CrossMark

Applied Psycholinguistics **36** (2015), 1247–1282
doi:10.1017/S0142716414000137

Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study

HYOJUNG LIM and ALINE GODFROID
Michigan State University

Received: July 28, 2013 Accepted for publication: February 9, 2014

ADDRESS FOR CORRESPONDENCE
Hyojung Lim, Second Language Studies, Michigan State University, B270 Wells Hall, East Lansing, MI 48824. E-mail: hyojung@msu.edu; or Aline Godfroid, Department of Linguistics and Languages, Michigan State University, B253 Wells Hall, East Lansing, MI 48824. E-mail: godfroid@msu.edu

ABSTRACT
This study investigates the automatization of sentence processing using the coefficient of variation (CV), a measure of intraindividual processing stability (Segalowitz & Segalowitz, 1993). A smaller CV (i.e., standard deviation to reaction time [RT] ratio) and a positive CV-RT correlation are taken to index increased automatization. Hulstijn, Van Gelderen, and Schoonen (2009) were the first to try to validate the use of CV at the sentence level; however, they did not find any evidence for automatization in their CV analyses. Forty Korean English as a second language students (20 intermediate, 20 advanced) and 20 native speakers performed three speeded tasks in English: a semantic classification task, a sentence verification task, and a sentence construction task. The results revealed that, consistent with findings from previous word recognition studies, the CV in the sentence-level tasks decreased as participants' proficiency level increased. Although the CV-RT correlation in the sentence verification task was not always significant, no counterevidence against Segalowitz and Segalowitz' (1993) hypothesis was found. The sentence construction task discriminated better between groups than the sentence verification task. We argue that the CV may be a valid measure of automatization at the sentence level, provided the tasks used target lower-level processes such as word recognition, parsing, and semantic proposition formation.

Essential to fluent language use is the automatic execution of various component subskills (McLaughlin, 1987). However, skills such as language are cognitively complex. They consist of a hierarchically organized set of subskills, in which automatic and attention-controlled processes can coexist at different levels (McLaughlin, 1987; Segalowitz & Segalowitz, 1993). Skill acquisition theory thus views increases in language proficiency, and especially in language fluency,

© Cambridge University Press 2014 0142-7164/14 $15.00

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

as the transition from more controlled to more automatic processes. Automaticity is a cognitive basis of second language (L2) fluency (Segalowitz, 2010); it plays "a significant role in distinguishing fluent from nonfluent abilities" (Segalowitz & Hulstijn, 2005, p. 373). Thus, this study will investigate the development of L2 automaticity as evidenced in native and nonnative speakers' fluent performance of two different sentence processing tasks.

The abstract notion of automaticity has attracted the interest of many second language acquisition (SLA) scholars because a high degree of L2 fluency is the ultimate goal for L2 learners. In different SLA theories, the transition from explicit to implicit knowledge (Bialystok, 1978), from declarative to procedural knowledge (Anderson, 1983), and from nonintegrated to integrated knowledge (Jiang, 2007) all seem to depend on the extent to which L2 linguistic processes become automatized. For instance, implicit knowledge is often described as knowledge accessible by means of automatic processing (e.g., Dörnyei, 2009; R. Ellis, 2005), and similarly, procedural knowledge is defined as a resource available for unconscious language use (e.g., Anderson, 1983). Integrated knowledge is also considered synonymous with automatic competence in adult SLA (e.g., Jiang, 2007). Though different scholars have used different terms, the common element that discriminates types of knowledge seems to be the presence or absence of automatic processing in learners' use of the language.

In view of the prominent role of automaticity in SLA, it is surprising how little literature on learners' automatic processing in the L2 there is. Many important questions have remained unanswered as to how L2 automaticity develops over the course of language learning, how long it takes for L2 learners to reach the fully automatized phase, how the development of automaticity varies depending on linguistic features, or how automaticity can be validly measured. The present study is particularly interested in a methodological issue: how to assess the development of automaticity in L2 sentence processing. Although automaticity research on L2 word recognition has steadily accumulated since the 1980s, only a few studies have attempted to measure automatic L2 sentence processing (DeKeyser, 1997; Hulstijn, Van Gelderen, & Schoonen, 2009; Jiang, 2007; Robinson, 1997; Rodgers, 2011; Serafini, 2013). The current study continues and expands the latter line of inquiry by extending a measure typically used to assess automatization of L2 word recognition to the sentence processing level with written stimuli.

BACKGROUND

Automaticity in SLA

In cognitive science and psychology, automatic processing has been defined traditionally as relatively rapid, effortless, unconscious, attention-free, ballistic (unstoppable), uncontrollable, stereotypic, and parallel processing. The study of SLA explores the same properties of automaticity, yet focuses on linguistic processes: how L2 users map forms to meanings during lexical retrieval, how they acquire and apply grammar rules, and how they become skilled at communicating in the L2. Given the complex, hierarchical nature of all of these activities, some linguistic subprocesses may conceivably be more amenable to automatization than others.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

In reading studies, for instance, the importance of efficient (i.e., automatic) word recognition and syntactic parsing has been accentuated for good reading comprehension (Breznitz, 2006; Grabe & Stoller, 2011; Koda, 2005, 2007; Van Gelderen, Schoonen, Stoel, De Glopper & Hulstijn, 2007). Segalowitz and Hulstijn (2005) posit that "it is especially at the intermediate levels of syntactic, morphological, and phonological encoding/decoding, as well as at the lower levels of articulation and perception of acoustic or orthographic signals, that component processes can become automatic to a large extent" (p. 385). Automatization studies and research on the four communicative skills (i.e., reading, writing, listening, speaking) can thus cross-fertilize each other, as a good understanding of automatization requires a theoretical model of the skill being automatized. However, few researchers have attempted to bridge this disciplinary divide.

One approach that holds potential in this regard is Segalowitz and Segalowitz' (1993) coefficient of variation (CV). CV is a measure of processing stability (to be defined below) that can in principle be applied to any type of human or nonhuman behavior. It is important to note that the CV is also compatible with more prevalent approaches to automatization in SLA, such as Anderson's adaptive control of thought (e.g., Anderson, 1976, 1983, 1993, 2007; DeKeyser, 1997) and Logan's instance theory (Logan, 1988, 2002; Robinson, 1997). These theories emphasize that (a) automatization necessitates a qualitative change either from declarative to procedural knowledge (Anderson) or from algorithm-based performance to memory-based performance (Logan), (b) automatization results from intensive and extensive practice, (c) automatization requires strong linguistic knowledge, and (d) automatization follows the power law of practice[1] (Anderson, 1992, p. 172; Logan, 1988). Although characteristics (b) through (d) seem readily applicable to the study of communicative skills, it is less clear what a qualitative change in processing would entail in this area.

According to Segalowitz (2010; Segalowitz & Segalowitz, 1993) the qualitative change that characterizes automaticity is restructuring: changing and/or organizing the underlying processing routines of a task differently (McLaughlin, 1987). Segalowitz (2010) subsumes the many characteristics of skilled processing under three key properties: processing speed, processing stability, and processing flexibility. It is intuitively known that automatic processing is fast processing. However, although processing speed is a prominent feature of automatized processing, Segalowitz (2010; Segalowitz & Segalowitz, 1993) emphasizes that speed alone is not sufficient for genuine automatization. In addition, automatization involves restructuring of the cognitive subcomponents of a task, which will improve processing stability and efficiency. "A process is automatic, even if one cannot specify the exact nature of that automaticity, as long as one can say that the process is functioning more efficiently in some meaningful way compared to how it would have functioned if it were non-automatic" (Segalowitz, 2010, p. 79).

Segalowitz and Segalowitz (1993) proposed the CV as a measure of processing stability. CV takes into account both processing stability (intraindividual standard deviation) and processing speed, or reaction time (RT), by computing the standard deviation to RT ratio. The resulting value reflects a person's performance stability corrected for his or her performance speed, with smaller CVs indicating higher levels of processing stability. Because the validity of the CV as a measure of

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

processing stability is of particular interest to the current study, the relevant literature will be discussed in the next two sections.

Finally, as far as grammatical processing is concerned, Segalowitz (2010) argues that flexibility of attentional control is also essential to skilled performance. Unlike processing speed and stability, processing flexibility is an attention-based mechanism that complements and supports the automatic aspects of processing (Segalowitz, 2010). Fluency entails the need to switch one's attentional focus from one task to another quickly and accurately in response to the linguistic information encountered in the input. For example, conjunctions such as *because* or *despite* carry information about how the speaker wishes the recipient to construe the link between two events, as a causal or noncausal relation (Segalowitz, 2010). Skilled language users are flexible in adapting their attentional focus accordingly and in integrating this new formation into the meaning representation they are constructing in real time. This type of specifically linguistic cognition has been termed language-directed attention (not to be confounded with attention to language; Fauconnier, 1994; Talmy, 2008). It is one of the factors underpinning fluent sentence processing.

CV

Although the use of CV in SLA has not spread much beyond Segalowitz and his collaborators, it is a common measure in the behavioral sciences to assess performance variability by virtue of its natural immunity to different units of measurement or widely different means (Bekoff, 1997; Carlton, Robertson, Carlton, & Newell, 1985; Dawkins & Dawkins, 1973). In particular, some psychologists show a strong preference for CV as a measure of intraindividual variability, pointing out that unlike the standard deviation, the CV is rarely affected by repeated testing or practice (Flehming, Steinborn, Langner, Scholz, & Westhoff, 2007). CV applications can be found in a wide range of fields of study: from measuring cardiac autonomic function (e.g., Toichi, Suguiura, Murai, & Sengoku, 1997) and predicting risk sensitivity of humans and other animals (e.g., Weber, Shafir, & Blais, 2004) to dissociating between speed and regularity in tapping behaviors (e.g., McManus, Kemp, & Grant, 1986).

Segalowitz and Segalowitz (1993) were the first to advocate the use of CV as a measure of L2 automatization. Central to Segalowitz and Segalowitz' proposal is that automatization involves a qualitative change in cognitive processing; that is, a restructuring (deletion and/or suppletion) of processing components above and beyond simple speedup. They argued that only this scenario results in a disproportionate decrease in standard deviation and concomitant CV reduction.

Let us take L2 visual word recognition as an example. Suppose that word recognition training helps to decrease the RTs associated with each subcomponent of the word recognition process (e.g., orthographic processing, phonological processing, translation into the native language, first language [L1]-mediated concept retrieval) and thus reduces the total RT for the word recognition behavior as a whole. Given that the default is for standard deviations and RTs to be linearly related (Anderson, 1992; Logan, 1988; Wagenmakers & Brown, 2007), standard deviations will decrease as RTs decrease. Because CV equals standard deviation/RT,

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

this means that CV will remain unchanged if standard deviations drop in proportion to RTs for the word recognition behavior as a whole. In Segalowitz and Segalowitz' view, no change in CV indicates speedup: each subcomponent has become accelerated but nonetheless functions in the same manner, just like when somebody becomes faster at cooking after having done it many times but still goes through all the same steps he or she did originally.

