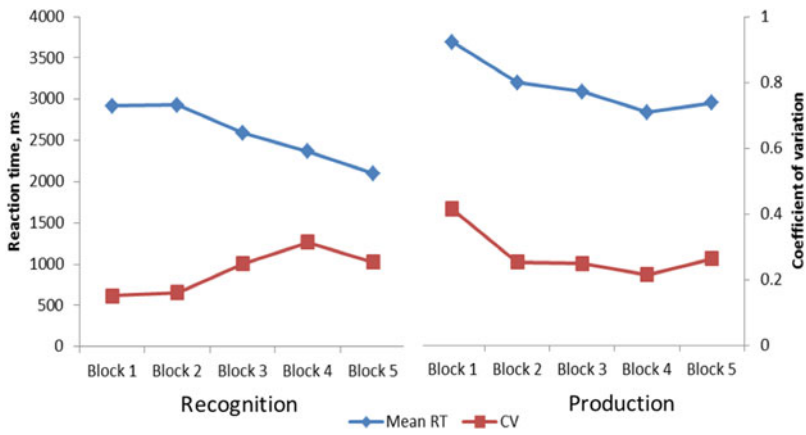


FIGURE 4. Mean RT and variability changes in Bartolotti & Marian (2014) on the recognition and production tasks.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

Our results paint a picture more complex than a steady decline in RTs accompanied by increased automatization or speed-up: Despite decreases in raw RTs, there is evidence of processing becoming more effortful due to subtle variations in task demands (the presence of mismatched pairs) and the growing familiarity of nonword stimuli. Thus, rather than simply automatizing or speeding up no-responses to nonwords, a conflict seems to occur between the practiced response becoming faster over time and the increasing familiarity of nonword stimuli making them more "wordlike" and difficult to reject. However, even in experiments where the required response did not involve a conflict, such as in Brown and Gaskell's study (2014), which required responses to novel words' lexical neighbors and not to the novel words, the observed pattern of CV changes is remarkably similar to ours. Similarly, in the case of Bartolotti and Marian (2014), the task involved matching novel words to pictures, which also does not pit lexical status (word, nonword) and familiarity against each other. Therefore, we conclude that increased variability accompanying speeded-up RTs is not simply an artifact of any conflict between the required response and the changing lexical status of novel stimuli that our experiment could have introduced. As demonstrated by our reanalysis of the Brown and Gaskell (2014) data, in addition to effort or automaticity, variability may capture growing *differentiation* within a stimulus set, which would be characteristic of knowledge restructuring and would not yet discernible in mean RTs alone.

Analyses of variability employing the CV can meaningfully augment mean RT analyses and be sensitive to changes in the familiarity and status of stimuli after short exposures. Future research can examine our proposals in more detail: One of the limitations of the present research is that we did not include an independent measure of processing speed on a task such as perceptual detection, as was done, for example, by Segalowitz and Segalowitz (1993). At the same time, to some extent changes in perceptual (and other kinds of lower-level processing) are built into RT declines. Further research will be needed to clarify the contexts in which RTs and CVs may move in opposite directions: Whether, for instance, it is possible for RTs to show a net increase in situations involving knowledge disruption after learning. In the present case, we were operating under the assumption of RT declines (which were borne out by the data) and only examined CV changes. It is conceivable, however, that a task from the more complex group, involving a constellation of mental processes, may indeed be characterized by a net increase in RTs and possibly also CVs.

The pattern of findings in our study extends the interpretation of changes in variability initially proposed by Segalowitz and Segalowitz (1993) and is largely consistent with it. If changes in variability, as indexed by CVs, reflect the number and efficiency of the underlying mental processes, then its reduction may signal increased efficiency and the elimination of slower processes, as was originally proposed by Segalowitz and Segalowitz. The present results suggest that its increases would signify the opposite—namely, the establishment of novel representations in memory, which necessitates additional component processes.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S0272263117000043

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press

REFERENCES

Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics*, *29*, 1–19.

Altarriba, J., & Basnight-Brown, D. M. (2007). Methodological considerations in performing semantic- and translation-priming experiments across languages. *Behavior Research Methods*, *39*, 1–18.

Anderson, J. R. (1983). *The architecture of cognition*. Mahwah, NJ: Erlbaum.

Bartolotti, J., & Marian, V. (2014). Wordlikeness and novel word learning. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 146–151). Austin, TX: Cognitive Science Society.

Brown, H., & Gaskell, G. (2014). The time-course of talker-specificity and lexical competition effects during word learning. *Language, Cognition and Neuroscience*, *29*, 1163–1179. doi: 10.1080/23273798.2014.916409

Fukkink, R. G., Hulstijn, J., & Simis, A. (2005). Does training of second-language word recognition skills affect reading comprehension? An experimental study. *The Modern Language Journal*, *89*, 54–75.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments*, *35*, 116–124.

Henson, R. N., Eckstein, D., Waszak, F., Frings, C., & Horner, A. J. (2014). Stimulus-response bindings in priming. *Trends in Cognitive Sciences*, *18*, 376–384.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*, 555–582. doi:10.1017/S0142716409990014

Hultsch, D. F., MacDonald, S. W., Hunter, M. A., Levy-Bencheton, J., & Strauss, E. (2000). Intraindividual variability in cognitive performance in older adults: Comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology*, *14*, 588.

Leung, C. Y., & Yoshikawa, L. (2015). What does the coefficient of variation computed from eye-movement measures tell us about automatization in L2 sentence processing? *Paper presented at the Annual Meeting of the American Association for Applied Linguistics*, Toronto, Canada.

Lim, H., & Godfroid, A. (2014). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*, 1247–1282. doi:10.1017/S0142716414000137

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Medler, D. A., & Binder, J. R. (2005). MCWord: An on-line orthographic database of the English language. Retrieved from http://www.neuro.mcw.edu/mcword/.

Meisel, J. M., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition*, *3*, 109–135. doi:10.1017/S0272263100004137

Rastatter, M., & Blair, B. (1984). Developmental influences on speech-motor equivalence: Some implications for articulatory disordered children. *Perceptual and Motor Skills*, *59*, 995–998.

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed SLA. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *49*, 295–319.

Segalowitz, N. S., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at-home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173–199.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*, 369–385. doi:10.1017/S0142716400010845

Segalowitz, N. S., Watson, V., & Segalowitz, S. (1995). Vocabulary skill: Single case assessment of automaticity of word recognition in a timed lexical decision task. *Second Language Research*, *11*, 121–136.

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*, 53–67.

https://doi.org/10.1017/S0272263117000043 Published online by Cambridge University Press