The Coefficient of Variance as an index of L2 lexical processing skill

This is a working draft. Please do not quote or circulate without permission.
Comments welcome.

December 12, 2005

Michael Harrington
Linguistics, School of English, Level 4, Michie
University of Queensland, Brisbane, QLD 4103
Australia
mwharr@uq.edu.au

The Coefficient of Variance (mean standard deviation/mean Response time) is a measure of response time variability that corrects for differences in mean Response time (RT) (Segalowitz & Segalowitz, 1993). A positive correlation between decreasing mean RTs and CVs ($r_{\text{CV-RT}}$) has been proposed as an indicator of L2 automaticity and more generally as an index of processing efficiency. The current study evaluates this claim by examining lexical decision performance by individuals from three levels of English proficiency (Intermediate ESL, Advanced ESL and L1 controls) on stimuli from four levels of item familiarity, as defined by frequency of occurrence. A three-phase model of skill development defined by changing $r_{\text{CV-RT.}}$values was tested. Results showed that RTs and CVs systematically decreased as a function of increasing proficiency and frequency levels, with the $r_{\text{CV-RT}}$ serving as a stable indicator of individual differences in lexical decision performance. The $\mathrm{r}_{\text{CV-RT}}$ and automaticity/restructuring account is discussed in light of the findings. The CV is also evaluated as a more general quantitative index of processing efficiency in the L2.

Key words: coefficient of variance, second language processing, processing efficiency proficiency, automaticity

1. Introduction

Fluent performance in a second language (L2) requires the necessary linguistic knowledge and the ability to access that knowledge in a rapid and efficient manner. The processes subserving the latter are receiving increasing attention in SLA theory (Piennemann, 1998; Carroll, 2001; VanPatten, 2004; Sharwood-Smith & Truscott, 2005). As a result methodological issues have also emerged concerning the measurement of L2 processing performance and the relation of these measures to basic theoretical constructs in SLA theory in general, and in accounts of fluency development in particular (Martinis, 2003). Response time is a key measure of processing performance, with faster responses taken as ipso facto evidence for greater proficiency (Koda, 1996; Dijkstra, 2005). Relative response time differences, e.g., between groups, are typically the results of theoretical interest, but response variability, especially as it relates to absolute response time values, can also be informative. Segalowitz and colleagues have examined the relationship between response time variability, as reflected in the *coefficient of variance* (CV) and L2 processing skill development (Segalowitz & Segalowitz, 1993; Segalowitz, Watson & Segalowitz, 1995; Segalowitz, Segalowitz & Wood, 1998; Segalowitz & Hulstijn, 2005).

The CV is a ratio of variability to mean response time. It is measured by dividing the the mean response time (RT) by the standard deviation (SD) of the mean RT. Unlike the information provided by considering the mean RTs and SDs separately, the CV is a single index of processing independent of absolute mean RT values. The CV is a quantitative measure of processing efficiency that, according to Segalowitz et al, can also serve as an index of emerging automaticity in L2 performance (Segalowitz & Segalowitz, 1993). Automaticity in performance is said to be evident when forms (words, phrases etc) are made available quickly, with minimal attention or control required by the individual. The emergence of automaticity is assumed to result from a qualitative change, or *restructuring* of the underlying processes and representations responsible for the performance in the course of development (for reviews of automaticity in SLA see DeKeyser, 2002 and Segalowitz & Hulstijn, 2005).

Response speed is a necessary condition for identifying automaticity but it is not sufficient, and the authors make a distinction between *speed-up* and *automatization* in task performance (Segalowitz & Segalowitz, 1993). In instances of speed-up, faster RTs are accompanied by proportionally faster SDs, but the CV remains unchanged (a mean RT of 1000 milliseconds and SD of 250 msecs has the same CV (.25) as a mean RT of 800 msec and SD 200 msecs). In the case of automatization faster responses times (lower RTs) there must be a decrease in mean RT and a disproportionate decrease in the mean SD resulting in a positive correlation between the CV and the mean RT ($r_{\text{CV-RT}}$). The disproportionately lower CV is attributed to the restructuring underlying knowledge components that change, or eliminate altogether, resource-demanding processes (Segalowitz & Segalowitz, 1993; Segalowitz, Segalowitz & Wood, 1998; Segalowitz & Hulstijn, 2005). The performance in a given task not only gets faster, there is less variability in responses. The variability in performance at early phases of development is largely due to these effortful processes, and as they change variability decreases. In contrast, in instances where performance speeds up but the amount of relative variability (the CV) remains the same, it is assumed that the same underlying processes are involved, with response differences presumably due to other factors e.g., individual differences in base information processing rates or variability in attention across trials.