Alternatively, if extended practice leads to the elimination or dramatic reduction of subcomponents that involve much attentional effort (e.g., L1 translation, metalinguistic analysis), the proportional reduction in standard deviations exceeds the proportional reduction in RTs, and CVs decrease as a result. Slower, effortful processes tend to be more variable (i.e., have larger standard deviations); therefore, removing them will have a disproportionate effect on the standard deviation of the total RTs (for a numerical example, see Segalowitz & Segalowitz, 1993, their table 1). Segalowitz and Segalowitz assert that a CV reduction evidences the development of automaticity. Finally, as CV, RT, and the standard deviation are all assumed to decrease over the course of automatization, a positive correlation between CV and RT is a corollary of a simultaneous reduction in CV and RT.

The CV analysis has proven a valuable tool in SLA studies with diverse populations and for different research purposes (for a review and critique, see Hulstijn et al., 2009). However, all of the acquisition-related CV research until then focused on the processing of single words. Paradis (2004, 2009) questioned the findings of such studies on grounds of ecological validity and the special status of vocabulary knowledge within language: "Single words are the least likely candidates for investigating 'language' representation, given that what makes language most specific as a cognitive function, that is, the language system (phonology, morphology, syntax), is supported by procedural memory, whereas isolated words are supported by declarative memory" (Paradis, 2004, p. 173). Hulstijn et al. (2009) were the first to extend the scope of CV investigations beyond the word level. They investigated the automatization of word- and sentence-level processing, as evidenced by CV, based on a massive data set of 397 Dutch high school learners of English who were followed over 2 years.

Hulstijn, Van Gelderen, and Schoonen's experiment 1

Hulstijn and colleagues (2009) carried out a 2-year longitudinal study in a number of high schools in The Netherlands. The students (about 200 at each time of measurement) performed four computerized RT tasks in L1 Dutch and in L2 English in Grades 8, 9, and 10. The tasks consisted of a lexical decision and retrieval task and a sentence verification and production task. As this was a longitudinal study, time was used as an independent index of automaticity; that is, it was assumed that learners' fluency in the L2 would develop as they advanced to a higher grade and that, consistent with Segalowitz and Segalowitz' (1993) hypothesis, this would be reflected in a within-subject decrease in CV over the 2 years. Contrary to this prediction, however, Hulstijn et al. (2009) found that "in all eight tasks the CV remained more or less constant"[2] (p. 567). Similarly, the CV-RT correlation was not always in the expected direction (i.e., some correlations were negative) and not all positive correlations were significant.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Compared to previous CV studies, Hulstijn and colleagues (2009) submitted their RT data to a more thorough cleaning procedure before statistical analysis. It is important to discuss the different options for data cleaning in CV research in some detail because the choices one makes can affect the results (for an illustration, see Hulstijn et al., 2009, exp. 2).

The purpose of data cleaning is to minimize the noise inherent in RT data (e.g., Jiang, 2012). First, in many RT tasks such as lexical decision, it is standard practice to analyze correct trials only because knowledge retrieval on incorrect trials either failed or was nontarget-like. However, if accuracy rates do not approach 100% this may mean discarding a significant portion of the data and hence important information loss. In the analysis reported for the longitudinal data set (exp. 1), Hulstijn et al. (2009) circumvented this problem by coding the incorrect trials as missing values and subsequently estimating them with the expectation-maximization algorithm.[3] Second, RT experiments also require dealing with extremely slow or extremely fast responses, as "such RTs cannot be considered to be a valid measure of the processes under investigation" (Hulstijn et al., 2009, p. 578). A common practice is to "pull in" outliers; that is, to replace them by a less extreme value such as the item mean $\pm$ 2, 2.5, or 3 *SD*. This approach, however, seems undesirable for CV analyses because "the empirical evidence of investigations on CV reduction is formed by the standard deviation, which, first and foremost, *is* an index of variability. By cutting off the tails of the distribution, the researcher artificially reduces the variability in the raw data" (Hulstijn et al., 2009, p. 578, original emphasis). The solution that these authors propose is again to treat the outliers as missing values and have the algorithm estimate them. Third, and perhaps most crucially, Hulstijn et al. argued that only items with high hit rates and participants with high accuracy rates should be included for any given task "because the speed tests [are] intended to measure speed of processing and should not be confounded with accuracy of linguistic knowledge" (p. 566). Herein lies the tricky part, because inclusion criteria can be made more or less stringent. By way of illustration, Hulstijn et al. (2009, exp. 2) computed CVs on a more restricted data set that was prepared in six different ways. Only if participants who performed at least at 95% accuracy were included did the CV decrease significantly, but in one analysis, this amounted to eliminating 40% of the data! Although some data cleaning may be indispensable, treatment should be used sparingly so as to minimize the amount of information loss and ensure the data's validity.

For their first experiment, the longitudinal study, Hulstijn and colleagues (2009) generally did not find a decrease of CV over time. The authors attributed this null result to a potential confounding of knowledge accumulation and skill acquisition in CV measurement. The argument is that with increasing exposure, L2 learners not only consolidate existing linguistic knowledge but also acquire new, declarative knowledge (e.g., new words). While the former results in faster RTs, the latter typically does not, because novel information tends to be accessed slowly and effortfully (McLaughlin, 1987). Hence, "RT tests . . . inherently suffer from confounding measuring improvement in response efficiency with measuring improvement in declarative knowledge" (Hulstijn et al., 2009, p. 579). Given that high school students are still learning both L1 and L2 vocabulary and grammar

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

knowledge, this problem may have been particularly acute in the Hulstijn et al. study (see p. 580, note 1).

Because of the speed–accuracy trade-off, Hulstijn et al. questioned the reliability of the CV index as a tool to distinguish between speedup and automatization. Although we agree that the mathematical and conceptual bases of the CV deserve close scrutiny, we believe that it is premature to discredit the measure based on the observed null findings. A number of alternative accounts of the findings come to mind. First, we cannot rule out the possibility that Dutch secondary school students failed to automatize their English processing skills in 2 consecutive years. When the students completed the tests for the first time in Grade 8, most had "just started L2 reading" and "their L2 vocabulary and grammar knowledge as well as their efficiency in processing L2 words and sentences [were] at a low level, relative to their L1 skills" (Van Gelderen et al., 2007, p. 478). Because automatization is known to take place relatively late in language development (Anderson, 1983; Segalowitz & Hulstijn, 2005; Wickens, 1992), more advanced students, for instance from the final years of high school or postsecondary education, might be more suitable study subjects. Results from independent proficiency tests may also be helpful to interpret an apparent lack of automatization. Second, issues of positive transfer of orthographic and syntactic processing from L1 Dutch to L2 English, which are two closely related languages, may have obscured potential gains in automaticity in other linguistic domains. Third, it is possible that automatization did occur but that some properties of Hulstijn et al.'s stimuli made it difficult to detect through a CV analysis.

For all of these reasons, we felt that the time for a conceptual replication of Hulstijn et al. (2009) was ripe. We acknowledge at the outset that this study is not as ambitious in scope as the original study, yet we believe that because it addresses the issues pointed out above, it may still make a valuable contribution to the nascent CV literature on L2 sentence processing. Using relatively advanced English as a second language (ESL) learners who have a typologically different L1 (Korean), we sought to answer the following questions:

1. Does the CV index the degree of automatization in L2 sentence processing?
   a. Does CV decrease as the automaticity of sentence processing develops?
   b. Does the correlation between CV and RT become statistically positive as the automaticity of sentence processing develops?
2. Does the usefulness of CV as an index of the degree of automatization differ across different tasks (e.g., sentence verification task vs. sentence production/construction task)?

Two important assumptions underlie this study: first, sentence processing, which entails grammar processing, is amenable at least partly to automatization. If grammar use is truly an automatic behavior like lexical access (e.g., DeKeyser; 1997; Robinson, 1997; Rodgers, 2011; Serafini, 2013), the principles of CV should work equally well for sentence processing in the L2. Second, L2 automaticity increases with proficiency level. Just as a longitudinal study assumes the gradual development of fluency over the course of time, this cross-sectional study assumes

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

that advanced learners have a higher level of fluency than intermediate learners and that fluency is a behavioral index of automatization.

METHOD

Participants

Forty Korean ESL learners and 20 native speakers of English were recruited in Michigan and New York. Most of the participants were either graduate or undergraduate students. A small number of participants were enrolled in an English for Academic Purposes program at a university, in which they were readying themselves to attend regular university courses in the following semester.

Participants' proficiency level was assessed through DIALANG, an online diagnostic language test established by Lancaster University. Due to time restrictions, only vocabulary size and grammar knowledge were assessed, amounting to an overall test duration of 10–15 min. DIALANG follows the common European framework (Council of Europe, 2001), which divides learners' proficiency into three broad levels: the A level for basic users, the B level for independent users, and the C level for proficient users. Each level further branches into two sublevels, such as A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced), and C2 (proficient). The present study classified ESL learners as advanced if they scored at the C level and intermediate if they scored at the B level. By requiring at least intermediary-level mastery of English, we hoped to minimize the effect of learners' declarative knowledge on processing skills (see above). Table 1 summarizes the demographic information of the participants.

Materials

Three RT experiments were administered to measure participants' L2 word recognition and L2 sentence processing skills: a semantic classification task, a sentence construction task,[4] and a sentence verification task. For measures of sentence processing efficiency to be maximally informative, all participants should ideally have equally automatized lexical knowledge, so any differences in sentence processing skills can be attributed to the sentence level of language processing. Therefore, the purpose of the semantic classification task was to assess any preexisting differences between participants' automatization of lexical knowledge.

All tasks were conducted in English only. DMDX was used for the semantic classification and the sentence verification tasks, and the sentence construction task was run on Linger. RTs and accuracy rates were recorded for all three tasks. The original sentences from Hulstijn et al. (2009) were made available to us by the authors and were revised to be suitable for Korean ESL speakers studying at US universities. Major modifications are further described in the next three subsections. The experimental stimuli used in each task are given in Tables A.1, A.2, and A.3 in Appendix A.

Semantic classification task. Participants were asked to decide as quickly and as accurately as possible if a word on the computer screen referred to a living being (e.g., boy) or a nonliving artifact (e.g., book). Such animacy judgment tasks have

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table 1. *Description of participants*

	Intermediate Learners ($n = 20$)	Advanced Learners ($n = 20$)	Native Speakers ($n = 20$)
DIALANG scores	B1 ($n = 7$) B2 ($n = 13$)	C1 ($n = 17$) C2 ($n = 3$)	NA
Gender	Female ($n = 5$) Male ($n = 15$)	Female ($n = 12$) Male ($n = 8$)	Female ($n = 9$) Male ($n = 11$)
Age	Range = 18–34 Median = 24.5	Range = 20–46 Median = 30	Range = 20–37 Median = 23
Education	EAP ($n = 2$) BA ($n = 13$) MA/PhD ($n = 5$)	BA ($n = 2$) MA/PhD ($n = 18$)	BA ($n = 10$) MA/PhD ($n = 10$)
Onset of learning	Mean = 10.80 Median = 11	Mean = 9.76 Median = 11	
Onset of arrival	Mean = 20 Median = 19.5	Mean = 18.76 Median = 20	
LOR (year)	Median = 2.80	Median = 3.84	

Note: Onset of learning refers to the age at which learners first started learning English. Onset of arrival is the age at which learners first arrived in the United States. EAP, English for academic purposes; LOR, length of residence in the United States.

been intended to replace traditional lexical decision tasks, in which participants are asked to discriminate between real words and nonwords (Phillips, Segalowitz, O'Brien, & Yamasaki, 2004; Segalowitz & de Almeida, 2002; Segalowitz & Frenkiel-Fishman, 2005). Segalowitz and his associates posit that the animacy judgment task is a more authentic way of measuring learners' word recognition skills than a lexical decision task because it invokes relatively strong semantic processing.

All word stimuli were extracted from the subsequent sentence-level tasks in order to expose the participants to the words before the sentence processing tasks. By doing so, we hoped to make sure that individual differences in sentence processing efficiency did not stem from differences in either participants' lexical processing or their word knowledge. Word frequencies for all the stimuli, based on the trimmed version of the CELEX database MC Word: An Orthographic WordForm Database (Medler & Binder, 2005), are reported in Table A.1 in Appendix A. The task began with six practice items, after which 46 actual trials followed.

Sentence construction task. In this task, the beginning part of a sentence was given on the first screen (e.g., *After some time* . . .), and two possible options followed on the next screen (e.g., A. *works*, B. *she*). Participants were asked to read the sentence beginnings in a self-paced fashion but then to quickly choose the word(s) that best continued the phrase that they just read. Participants were told that the given options would not necessarily complete the sentence but that one of them was a good continuation.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

The sentence construction task required participants to process a sentence fragment and to select the continuation that best fit into the syntactic structure they created. Unlike in Hulstijn et al. (2009), we kept the stimuli (both the stem and the options) as short as possible; sentence beginnings ranged from one to three words and each answer option constituted a single word or two words, either a function word or a content word. Because the stimuli were so simple, it limited the amount of higher level semantic analysis that participants could bring to the task, focusing instead on participants' parsing skills. The task started with six warm-up trials, after which 40 real items followed.

Sentence verification task. This is a sentence plausibility judgment task, an extended form of the semantic classification task at the word level. Participants were asked to read a sentence and decide quickly whether it made sense. Unlike in Hulstijn et al. (2009), sentence length was controlled for, with each sentence consisting of 7–10 words. The content of each sentence was also revised to fit the geographical location of the current participants. Some sentences contained semantic errors instantiated at the lexical level as in *Most theaters in the US have only one chair*, and others entailed structural violations at the phrasal level (e.g., *My uncle made me for a snowman*) or at the clausal level (e.g., *He went inside, though it started to rain*).

Current event-related potentials (ERP) studies justify incorporating structural violations in the sentence verification task for the purpose of examining automatic L2 sentence processing. During syntactic processing, phrase structure building and syntactic relational processing are known to take place in a relatively automatic manner (Gunter, Hahne, & Mauth, 2004; Hasting & Kotz, 2008). Thus, those who can process a sentence structure automatically should be able to verify the implausibility of the sentence quickly. Furthermore, varying the focus of the sentence verification task seems appropriate for tapping into learners' automatic processing skills. Recall that the third attribute of skilled processing (Segalowitz, 2010) was the flexibility with which learners can shift their focus of attention while dealing with a variety of linguistic input. In this task, some sentences asked participants to utilize their factual knowledge (e.g., *Most theaters in the US have only one chair*), while others required them to invoke grammatical knowledge (e.g., *My mother was cooked by a cake*). Learners who possess automatic processing skills are expected to adjust their attentional focus quickly to the adequate level of representation, as directed by the linguistic input, and thus react equally fast to different sorts of sentences. Participants were instructed to focus on meaning, rather than grammar, during reading. The task began with 6 warm-up trials, after which 60 real trials followed. Thirty-four sentences were nonsensible and 26 were sensible.

Procedure

All participants visited the language lab individually. First, the researcher explained the entire procedure to a participant and collected a consent form. The three speeded tasks described above were administered in the following order: (a) the semantic classification task, (b) the sentence construction task, and (c) the sentence verification task. Upon completing the main tasks, participants took the English diagnostic language test, DIALANG. Finally, they filled out a language

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

background questionnaire that included questions about their educational background, age of first arrival in an English-speaking country, length of residence in the United States, and the like (see Appendix B). Participants were all volunteers and were paid $10 for their efforts.

Data analysis

Different statistical tests were used depending on whether the data met the normality assumption and the homogeneity of variances assumption. In the parametric option, we submitted the RTs, standard deviations, and CVs to one-way analyses of variance, followed by Tukey post hoc tests (equal variances) or Games–Howell tests (nonequal variances). The nonparametric solution consisted of a Kruskal–Wallis test followed by Mann–Whitney *U* tests. For the correlation analyses, we used Pearson's correlation (parametric) or Spearman's rho (nonparametric). It is important that all RT data were analyzed twice: once without data cleaning and the second time with data cleaning. The significance level α was set at .05 for all analyses.

For the sentence verification task, we conducted two analyses at each step. For the first analysis, we included all the correct responses, as we did for the other two tasks. This means that we considered both "hits" (sensible sentences judged sensible) and "correct rejections" (nonsensible sentences judged nonsensible) together. However, because automatic processing may be interrupted when individuals encounter something unexpected or problematic (Baars & Franklin, 2003; N. C. Ellis, 2005), it could be argued that the correct rejections do not reflect automatic processing but a (correct) reversion to attention-controlled processing. Therefore, we also looked at the hits separately. No such division between yes and no responses was necessary for the semantic classification or sentence construction tasks given the somewhat different nature of these tasks.

RESULTS

Raw data analysis: Clearest evidence of automatization in sentence construction task

For the first data analysis, we included only correct responses but did not do any additional data cleaning. We found that RTs, standard deviations, and CVs decreased in all tasks as English proficiency level increased (see Table 2). In the word-level semantic classification task, however, reduction in CVs was not statistically significant, $F\ (2,\ 57) = 2.16$, $p = .13$, $\eta^2 = 0.207$, meaning that we did not find any between-group differences in lexical processing efficiency. Although the decrease in RTs and standard deviations was statistically significant, such differences can be interpreted as reflecting differences in processing speed rather than degree of automaticity.

In both the sentence construction and verification tasks, RTs and standard deviations decreased significantly. The decrease in CVs in the sentence construction task was significant, $F\ (2,\ 57) = 18.66$, $p < .001$, $\eta^2 = 0.396$. Specifically, Games–Howell post hoc tests indicated that the CV of the native speaker group was statistically smaller than that of the advanced group ($p < .001$, $d = 0.93$) and

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table 2. *Statistical results for the three speeded measures before data cleaning*

Variable	Mean RT (*SD*)	Mean Stand. Dev. (*SD*)	Mean CV (*SD*)	Accu. Rate	*r* (CV, RT) For All	*r* (CV, RT) For Group
Semantic Classification Task, 46 Items						
Intermediate ($n = 20$)	815 (149)	242 (105)	0.29 (0.08)	0.98		.819*** ($p < .001$)
Advanced ($n = 20$)	752 (150)	190 (85)	0.24 (0.06)	0.98	.672*** ($p < .001$)	.724*** ($p < .001$)
Native ($n = 20$)	694 (148)	176 (67)	0.25 (0.06)	0.99		.407 ($p = .075$)
F (2, 57)	3.30	3.16	2.16			
p	.044	.050	.125			
η^2	0.104	0.491	0.207			
Sentence Construction Task, 40 Items						
Intermediate ($n = 20$)	2207[a,b] (673)	1254[a,b] (578)	0.56[a,b] (0.15)	0.92		.189 ($p = .287$)
Advanced ($n = 20$)	1572[a,c] (477)	707[a,c] (435)	0.42[a,c] (0.14)	0.96	.751*** ($p < .001$)	.713*** ($p = .001$)
Native ($n = 20$)	1231[b,c] (259)	390[b,c] (166)	0.31[b,c] (0.09)	0.96		.532** ($p = .016$)
	$\chi^2 = 27.63$	$\chi^2 = 30.35$	F (2, 57) = 18.66			
p	<.001	<.001	<.001			
η^2	0.468	0.514	0.396			

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table 2 (*cont.*)

Variable	Mean RT (*SD*)	Mean Stand. Dev. (*SD*)	Mean CV (*SD*)	Accu. Rate	*r* (CV, RT) For All	*r* (CV, RT) For Group
Sentence Verification Task, 60 Items (Both Plausible and Implausible Items)						
Intermediate ($n = 20$)	4595[a,b] (1321)	2011[a,b] (514)	0.46 (0.11)	0.90		−.602** ($p = .005$)
Advanced ($n = 20$)	3561[a,c] (930)	1509[a,c] (593)	0.43 (0.09)	0.96	.066 ($p = .619$)	.383 ($p = .095$)
Native ($n = 20$)	2653[b,c] (396)	1054[b,c] (237)	0.40 (0.08)	0.96		.015 ($p = .950$)
	$\chi^2 = 26.06$	$\chi^2 = 27.58$	$F(2, 57) = 1.71$			
p	<.001	<.001	.190			
η^2	0.442	0.467	0.057			
Sentence Verification Task, 25items (Plausible Items Only)						
Intermediate ($n = 20$)	4433[a,b] (1386)	1897[a,b] (619)	0.44 (0.11)	0.89		−.338 ($p = .146$)
Advanced ($n = 20$)	3449[a,c] (953)	1475[a,c] (588)	0.42 (0.10)	0.97	.125 ($p = .342$)	.096 ($p = .688$)
Native ($n = 20$)	2596[b,c] (431)	1102[b,c] (336)	0.42 (0.11)	0.97		−.241 ($p = .306$)
	$\chi^2 = 21.01$	$\chi^2 = 16.82$	$F(2, 57) = 0.155$			
p	<.001	<.001	.859			
η^2	0.356	0.285	0.005			

Note: The exact p values are given below each statistic. RT, reaction time; CV, coefficient of variation.
[a]A significant difference from the native speakers.
[b]A significant difference from the advanced group.
[c]A significant difference from the intermediate-level group.
*$p = .05$. **$p = .01$. ***$p = .001$.

the CV of the advanced group was statistically smaller than that of the intermediate group ($p = .004$, $d = 0.96$). The decrease in CVs in the sentence verification task was not significant: for both sensible and nonsensible items, F (2, 57) = 1.71, $p = .190$, $\eta^2 = 0.057$; for the sensible items only, F (2, 57) = 0.16, $p = .859$, $\eta^2 = 0.005$. Nonetheless, participants' CVs on the verification and construction tasks were positively related (with all items, $r = .343$, $p = .007$; with only sensible items in the sentence verification task, $r = .280$, $p = .030$).