The proposed connection between differences in the CV and the development of automaticity in the L2 has attracted the interest of SLA researchers (Akamatsu, 2001; Fukkink, Hulstijn, & Simis, 2005; Schoonen, van Gelderen, de Glopper, Hulstijn, Simis, Snelling, & Stevenson, 2003), and is the focus of this paper.

2. The CV in L2 lexical processing.

Empirical evidence for the speed up and automaticity distinction was first presented in Segalowitz & Segalowitz (1993). Response time performance by French learners of English as a foreign language (EFL) was compared on two tasks, a simple stimulus detection task and an English lexical decision task. The detection task required subjects to indicate when a shape appeared on a computer screen. In this task RTs and SDs differed across trials to a small degree, but the differences between the mean RTs and SDs were proportional, resulting in an unchanged CV across trials. In the lexical decision task subjects were tested on familiar L2 vocabulary at multiple times. In the main analysis subjects were divided into Fast and Slow responders on the basis of overall response times. The Fast responders had lower mean RTs and CVs and a significant $r_{\text{CV-RT}}$ at both initial and final presentations, the latter result taken as evidence of emerging automaticity in these individuals. In contrast, the Slower responders initially showed no correlation $r_{\text{CV-RT}}$, with a significant correlation and, it was argued, automaticity emerging only at the final test. See Table 1. For both groups performance on the lexical decision task contrasted with that on a simple stimulus detection task, where faster response times were not accompanied by lower CVs.

The findings have been replicated in studies examining group performance on a lexical decision task (Segalowitz, Watson & Segalowitz, 1995, Segalowitz, et al. 1998; Akamatsu, 2001; Fukkink, Hultsijn & Simis, 2005), L2 lexical development in a single individual over time (Segalowitz, Watson & Segalowitz, 1995), the effect of study abroad on learners (Segalowitz & Freed, 2004), bilingual priming effects (Phillips, Segalowitz, O'Brien, & Yamasaki, 2004) and attention-switching processes in L2 performance (Segalowitz, Poulsen & Segalowitz, 1999: Segalowitz & Frenkiel-Fishman, 2005). To generalise these findings, better performance is characterised by faster mean RTs, smaller SDs, and smaller CV values. For most of the studies a significant correlation between RT and CV in complex task performance also accompanied better performance both between and within the L2 groups examined. This correlation was taken as evidence of automaticity (Segalowitz & Hulstijn, 2005). An exception to this is Akamatsu (2001) who examined performance by Japanese university subjects on a English lexical decision test consisting of high and low familiarity items. No significant $r_{\text{CV-RT}}$ was evidentin performance on the high familiarity items but it did appear in the low familiarity items. The author attributed this pattern of results to the subjects having fully automatised the access processes for the high familiarity words, but not for the low familiarity ones. In terms of the automaticity account, the underlying changes had been completed for the high familiarity words, which were recognised quickly and with less variability. Given that the high-familiarity items used in the study were monosyllabic words regularly encountered in English texts in Japan (e.g.,. "yes", "such", "like") the fast, low variability responses are not unexpected. For discussions of the Akamatsu results see Segalowitz & Hulstijn (2005) and Fukkink, et al. (2005, p 58).

3. The $\mathbf{r_{CV\text{-}RT}}$ as an index of L2 processing: A three-phase account.

The findings suggest a developmental continuum in which the CV and $r_{\text{CV-RT}}$ relationship index the onset, emergence and ultimate attainment of automaticity. Three phases of RT-CV variability in the development of automaticity in the L2 are schematised in Figure 1. See Segalowitz & Segalowitz (1993, p 375) for a similar account. Performance at Phase 1, the *Controlled Processing* Phase, is characterised by relatively slow responses and no significant $r_{\text{CV-RT}}$. At Phase 2, the *Automatizing* Phase, responses are faster and a significant correlation emerges, presumably due to the restructuring of the underlying processes. This shift is evident in the Slower responders in Segalowitz & Segalowitz (1993) and Segalowitz, et al. (1998), for the low frequency items in Akamatsu (2001), and for the trained items in

Study 1 in Fukkink et al (2005). Performance becomes fully automatized in Phase 3, the *Automatic Processing* Phase, as floor is reached in response times and the significant r_{CV-RT} disappears, as for the high familiarity items in (and only in) Akamatsu (2001).[1]