In terms of the CV-RT correlation analyses, we found a positive and significant correlation between RTs and CVs for the semantic classification task ($r = .672$, $p < .001$) and the sentence construction task ($r = .751$, $p < .001$). The significant correlation for the lexical task is not very informative because it was not accompanied by a significant CV reduction, but the correlation for the sentence construction task is significant: it may indicate individual differences in L2 automaticity. Finally, the CV-RT correlation in the sentence verification task was not significant: for the entire items ($r = .066$, $p = .619$) and for the sensible items only ($r = .125$, $p = .342$).

The results of the first data analyses suggest that all participants were comparable at the lexical level, which was consistent with the authors' prediction but diverged in terms of sentence-level processing skills. An interesting finding was that the two sentence tasks yielded different outcomes. To reduce the potentially confounding effect of participants' linguistic knowledge on the CV computations (Hulstijn et al., 2009), we now submitted all RTs from all tasks to the data cleaning procedure described in Hulstijn et al. (2009, exp. 1). We then reanalyzed these tailored data using the same statistics.

Data cleaning

Whereas only correct responses were considered for the original analysis, we now estimated RTs for incorrect responses and outliers using multiple imputation and included these data as well. In addition, a number of low-performing items were removed from the analysis.

The data cleaning procedure consisted of four steps. First, RTs of more than 3 *SD* above the group mean for each item were considered as extremely slow and therefore defined as missing values. To set the fast extremes of nonnative speakers' RTs, the smallest RT that native speakers achieved was used as the point of reference. Then, nonnative speakers' RTs faster than the fastest RTs of native speakers were again defined as missing values. Second, to minimize the effect of linguistic knowledge on the processing measure, the items with high error rates were taken out. As Hulstijn et al. (2009) suggested, only the items with higher than 87.5% correct responses were included. This represents "a 75% correct score, with the chance level for guessing added" (Hulstijn et al., 2009, pp. 566–567). As a consequence, we excluded 16 sentence verification items out of 60 (2 plausible sentences, 14 implausible sentences) and 9 sentence construction items out of 40. The excluded items are listed in Appendix C. Third, in addition to these two steps, Hulstijn et al. treated the students who scored less than 62.5% correct answers on any speed test as missing for the whole test. In the current study, only one participant scored 52% correct on the plausible sentences in the sentence verification task (but above threshold on the task as a whole); he or she

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

was thus excluded from the second data analysis for this task. Fourth, all the erased RTs (of incorrect, or extremely slow or fast responses) were estimated by means of multiple imputation. Multiple imputation is the substitution of some value for a missing data point in multivariate analysis (Rubin, 1977).[5] We calculated the maximum and the minimum RTs for each individual item and defined these as the boundary points for the imputed values to fall in. In terms of the missing data patterns, 3.01% of the values were missing in the word recognition task, 4.96% in the sentence verification task (4.53% for the "yes" response items only), and 5.86% in the sentence construction task.

Cleaned data analysis mostly confirmed findings

The data cleaning procedure did not change the descriptive statistics substantially. All RT measures reduced slightly as a result of data cleaning as did their corresponding standard deviations (compare Table 2 and Table 3). As expected, the native speaker group recorded the smallest RTs, standard deviations, and CVs in all three tasks, whereas the intermediate group showed the largest RTs, standard deviations, and CVs (see Table 3).

A one-way analysis of variance indicated that all three measures for the semantic classification task decreased significantly as proficiency level increased ($p = .013$, $\eta^2 = 0.142$ for mean RT, $p = .004$, $\eta^2 = 0.174$ for mean standard deviation, and $p = .027$, $\eta^2 = 0.119$ for mean CV). The significant decrease for CV was new compared to the earlier analysis. Tukey post hoc tests revealed that the intermediate ESL group was significantly slower, more variable and less efficient than the native speaker group: the differences between the two groups were 123 ms for mean RT ($p = .009$, $d = 0.96$), 73 ms for mean standard deviation ($p = .011$, $d = 0.99$), and 0.05 for mean CV ($p = .046$, $d = 0.77$). The advanced ESL group did not differ from the intermediate learners or the native speakers on any measure. In terms of the CV-RT correlation, a medium-size, positive correlation was found in the entire group ($r = .549$, $p < .001$). Within groups, CVs of the intermediate and the advanced group were highly correlated with RTs (intermediate: $r = .772$, $p < .001$; advanced: $r = .619$, $p = .004$). In contrast, the CV-RT correlation in the native speaker group was virtually zero ($r = .031$, $p = .878$).

The results for the sentence construction task were consistent with the results before the data cleaning: the groups differed significantly from one another for all three measures. The native speakers showed the smallest RTs, standard deviations, and CVs, whereas the intermediate learners recorded the largest RTs, standard deviations, and CVs. In terms of the correlation analyses, the entire set of RTs were significantly correlated with the entire set of CVs ($r = .642$, $p < .001$). At the group level, the CV-RT correlation was positive and significant for the advanced learners ($r = .541$, $p = .001$) but neither for the intermediate learners ($r = .191$, $p = .419$) or the native speakers ($r = .049$, $p = .838$). By deleting nine items with high error rates, accuracy rose to 95% for the intermediate, 98% for the advanced, and 99% for the native group. The differences in accuracy between the three groups were not statistically significant, $F(2, 57) = 2.28$, $p = .112$.

Finally, general patterns for the sentence verification task did not change after the data cleaning: mean RT, standard deviation, and CV decreased as proficiency level increased. For mean RTs and standard deviations, all groups differed significantly

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table 3. *Statistical results for the speeded measures after data cleaning*

Variable	Mean RT (*SD*)	Mean Stand. Dev. (*SD*)	Mean CV (*SD*)	Accu. Rate	*r* (CV, RT) For All	*r* (CV, RT) For Group
Semantic Classification Task, 46 Items						
Intermediate ($n = 20$)	806[a]	233[a,b]	0.28[a]	0.98		.772***
	(132)	(93)	(0.07)			($p < .001$)
Advanced ($n = 20$)	743	178[c]	0.24	0.98	.549***	.619**
	(125)	(61)	(0.05)		($p < .001$)	($p = .004$)
Native ($n = 20$)	683[c]	160[c]	0.23[c]	0.99		.031
	(123)	(48)	(0.06)			($p = .878$)
F (2, 57)	4.70	5.95	3.83			
p	.013	.004	.027			
η^2	0.142	0.174	0.119			
Sentence Construction Task, 31 Items						
Intermediate ($n = 20$)	2016[a,b]	893[a,b]	0.44[a,b]	0.95		.191
	(440)	(307)	(0.10)			($p = .419$)
Advanced ($n = 20$)	1494[a,c]	523[a,c]	0.34[a,c]	0.98	.642***	.541*
	(307)	(186)	(0.07)		($p < .001$)	($p = .014$)
Native ($n = 20$)	1164[b,c]	304[b,c]	0.26[b,c]	0.99		.049
	(194)	(92)	(0.06)			($p = .838$)
	$\chi^2 = 33.58$	$\chi^2 = 38.24$	$\chi^2 = 29.00$			
p	<.001	<.001	<.001			
η^2	0.569	0.648	0.492			

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table 3 (*cont.*)

Variable	Mean RT (*SD*)	Mean Stand. Dev. (*SD*)	Mean CV (*SD*)	Accu. Rate	*r* (CV, RT) For All	*r* (CV, RT) For Group
Sentence Verification Task, 44 Items (Plausible And Implausible Items)						
Intermediate ($n = 20$)	4305[a,b] (1202)	1670[a,b] (442)	0.40[a] (0.07)	0.96		−.516* ($p = .020$)
Advanced ($n = 20$)	3269[a,c] (673)	1247[a,c] (404)	0.38[a] (0.07)	0.97	.203 ($p = .120$)	.361 ($p = .118$)
Native ($n = 20$)	2445[b,c] (358)	774[b,c] (158)	0.32[b,c] (0.06)	0.98		−.352 ($p = .128$)
	$\chi^2 = 30.97$	$\chi^2 = 36.84$	$F(2, 57) = 8.10$			
p	<.001	<.001	.001			
η^2	0.525	0.624	0.221			
Sentence Verification Task, 23 Items (Plausible Items Only)						
Intermediate ($n = 19$)	4105[a,b] (1122)	1862[a] (614)	0.39 (0.07)	0.93		−.156 ($p = .523$)
Advanced ($n = 20$)	3332[a,c] (804)	1475[a] (588)	0.40 (0.10)	0.96	−.066 ($p = .621$)	.061 ($p = .797$)
Native ($n = 20$)	2533[b,c] (415)	1102[b,c] (336)	0.40 (0.07)	0.97		−.113 ($p = .635$)
	$\chi^2 = 19.87$	$\chi^2 = 15.56$	$F(2, 57) = 0.09$			
p	<.001	<.001	.913			
η^2	0.343	0.268	0.003			

Note: The exact p values are given below each statistic. RT, reaction time; CV, coefficient of variation.
[a]A significant difference from the native speakers.
[b]A significant difference from the advanced group.
[c]A significant difference from the intermediate-level group.
*$p = .05$. **$p = .01$. ***$p = .001$.

from one another. It is interesting that the results of the CV analyses differed somewhat depending on whether only the plausible sentences or both the plausible and the implausible sentences were analyzed. For the plausible sentences only, CVs did not differ between groups ($p = .913$, $\eta^2 = 0.003$) and did not correlate with participants' CVs on the sentence construction task ($r = –.035$, $p = .795$). However, when plausible and implausible sentences were combined, CVs did correlate strongly with those for the sentence construction task ($r = .540$, $p < .001$). Nonnative speakers' CVs were also significantly larger than native speakers' ($p = .001$, $d = 1.23$ for intermediate vs. native, and $p = .013$, $d = 0.92$ for advanced vs. native), yet the CVs for the two ESL groups did not differ significantly ($p = .629$).

As regards the CV-RT correlation for the sentence verification task, correlations for the plausible sentences only were all close to zero (across the entire group: $r = –.066$, $p = .621$; native speakers: $r = –.113$, $p = .635$; advanced learners: $r = .061$, $p = .797$; intermediate learners: $r = –.156$, $p = .523$). When we included both plausible and implausible sentences, a negative CV-RT correlation was statistical in the intermediate group ($r = –.516$, $p = .020$) but not so among the native speakers ($r = –.352$, $p = .128$). CV-RT correlations for the advanced learners ($r = .361$, $p = .118$) and for the entire sample ($r = .203$, $p = .120$) were positive but nonsignificant. After deleting 2 plausible and 14 implausible items with high error rates, accuracy increased to 96% for the intermediate, 97% for the advanced, and 98% for the native group. As a result, group differences became insignificant, $F(2, 57) = 3.08$, $p = .054$. Accuracy differed significantly as a function of item plausibility among native speakers, $t(19) = –2.84$, $p = .010$, $d = 0.077$, but not in the ESL groups, intermediate: $t(19) = –1.70$, $p = .106$; advanced: $t(19) = –1.14$, $p = .267$. The native speakers misjudged the implausible items significantly more often than the plausible items, whereas this was not the case for the nonnative speakers.

Summary

The results of the current study supported Segalowitz and Segalowitz' (1993) hypothesis (CV decreased as speakers developed automatic processing skills), regardless of whether the data were cleaned (see Table 4). Data cleaning did not affect overall patterns of RTs, standard deviations, and CVs in any task, yet it boosted accuracy rates and caused some contrasts that had initially been nonsignificant to become significant, notably in the semantic classification and sentence verification tasks. Thus, only after data cleaning were the intermediate learners found to retrieve word meanings with more difficulty than the native speakers ($p = .009$ for RTs, $p = .011$ for standard deviations, $p = .046$ for CVs). No such difference was found for the advanced learners versus the native speakers. Similarly, after data cleaning, but not before, nonnative speakers appeared to be less efficient at sentence verification than native speakers, at least when both plausible and implausible sentences were considered together. Finally, the findings of the sentence construction task seemed to be the most robust, given that CV decreased with increasing proficiency level both before and after data cleaning.

Table 4 summarizes the important findings of the study and compares them with those of Hulstijn et al. (2009). Note that Hulstijn et al. did not report the

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table 4. *Comparisons of findings in Hulstijn et al. (2009) and the present study*

	Mean RT	Mean Stand. Dev.	Mean CV
Receptive Lexical Task (Lexical Decision or Semantic Classification)			
Hulstijn et al. (2009)	Grade 8 > Grade 9 > Grade 10 ($p < .001$)	Grade 8 > Grade 9 > Grade 10 ($p < .001$)	Grade 9 > Grade 8 > Grade 10 (*ns*)
This study			
Raw data	Intermed. > advanced > NS ($p = .044$)	Intermed. > advanced > NS ($p = .050$)	Intermed. > advanced > NS ($p = .125$)
Cleaned data	Intermed. > advanced > NS ($p = .013$)	Intermed. > advanced > NS ($p = .004$)	Intermed. > advanced > NS ($p = .027$)
Sentence Production/Construction Task			
Hulstijn et al. (2009)	Grade 8 > Grade 9 > Grade 10 ($p < .001$)	Grade 8 > Grade 9 > Grade 10 ($p < .001$)	Grade 8 < Grade 9 < Grade 10 (*ns*)
This study			
Raw data	Intermed. » advanced > NS ($p < .001$)	Intermed. » advanced > NS ($p < .001$)	Intermed. » advanced » NS ($p < .001$)
Cleaned data	Intermed. » advanced » NS ($p < .001$)	Intermed. » advanced » NS ($p < .001$)	Intermed. » advanced » NS ($p < .001$)
Sentence Verification Task			
Hulstijn et al. (2009).	Grade 8 > Grade 9 > Grade 10 ($p < .001$)	Grade 8 > Grade 9 > Grade 10 ($p < .001$)	Grade 8 < Grade 9 < Grade 10 ($p < .05$)
This study			
Raw data	Intermed. » advanced » NS ($p < .001$)	Intermed. » advanced » NS ($p < .001$)	Intermed. > advanced > NS ($p = .190$)
Cleaned data	Intermed. » advanced » NS ($p \leq .001$)	Intermed. » advanced » NS ($p \leq .001$)	Intermed. > advanced » NS ($p \leq .001$)

Note: The much greater than symbol » indicates a significant difference between groups based on the post hoc analyses. NS, native speaker. RT, reaction time; CV, coefficient of variation.

results of the post hoc analyses. Thus, we do not know which differences between measurement times in Hulstijn et al. were significant.

DISCUSSION

Comparison of findings with Hulstijn et al.'s experiment 1

In this study, we reported the findings for two sets of statistical analyses that were conducted before and after cleaning participants' RTs for three speeded tasks. The general finding was that RTs, standard deviations, and CVs all decreased as a function of participants' English proficiency level: from intermediate ESL learners (L1 Korean), to advanced ESL learners, to native English speakers. This finding was in stark contrast to Hulstijn et al.'s (2009) study, who failed to find a decrease in Dutch high school students' CV over three consecutive grade years, in which the students received ESL instruction in school and plentiful exposure to English (e.g., through the media) outside the classroom. The unvarying CV in Hulstijn et al. (2009) should not be taken as an index that the Dutch high school students did not progress in their knowledge of the English language (see the appendix in Van Gelderen et al., 2007); however, it might suggest that they were still in a relatively early stage of their language learning trajectory, in which declarative knowledge accumulation (vocabulary and grammar learning) dominated.

Participants in the present study were intermediate and advanced ESL learners enrolled at a US university. Except for a small subset of English for Academic Purposes program students, none of the participants were taking dedicated English language courses anymore. They were immersed in an English-speaking environment in which they were functioning as regular undergraduate and graduate students. The communicative pressures that they encountered in their daily lives may thus have pushed them to practice and proceduralize or automatize their existing English knowledge. Furthermore, participants seemed to have a sound knowledge base of English, which is a prerequisite for automatization (Anderson, 1992), as shown in the high accuracy scores on the three speeded tasks (range = 90%–98%).

Given their findings, Hulstijn et al. (2009) questioned the validity of the CV as a measure of automatization in sentence processing. "Although *conceptually* skill acquisition can be distinguished from knowledge accumulation, *in reality,* knowledge accumulation forms part of skill acquisition because, in real L2 learning, exposure to new words goes hand in hand with exposure to words encountered before" (Hulstijn et al., 2009, p. 576, original emphasis). They argued that response efficiency and knowledge accumulation are inherently confounded in RT tasks and that data cleaning is therefore strongly recommended. While the usefulness of data cleaning will depend on the specifics of the study, we emphasize that, in this study, data cleaning was not necessary. Had we not been replicating Hulstijn et al. (2009) conceptually, we would have stopped after the first, raw-data analysis of the correct responses.

The raw-data analysis indicated that the ESL learners were indistinguishable from native speakers in terms of their automatic processing of individual words but not so in their processing of phrasal and clausal units. Particularly, processing

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

efficiency on the sentence construction task decreased significantly for the lower proficiency levels. CVs on the sentence verification task did not differ significantly between groups, yet the numerical pattern was the same as for the sentence construction task. As we explain in the next section, we believe that this interpretation may be the most solid account of our data; CV proved to be a useful measure of automatization of this sample's L2 sentence processing skills.

Why data cleaning may not be necessary

On the semantic classification task, any data cleaning beyond the removal of incorrect responses seemed unnecessary because both groups of ESL learners performed near ceiling (98% accuracy), which leaves very little room for a knowledge confound to influence findings. For the sentence construction task, the results did not change before or after data cleaning. Likewise, the CV-RT correlation analyses accorded well with the CV analyses in both cases (compare Tables 2 and 3), and we would argue that the simpler, raw-data analysis is to be preferred. However, for the sentence verification task the outcome did change as a result of data cleaning. Removal of 16 test items (i.e., 27% of the data) and outlier substitutions caused the difference in CV between native and nonnative speakers (both advanced and intermediate) to become significant. However, we believe that this effect should be interpreted with caution, as it coincided with a significant, *negative* CV-RT correlation in the intermediate ESL group (see Table 3). This signals that intermediate learners who responded more slowly tended to produce less variable RTs and that faster responders tended to show greater variability. Such patterns are at odds with the principle of automatization that fast and reliable processing go hand in hand (Segalowitz, 2010).

Sentence construction proved a better task than sentence verification

Following Hulstijn et al. (2009), we adopted a sentence verification and a sentence construction task to examine participants' automatization of L2 sentence processing. The results of the two tasks were compatible (as the positive correlation coefficients between CVs in the two tasks indicate) but not identical. In the construction task, the advanced group exhibited more automatized processing than the intermediate group, whereas in the sentence verification task, the advanced group only responded faster but not more automatically (see Table 4). These results were qualified by an interesting item effect: when only plausible items in the sentence verification task were analyzed, the correlation with the sentence construction task disappeared and the mean CV for the different groups became very similar. At this juncture, a natural question arises: which underlying processes does each task and item type tap into?

Sentence-level processing is more complex than word-level processing because in addition to identifying the individual words that make up the sentence (lexical access), readers should establish how they belong together structurally (syntactic parsing) and draw meaning from them. (At the clause level, the third process is called semantic proposition formation; Grabe & Stoller, 2011.) Grabe and Stoller defined lexical access, syntactic parsing and semantic proposition formation as

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

lower level linguistic processes.[6] They may differ from higher level processes, such as the use of background knowledge to draw inferences, in that they "occur relatively automatically for the fluent reader" (Grabe & Stoller, 2011, p. 19). Grabe and Stoller do not really discuss the possibility of automatizing higher level processes, but we believe that these processes cannot be automatized for two reasons: first, they require conscious processing, and second, they are highly context dependent. For instance, an inference one makes in one reading context will seldom be reusable in a different reading context, and so it is difficult to practice and automatize that specific procedure. There is no doubt that inferencing, using background knowledge, building a text model, and so on, are integral components of skilled reading (e.g., Grabe & Stoller, 2011; Koda, 2005; Schmalhofer & Perfetti, 2007); however, we question that the CV is as suited for measuring these skills as it is for measuring participants' lower level processing skills.

The implication of this for CV studies on sentence processing is that tasks that reduce the influence of higher level processes will likely produce clearer findings. In our study, this was the sentence construction task. We would argue that besides word recognition, the construction task primarily assessed participants' parsing, which is a "very [rapid process] without much effort or conscious attention" (Grabe & Stoller, 2011, p. 17) in fluent language users. Predicting the next word after a given phrase likely forced participants to build a sentence structure incrementally. Then, to predict the next word, participants had to make use of word order and probabilistic information expeditiously so they could select the response option that belonged to the syntactic category that they predicted would follow. Given that the syntactic structure of Korean is distinct from that of English (e.g., head-final vs. head-initial language) and that, therefore, L1 syntactic transfer will be negative in the beginning stages of learning (Grabe & Stoller, 2011; Koda, 2005), syntactic parsing may have been a particularly informative subskill to investigate in relation to Korean speakers' automatization of L2 English. In general, there is a case for studying automatization in bilingual speakers of typologically distinct languages as it will avoid issues of positive orthographic, lexical, and syntactic transfer (for insightful discussion, see Koda, 2005, 2007).

Whereas the sentence construction task made heavy demands on syntactic parsing, the sentence verification task may have invited the use of top-down, concept-driven knowledge (i.e., world knowledge, background knowledge). Participants were asked to evaluate the plausibility of sentences, which often required their use of factual knowledge or personal experience. For instance, one participant commented that the sentence *Jim passed the exam because he was happy* could be either sensible or nonsensible. Though experience tells us that passing an exam makes people feel happy, and not the other way around, a reader might make the additional inference that *Jim passed the exam because he was happy and happy people are more productive.* It is important to note that this kind of additional step in the interpretation process would only have occurred in items that were intended to be nonsensible (because the sensible items were inherently straightforward). That only the native speakers made disproportionately more errors on these nonsensible items (by judging them as sensible) attests to the mental flexibility that skilled readers bring to the meaning making task (e.g., Grabe & Stoller, 2011; Koda, 2005; Schmalhofer & Perfetti, 2007).

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Given that default processing did not yield a congruent interpretation for the nonsensible items, it is all the more surprising that these items were indispensable in discriminating between the different groups in terms of their automaticity. Recall that when we analyzed the sensible items (hits) only, all the patterns expected under an automatization account disappeared. However, when RTs for both hits and correct rejections were considered, CVs decreased with increasing proficiency level and correlated strongly with the CVs on the sentence construction task. Although further research would be needed to clarify this issue, this may suggest that the sentence verification task benefitted from having sentences that tested participants' processing flexibility (i.e., the nonsensible sentences, which included different types of violations) and that, perhaps, processing flexibility factors into the CV when computed for more complex tasks such as sentence verification.

The discussion of plausible and implausible items resonates with a similar question in the area of implicit and explicit knowledge, namely, whether grammatical and ungrammatical grammaticality judgment test items measure the same or different types of knowledge. Researchers who have found that grammatical and ungrammatical items perform differently (R. Ellis, 2005; Godfroid et al., 2013; Gutiérrez, 2013; Zhang, 2013) have taken this to mean that grammatical items measure implicit knowledge and ungrammatical items measure explicit knowledge. Therefore, we had not anticipated that the implausible sentence verification items, which are conceptually similar to ungrammatical grammaticality judgment test items, would help to differentiate our three groups' CVs, as automatic processes (captured by the CV) are said to operate on implicit or procedural, rather than analyzed explicit, knowledge (e.g., Anderson, 2007; DeKeyser, 2007b; R. Ellis, 2005; Paradis, 2004, 2009; Segalowitz & Hulstijn, 2005).

Longitudinal versus cross-sectional studies of automatization

Ortega and Iberri-Shea (2005) noted that "many questions concerning L2 learning are fundamentally questions of time and timing" (p. 27). The development of automaticity is no exception, as it requires months or years of deliberate practice to occur (DeKeyser, 2007a). Because automatization is a developmental phenomenon, it lends itself naturally to being studied with a within-subject design, longitudinally, or in a training study (for examples see Akamatsu, 2008; DeKeyser, 1997; Hulstijn et al., 2009; Robinson, 1997; Segalowitz & Segalowitz, 1993, exp. 2; Segalowitz, Segalowitz, & Wood, 1998; Segalowitz, Watson, & Segalowitz, 1995; Serafini, 2013). However, given the complexities involved in sentence-level processing, the time span needed to observe significant progress in L2 automatization may well exceed a couple of years,[7] which raises the question of how feasible it is for many researchers, especially in early career, to trace the development of automaticity longitudinally. As stated in the introductory section, a possible reason why Hulstijn et al. (2009) did not find a significant within-subject CV reduction in high school students at Grades 8, 9, and 10 is that 2 years may not have sufficed for processing efficiency in these intermediate-level learners to develop.

The present study examined the development of L2 automaticity by using a cross-sectional design (see also Harrington, 2006; Phillips et al., 2004; Rodgers,

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

2011; Segalowitz & Freed, 2004; Segalowitz & Segalowitz, 1993; Serafini, 2013). An advantage of cross-sectional studies is that it is easier to sample participants from more clearly distinct developmental stages; however, between-subject experiments do rest on two assumptions. First, all participants are assumed to follow a common developmental path in terms of automatization. Second, a higher proficiency level means a higher level of automaticity. Regarding the first assumption, Anderson's ACT-R theory suggests a common trajectory of L2 skill development, whereby all learners move from the declarative, through the procedural, to the automatic stage (e.g., Anderson, 2007). The distinction between speedup and automatization (Segalowitz & Segalowitz, 1993) also carries implications for the developmental aspect of skill acquisition, namely, mere speedup precedes automatization (Wickens, 1992). Crucial in between-subject designs is the second assumption that level of proficiency indexes level of automaticity. Although the RT data of the current study satisfied this assumption, it might not be always the case. As documented in Swain (2000), a French immersion program attracted much criticism because immersion students consistently had problems with accuracy, while showing nativelike fluency in their oral production. In this case, the immersion students had achieved a high level of L2 automaticity, whereas their proficiency level, which takes both accuracy and fluency into account, was lagging behind. The reverse scenario of high accuracy and low fluency also exists, for example, in many ESL programs in East Asia, which prepare students for passing high-stakes English language tests but do not give them enough opportunity to put their linguistic knowledge into practice. Thus, the link between proficiency level and degree of automatization ought to be evaluated within the context in which language learning takes place.

It is important to keep in mind that automatization occurs in a relatively late stage of L2 development (Anderson, 1983; Segalowitz & Hulstijn, 2005; Wickens, 1992). This is also illustrated in Anderson's skill acquisition theory where learners in the earlier phases primarily focus on acquiring rules and facts, while those in the later phases are more likely to fine tune their processing skills. Because beginning or even intermediate learners are likely to work on both knowledge accumulation and skill development (Hulstijn et al., 2009), the development of automatic processing skills might be more visible among relatively advanced learners.

CV as a measure of automatization in reading

The use of CV as a measure of automatization rests on a subtle distinction between a proportional reduction of processing variability (speedup) and a disproportional reduction (automatization) relative to an increase in processing speed. Segalowitz argued that a disproportional reduction in standard deviation, and hence a smaller CV, will only occur if the underlying processes are restructured (Segalowitz, 2010; Segalowitz & Segalowitz, 1993). Thus, in Segalowitz' view, the restructuring of cognitive processes that enable skilled performance is key to automatization.

However, in an early account of automaticity, McLaughlin (1987) observed that "[cognitive] theory is not precise enough about when and how restructuring occurs" (p. 152). Despite the advances that have been made since then, we argue his

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

general assessment still holds true, especially as regards the restructuring of communicative competence in the four primary skills (i.e., speaking, listening, writing or reading) rather than the restructuring of rule knowledge. A reason for this is that communicative skills are so complex. They comprise a variety of subskills, which draw on distinct linguistic (and sometimes nonlinguistic) knowledge. Therefore, in theory automatization could entail reorganizing the different subskills involved (e.g., by eliminating a subcomponent) or it could consist of the automatization of one particular subskill only (e.g., more efficient word recognition in reading). Finally, because the subskills themselves differ in the knowledge sources they require, some component processes likely cannot be automatized at all.

If we look at the automatization of reading, which was the main skill targeted in this study, it appears that the phenomenon can be understood best in terms of a component skills approach (Carr & Levy, 1990; Koda, 2005, 2007) and, optionally, the assumption of limited attentional capacity (LaBerge & Samuels, 1974). The component skills approach to reading holds that reading is "the product of a complex information-processing system, involving a constellation of closely related mental operations" (Koda, 2005, p. 19; see also Carr, Brown, Vavrus, & Evans, 1990, p. 5). The different mental operations that make up reading are widely believed to be organized hierarchically (e.g., Breznitz, 2006; Grabe & Stoller, 2011; Koda, 2005; Stanovich, 2000), from lower level, data-driven processes to higher level, conceptual processes. On the assumption that attention is a limited resource (LaBerge & Samuels, 1974), the inefficient, effortful execution of lower level processes (e.g., word decoding) will negatively affect the amount of attention available for higher level reading processes. "Because inefficient decoding is resource demanding, it severely restricts readers' involvement in higher order comprehension operations, such as text-information integration, inference, and reasoning" (Koda, 2007, p. 23).

Although not all accounts of automatization in reading assume that attention is capacity limited (see Stanovich, 2000), there does seem to be a consensus that reading subprocesses differ in whether they can be automatized. Likewise, there is theoretical agreement that some reading subprocesses may be automatized before others. From a component skills perspective, then, it seems important to analyze the specific cognitive demands of one's sentence processing tasks, as it is the automatization of these subskills that CV will measure primarily. In effect, this amounts to measuring the automatization of a particular reading subskill; as of yet, we do not see how CV computations for sentence-level processing tasks could address the question of how the different subskills are organized (and potentially reorganized or restructured) within the information-processing system as a result of practice.

Finally, because of the high temporal sensitivity of CV we believe it may work best with tightly controlled tasks such as lexical decision, semantic classification, and sentence construction. Given that other fluency measures are available (see Andringa, Olsthoorn, Van Beuningen, Schoonen, & Hulstijn, 2012; De Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2011; Schoonen et al., 2003; Van Gelderen et al., 2004) a challenge for future researchers is to determine for what tasks CV is a good measure of cognitive fluency and in which cases other, more global fluency measures may be more informative.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Limitations and future prospects

The current study did not use all of the tasks that Hulstijn et al. (2009) did. Because our participants performed word and sentence processing tasks in L2 English only, we cannot make any claims about the automaticity of their L1 Korean processing skills. It is conceivable that the ESL learners transferred some of their L1 reading skills to the L2 sentence processing tasks, even though the typological distance between Korean and English would seem to preclude large amounts of positive transfer. Nonetheless, it would be prudent to include the corresponding L1 tasks in future studies so as to obtain measures of L2 specific cognitive fluency.

In the preceding discussion sections, we emphasized the importance of the task and task characteristics in CV studies of automatization. Although we analyzed the sentence construction and sentence verification tasks at some length, these analyses may benefit from additional, independent support. For example, RT experiments can be triangulated with other psycholinguistic measures, such as ERPs, to obtain converging evidence. As a case in point, Phillips et al. (2004) validated the use of CV by comparing the CV analysis with ERP measures that were recorded during a semantic priming task measuring word recognition skills. To justify the use of CV in sentence processing, similar comparisons need to be made at the sentence level.

CONCLUSION

The purpose of the study was to investigate the development of L2 automaticity at the sentence level and to validate the use of CV as a measure of L2 automatization in sentence processing. The study was a partial conceptual replication of Hulstijn et al. (2009), using the same analyses on a subset of the same speeded tasks. Compared to Hulstijn et al., however, the current study made important changes to the data collection and data analysis procedures: (a) participants' L1 background was restricted to Korean, which is typologically distinct from English; (b) learners' English proficiency was assessed independently via an online English diagnostic test; (c) participants were likely more advanced than the high school students in Hulstijn et al.; and (d) the sentence stimuli in the speeded tasks were made shorter.

As a result, the CV in the sentence-related tasks did decrease as learners' proficiency level increased, suggesting that the different ESL groups in this study differed in the automatization of their L2 processing skills. This result was found regardless of whether the data were cleaned; it lends support to Segalowitz and Segalowitz' (1993) hypothesis. Hence, the data from the Korean college-level ESL students seem to justify the expansion of the CV measure to the sentence processing level.

However, whether a smaller CV really indicates increased automaticity is still an open question. The assumptions on which Segalowitz and Segalowitz' (1993) logic is based require further theoretical elaboration and empirical support; in particular, the jury is still out on whether the elimination of initially slow subcomponents really causes a disproportionate reduction in standard deviation and whether the argument also goes the other way. By the same token, it seems to be difficult to

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

falsify Segalowitz and Segalowitz' hypothesis. If CV did not decrease, it is taken to indicate mere speedup, whereas if CV decreased, it is a sign of automatization. As the theoretical foundation for CV (a smaller CV means a higher level of automaticity) becomes a reason for data interpretation, the argument for the validity of CV appears to be somewhat circular. In addition to crunching numbers, we will need psycholinguists' insights into how L2 learners process linguistic information, what cognitive subcomponents can be actually removed through practice, and what aspects of L2 processing can or cannot be automatized. Although CV experiments are far from solving the automaticity puzzle, they certainly provide interesting behavioral pieces that will keep psycholinguists busy for some time to come.

APPENDIX A

Table A.1. *Word stimuli for the semantic classification task*

Words	Frequency	Words	Frequency
Americans	55.98	Ink	9.58
Animal	119.16	Letter	121.43
Baby	191.21	Mailman	0.18
Bed	254.39	Men	679.65
Bicycle	18.44	Mother	446.20
Bill	55.45	People	1317.05
Bird	43.25	Player	29.87
Book	252.25	Card	44.68
Bottle	87.22	Servant	25.22
Boy	223.87	Shoes	67.88
Bread	76.98	Skirt	21.77
Bus	65.09	Snow	60.03
Car	274.92	Soccer	3.87
Cat	43.31	Student	94.12
Chair	110.66	Summer	122.50
Desk	86.50	Sunglasses	3.09
Dog	73.24	Teacher	71.75
Exam	4.88	Telephone	102.92
Father	343.99	Train	74.43
Girl	284.73	Uncle	61.58
Grandfather	23.32	Wheel	28.91
Grandmother	16.36	Women	520.80
Horse	87.87	Mean	152.36
Human	308.17	*SD*	227.88

Note: Word frequencies (occurrences per million) are based on the trimmed version of the CELEX database *MC Word: An Orthographic WordForm Database* (Medler & Binder, 2005).

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table A.2. *Stimuli for the sentence verification task*

No.	Sentence	Structure	Key
1	A horse is an animal that can fly.	Relative clause	NO
2	Summer is the hottest time of a year.	SVO	YES
3	Soccer is a sport for two or four players.	SVO	NO
4	Jane was written by the letter.	Passive	NO
5	Most bicycles have seven wheels.	SVO	NO
6	Jim went to bed because he was sleepy.	Subordinate clause	YES
7	Jim was sleepy although he went to bed.	Subordinate clause	NO
8	My mother had me write a letter.	Causative	YES
9	She believed her children to be honest.	Exceptional verb	YES
10	The boy whose sister plays soccer is tall.	Relative clause	YES
11	People always wear shoes before socks.	SVO	NO
12	The book that you have is delicious.	Relative clause	NO
13	Good teachers are believed to be helpful.	Exceptional verb	YES
14	Jim was happy because he passed the exam.	Subordinate clause	YES
15	Jim passed the exam because he was happy.	Subordinate clause	NO
16	A mailman is delivered by many postcards.	Passive	NO
17	The girl who has lots of CDs loves music.	Relative clause	YES
18	The letter was written by my grandmother.	Passive	YES
19	If you want to be clean, take a shower.	Subordinate clause	YES
20	If you study hard, you will fail the exam.	Subordinate clause	NO
21	Jane crossed the ocean by riding a bike.	SVO	NO
22	Jane had her dog pay the bill for her.	Causative	NO
23	She expects her mother to come tomorrow.	Causative	YES
24	Americans go to work seven days a week.	SVO	NO
25	In the States, most cars are made of wood.	Passive	NO
26	He went to a post office to buy some bread.	SVO	NO
27	The student who came late left very early.	Relative clause	YES
28	Take off your jacket if you feel cold.	Subordinate clause	NO
29	Women in the US never wear skirts.	S negVO	NO
30	Why don't you study harder if you feel tired?	Subordinate clause	NO
31	The boy who likes music has lots of CDs.	Relative clause	YES
32	The book that I taught math is my teacher.	Relative clause	NO
33	Sick people help a doctor to get better.	Causative	NO
34	If you eat too much, you can get too thin.	Subordinate clause	NO
35	You can drink water if you feel thirsty.	Subordinate clause	YES
36	Most theaters in the US have only one chair.	S adverbs VO	NO
37	Mothers are always older than their babies.	Comparative	YES
38	The car that I bought went to a vacation.	Relative clause	NO
39	This is your pencil that I borrowed.	Relative clause	YES
40	Humans are smarter than animals.	Comparative	YES
41	The baby that I saw yesterday is her father.	Relative clause	NO
42	English was taught by the new teacher.	Passive	YES
43	In general, trains are shorter than buses.	Comparative	NO
44	People wear a coat when they go to bed.	Subordinate clause	NO
45	The boy that I liked moved to Canada.	Relative clause	YES
46	The girl picked up the phone when it rang.	Subordínate clause	YES
47	My mother was cooked by a cake.	Passive	NO

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table A.2 (*cont.*)

No.	Sentence	Structure	Key
48	April comes right before March.	SVO	NO
49	You shouldn't believe everything you hear.	Relative clause	YES
50	Jane expects a church to go to Sunday.	Causative	NO
51	He woke up early because he slept late.	Subordinate	NO
52	He went inside, though it started to rain.	Subordinate	NO
53	A telephone is often used to open a bottle.	Passive	NO
54	The man ate lunch because he was hungry.	Subordinate	YES
55	The teacher had the students sit down.	Causative	YES
56	Jane had the desk fix the broken legs.	Causative	NO
57	My uncle made me for a snowman.	Dative	NO
58	The homework gave me a headache.	Dative	YES
59	The letters were written in black ink.	Passive	YES
60	People go to bed wearing sunglasses.	SVO	NO

Note: Modified from Hulstijn et al. (2009). SVO, subject–verb–object.

Table A.3. *Stimuli for the sentence construction task*

No.	On the First Screen	Options		Key
1	I wonder what . . .	1) does	2) he	2N
2	Mr. and Ms. Smith . . .	1) go	2) to	1Y
3	What does that . . .	1) she	2) mean	2N
4	Jane's cat . . .	1) cute	2) is	2N
5	The students . . .	1) went	2) angry	1Y
6	I know what	1) she	2) does	1Y
7	I expect . . .	1) them	2) go	1Y
8	Jane had him . . .	1) to	2) go	2N
9	In the past. . .	1) happy	2) she	2N
10	Does . . .	1) Jane	2) come	1Y
11	Aren't . . .	1) busy	2) you	2N
12	Why . . .	1) she	2) can't	2N
13	I know where . . .	1) is	2) he	1Y
14	It was . . .	1) write	2) written	2N
15	Where . . .	1) he	2) will	2N
16	Would you . . .	1) a book	2) bring	2N
17	He came . . .	1) often	2) home	2N
18	It is not true . . .	1) you	2) what	2N
19	He gave me	1) a letter	2) to	1Y
20	Does he . . .	1) know	2) Jane	1Y
21	Isn't it . . .	1) can	2) true	2N
22	Where . . .	1) do	2) you	1Y
23	Students are . . .	1) leave	2) asked	2N
24	These students . . .	1) are	2) angry	1Y
25	Don't . . .	1) that	2) do	2N
26	How often . . .	1) it	2) does	2N

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Table A.3 (*cont.*)

No.	On the First Screen	Options		Key
27	They believe that . . .	1) God	2) exists	1Y
28	Yesterday . . .	1) happy	2) Mary	2N
29	Would you . . .	1) do	2) going	1Y
30	The boy who . . .	1) kind	2) is	2N
31	Jane let him . . .	1) go	2) going	1Y
32	The letter . . .	1) was	2) writes	1Y
33	He made me . . .	1) for	2) a box	2N
34	Football players . . .	1) make	2) lots	1Y
35	After some time . . .	1) woke	2) she	2N
36	It is . . .	1) not	2) does	1Y
37	They are . . .	1) expected	2) expect	1Y
38	All students . . .	1) have	2) not	1Y
39	You think that . . .	1) Jane	2) does	1Y
40	I wonder . . .	1) if	2) he	1Y

Note: Modified from Hulstijn et al. (2009).

APPENDIX B

English Language Learning Background Questionnaire

1. Age: _____
2. Gender: Male ☐ Female ☐
3. Mother tongue (First language): _____
4. Other language spoken at home as a child:_____
5. Academic status (please choose all applied to you)
   ☐ Freshman ☐ Sophomore ☐ Junior ☐ Senior ☐ MA ☐ PhD ☐ EAP (advanced ESL)
   ☐ Other (please specify) _____
6. How long have you been studying in the English-speaking countries (including previous education, if applicable)?
   For _____ years _____ months.
7. **At what age** did you **begin studying English**? _____
8. **At what age** were you **first exposed** to English communication environment? _____
   For how long? _____
9. Have you learned (or studied) any foreign languages other than English? ☐ Yes ☐ No
   If yes, please specify all information including 1) which language, 2) how long, and 3) in what context(s)
   I have learned (or studied) _____ for _____ month(s) context:
   _____
   I have learned (or studied) _____ for _____ month(s) context:
   _____
   [Example] *I have learned French for 3 years as a part of my coursework in high school. I have learned German for 2 year because my family moved to Germany during that time*

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

10. English Proficiency Test Score (if you remember)
    Test: _____ (e.g., TOEFL, TOEIC, ILETS, TEPS, MSUELT)
    Total score: _____
    Test Date: _____ (i.e., summer 2007) (Reading score:_____)
11. How much time do you usually spend using English from outside of the classroom?

Language Skills	Description	Use of Time (Weekly)
Reading	Newspaper (including online or internet news), magazines, books (including textbooks), academic journals, and any other types of reading activities in English	(about _____ hours) per week
Writing	Homework assignment, essay, diary, English chatting, and any other types of writing activities in English	(about _____ hours) per week
Listening	Pop music (in English), lectures, TV, radio, and any other types of listening activities in English	(about _____ hours) per week
Speaking	Conversing with friends who are native speakers of English, in-class discussions in English, and any other types of speaking activities	(about _____ hours) per week

Your privacy will be protected. Thank you for your participation.

APPENDIX C

Deleted items from the sentence verification task (16 out of 60 items)

6. The boy whose sister plays soccer is tall. (sensible)
8. Good teachers are believed to be helpful. (sensible)
11. Jim passed the exam because he was happy. (nonsensible)
15. Jim was borrowed the note from his friend. (nonsensible)
16. A mailman is delivered by many postcards. (nonsensible)
20. If you study hard, you will fail the exam. (nonsensible
24. Americans go to work seven days a week. (nonsensible)
27. The student who came late left very early. (sensible)
28. Take off your jacket if you feel cold. (nonsensible)
30. Why don't you study harder if you feel tired? (nonsensible)

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

42. English was taught by the new teacher. (sensible)
48. April comes right before March. (nonsensible)
50. Jane expects a church to go to Sunday. (nonsensible)
52. He went inside, though it started to rain. (nonsensible)
56. Jane had the desk fix the broken legs. (nonsensible)
57. My uncle made me for a snowman. (nonsensible)

Deleted plausible items from the sentence verification task (2 out of 25 items)

6. The boy whose sister plays soccer is tall. (sensible)
42. English was taught by the new teacher. (sensible)

Deleted items from the sentence construction task (9 Items)

No.	On the First Screen	Options		Key
1	I wonder what . . .	1) does	2) he	2N
8	Jane had him . . .	1) to	2) go	2N
12	Why . . .	1) she	2) can't	2N
13	I know where . . .	1) is	2) he	1Y
15	Where . . .	1) he	2) will	2N
17	He came . . .	1) often	2) home	2N
25	Don't . . .	1) that	2) do	2N
26	How often . . .	1) it	2) does	2N
33	He made me. . .	1) for	2) a box	2N

ACKNOWLEDGMENTS

We thank Jan Hulstijn, Amos Van Gelderen, and Rob Schoonen for sharing their stimuli with us and kindly taking the time to answer our questions.

NOTES

1. The power law of practice describes a learning curve where RTs for a particular task decrease exponentially with the number of practice trials (Delaney et al., 1998).
2. On some tasks, the CV increased or decreased significantly, but the effect sizes were very small (the largest partial $\eta^2 = 0.038$).
3. An expectation-maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on unobserved latent variables.
4. This task was the same as Hulstijn et al.'s (2009) sentence production task. However, we prefer the term *sentence construction task* to the term *sentence production task* because the language-production requirements were minimal. When individuals read a text (which is a receptive language activity), they are also known to generate predictions about the upcoming words (e.g., Ehrlich & Rayner, 1981), as they had to do in this task.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

5. The multiple imputation method was chosen rather than the EM algorithm (used by Hulstijn et al., 2009) for two reasons: first, it is as rigorous a method as the EM algorithm for missing data analysis (King & Scheve, 2001). Second, the EM algorithm in the Statistical Package for the Social Sciences does not have a function to put constraints on data generation. It is necessary to specify the minimum and maximum allowable imputed values for data generation because RTs only fall within a given range (e.g., no RTs in a word recognition task smaller than 10 ms or larger than 10000 ms, and no negative values).
6. The distinction between lower level and higher level processes in reading is common; however, different researchers draw the boundary between low and high at different levels (compare, for instance, Breznitz, 2006). For the purposes of this paper, the important part is that there is a subset of reading processes (lower level processes in Grabe and Stoller's, 2011, terms) that can operate automatically, whereas other, more cognitively complex reading processes may not be amenable to automatization.
7. DeKeyser (2000) posited that it takes 10 years of residence in the target-language community for L2 users to reach their ultimate attainment in morphosyntactic knowledge and processing.

REFERENCES

Anderson, J. R. (1976). *Language, memory, and thought*. Mahwah, NJ: Erlbaum.
Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
Anderson, J. R. (1992). Automaticity and the ACT theory. *American Journal of Psychology, 105*, 165–180.
Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.
Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.
Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning, 62*, 49–78.
Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics, 29*, 175–193.
Baars, B. J., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences, 7*, 166–172.
Bekoff, M. (1997). Quantitative studies of three areas of classical ethology: Social dominance, behavioral taxonomy and behavioral variability. In B. A. Hazlett (Ed.), *Quantitative methods in the study of animal behavior* (pp. 1–46). New York: Academic Press.
Bialystok, E. B. (1978). A theoretical model of second language learning. *Language Learning, 28*, 69–83.
Breznitz, Z. (2006). *Reading fluency: Synchronization of processes*. Mahwah, NJ: Erlbaum.
Carlton, M., Robertson, R., Carlton, L., & Newell, K. (1985). Response timing variability: Coherence of kinematic and EMG parameters. *Journal of Motor Behavior, 17*, 301–319.
Carr, T. H., Brown, T. L., Vavrus, L. G., & Evans, M. A. (1990). Cognitive skill maps and cognitive skill profiles: Componential analysis of individual differences in children's reading efficiency. In T. H. Carr & B. A. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 1–55). San Diego, CA: Academic Press.
Carr, T. H., & Levy, B. A. E. (1990). *Reading and its development: Component skills approaches*. San Diego, CA: Academic Press.
Council of Europe. (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. New York: Cambridge University Press.
Dawkins, R., & Dawkins, M. (1973). Decisions and the uncertainty of behavior. *Behaviour, 45*, 83–103.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2011). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*, 4–34.
DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19*, 195–221.
DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22*, 499–533.
DeKeyser, R. (Ed.). (2007a). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. New York: Cambridge University Press.
DeKeyser, R. (2007b). Skill acquisition theory. In J. Williams & B. Van Patten (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Erlbaum.
Dörnyei, Z. (2009). *The Psychology of Second Language Acquisition*. Oxford: Oxford University Press.
Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*, 305–352.
Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition, 27*, 141–172.
Fauconnier, G. (1994). *Mental spaces*. Cambridge: Cambridge University Press.
Flehming, H. C., Steinborn, M., Langner, R., Scholz, A., & Westhoff, K. (2007). Response time variability in serial choice RT tasks. *Psychology Science, 49*, 132–149.
Godfroid, A., Loewen, S., Jung, S., Park, J., Gass, S. M., & Ellis, R. (2013, October). *Multiple perspectives on grammaticality judgments: Do different types of knowledge imply different types of processing?* Paper presented at the Second Language Research Forum, Provo, UT.
Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading*. Harlow: Longman.
Gunter, T. C., Hahne, A., & Mauth, K. (2004). The relative timing of syntactic and semantic processes in sentence comprehension. *Syntax, 15*, 1–5. doi:10.1097/01.wnr.0010093585.33576.65
Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition, 37*, 1–27.
Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook, 6*, 147–168. doi:10.1075/eurosla.6.10har
Hasting, A. S., & Kotz, S. A. (2008). Speeding up syntax: On the relative timing and automaticity of local phrase structure and morphosyntactic processing as reflected in event-related brain potentials. *Journal of Cognitive Neuroscience, 20*, 1207–1219. doi:10.1162/jocn.2008.20083
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
Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning, 57*, 1–33. doi:10.1111/j.1467-9922.2007.00397.x
Jiang, N. (2012). *Conducting reaction time research in second language studies*. New York: Routledge.
King, G., & Scheve, K. (2001). Analyzing incomplete political science data: An Alternative algorithm for. *American Political Science Review, 95*, 49–69.
Koda, K. (2005). *Insights into second language reading*. New York: Cambridge University Press.
Koda, K. (2007). Reading and language learning: Cross-linguistic constraints on second language reading development. *Language Learning, 57*, 1–44.
LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology, 6*, 293–323.
Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review, 95*, 492–527.
Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review, 109*, 376–400.
McLaughlin, B. (1987). *Theories of second language learning*. London: Arnold.
McManus, I., Kemp, R. & Grant, J. (1986). Differences between fingers and hands in tapping ability: Dissociation between speed and regularity. *Cortex, 22*, 461–473.
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/
Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26–45.
Paradis, M. (2004). *A neurolinguistic theory of bilingualism* (Vol. 18). Amsterdam: John Benjamins.
Paradis, M. (2009). *Declarative and procedural determinants of second languages* (Vol. 40). Amsterdam: John Benjamins.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Phillips, N., Segalowitz, N., O'Brien, I., & Yamasaki, N. (2004). Semantic priming in a first and second language: Evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics, 17*, 237–262. doi:10.1016/S0911-6044(03)00055-1

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced and instructed conditions. *Studies in Second Language Acquisition, 19*, 223–247.

Rodgers, D. M. (2011). The automatization of verbal morphology in instructed second language acquisition. *International Review of Applied Linguistics, 49*, 295–319.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association, 72*, 538–543.

Schmalhofer, F., & Perfetti, C. A. (Eds.). (2007). *Higher level language processes in the brain: Inference and comprehension processes*. Mahwah, NJ: Erlbaum.

Schoonen, R., Gelderen, A. V., Glopper, K. D., Hulstijn, J., Simis, A., Snellings, P., et al. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53*, 165–202.

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N., & de Almeida, R. G. (2002). Conceptual representation of verbs in bilinguals: Semantic field effects and a second-language performance paradox. *Brain and Language, 81*, 517–531. doi:10.1006/brln.2001.2544

Segalowitz, N., & 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., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition, 33*, 644–653.

Segalowitz, N., & Hulstijn, J. (2005). Automaticity in second language learning. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 371–388). Oxford: Oxford University Press.

Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice and the differentiation of speedup from automatization effects. *Applied Linguistics, 14*, 369–385.

Segalowitz, N., 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. doi:10.1017/S0142716400010572

Serafini, E. (2013). *Cognitive and psychosocial factors in the long-term development of implicit and explicit second language knowledge in adult learners of Spanish at increasing proficiency*. Ann Arbor, MI: ProQuest.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford Press.

Swain, M. (2000). French immersion research in Canada: Recent contributions to SLA and applied linguistics. *Annual Review of Applied Linguistics, 20*, 199–212.

Talmy, L. (2008). Aspects of attention in language. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 27–38). New York: Routledge.

Toichi, M., Suguiura, T., Murai, T., & Sengoku, A. (1997). A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of R-R interval. *Journal of the Autonomic Nervous System, 62*, 79–84. doi:10.1016/S0165-1838(96)0012-9

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., et al. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first-and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology, 96*, 19–30.

Van Gelderen, A., Schoonen, R., Stoel, R. D., De Glopper, K., & Hulstijn, J. (2007). Development of adolescent reading comprehension in language 1 and language 2: A longitudinal analysis of constituent components. *Journal of Educational Psychology, 99*, 477–491.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press

Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review, 114*, 830–841. doi:10.1037/0033-295X.114.3.830

Weber, E., Shafir, S., & Blais, A. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111*, 430–445.

Wickens (1992). *Engineering psychology and human performance* (2nd ed.). New York: Harper Collins.

Zhang, R. (2013). *The relationship between L2 knowledge and general L2 proficiency in an EFL context.* Unpublished doctoral dissertation, University of Auckland.

https://doi.org/10.1017/S0142716414000137 Published online by Cambridge University Press