The study presented below evaluates the three-phase account by examining performance by individuals at three levels of English proficiency (Intermediate L2, Advanced L2 and L1 controls) on an English lexical decision task in which item difficulty (as reflected in frequency of occurrence) is systematically varied. Unlike the previous studies, where the subjects were homogeneous and the test items familiar, the research here examines changes in the CV and CV-RT relationship across levels of proficiency and item familiarity. By varying proficiency and item familiarity, the current study will directly assess the effect of these factors on the CV measures. The study attempts to extend the earlier findings and evaluate the CV, as it co-varies with Accuracy and RT measures, as a possible index of proficiency in L2 lexical processing. A better understanding of how the CV relates to proficiency, that is, the accuracy and speed of L2 performance, will have potential implications for the use of the CV in research and testing beyond the laboratory. The findings will also be relevant to the automaticity/restructuring account put forth by Segalowitz et al.

The main focus of the study is on how well the CV serves to discriminate between proficiency and frequency levels in lexical decision performance. Prior to examining these issues however, it is necessary to establish that RT (and hence the CV) reflects a processing dimension that is negatively correlated with accuracy, that is, there is no tradeoff between speed and accuracy in performance (Sternberg, 1998). The first research question the study will answer is thus,

(1) *Do accuracy and Response time measures reflect systematic changes in performance by levels of proficiency and item familiarity?*

Question 1 seeks to establish that RT and Accuracy performance are negatively correlated dimensions of L2 lexical knowledge. This is particularly important for the two ESL groups, where evidence for the strategic allocation to speed or accuracy in performance would make the RT (and CV) result difficult to interpret (Sternberg, 1998). Questions 2 and 3 evaluate the CV and r_{CV-RT} measures as indices of L2 lexical processing.

(2) *How well do the RT and CV measures discriminate between proficiency and word frequency levels?*

The CV will be compared with mean RT and Accuracy measures across proficiency and familiarity levels.. Of interest is whether the CV values systematically reflect differences in proficiency (low CV = high proficiency) and item familiarity as reflected in word frequency (low CV = high familiarity).

(3) *Do changes in r_{CV-RT} reflect a developmental continuum?*

The three-phase account predicts a changing RT-CV relationship as proficiency develops and/or items become easier. No correlation (Phase 1) is expected for lower proficiency subjects and performance on less familiar items (e.g., performance by Intermediate ESL subjects on all but the highest familiarity items, or by Advanced ESL subjects on low familiarity items). A significant r_{CV-RT} will emerge as proficiency increases and/or difficulty decreases (e.g., Advanced ESL performance on more familiar items, English L1 controls on low familiarity items) and will then disappear as RT responses are fastest and variability minimal (e.g., English L1 controls on high familiarity items). Evidence for this would be consistent with Akamatsu (2001) and indicate that the CV-RT relationship can provide a quantitative index of processing skill development that approximates a u-shaped curve.[2] The lack of evidence for Phase 3 would leave open what conditions are necessary for reaching full skilled processing. Although the lexical decision task is complex, it is far more constrained than sentence completion, grammaticality judgments, and sentence and discourse comprehension. If Phase 3 is not evident in the lexical decision performance examined here, it seems unlikely to appear in more complex tasks.[3]

4. The study.

Method

Participants. Data were collected from 110 participants from groups representing three proficiency levels. The first group consisted Intermediate ESL learners (n=32) studying at an English language institute at an Australian university. The learners were from mixed Asian L1s and drawn from classes at the same level as set by the institute's placement test (5th out of 7 levels, with Level 7 corresponding to the threshold for entrance to tertiary study in Australia). The Intermediate group participated as part of a class activity. The second group were Advanced ESL learners (n=36) also from mixed Asian L1s who were undergraduate and graduate students studying at the same university, as were the third group of English L1 speakers (n=42). Groups 2 and 3 participated for course credit as part of an introductory linguistics course.

Materials. The 150-item lexical decision task contained 90 words and 60 pseudowords. The 90 words consisted of 18 items from each of four frequency of occurrence bands (2000 most frequent (2K), 3000 (3K), 5000 (5K) and 10,000 (10K) words, and from the Academic Word list (Coxhead, 2001). The test items were drawn from the Vocabulary Levels Test, a standard test of receptive L2 vocabulary (Schmitt, Schmitt & Clapham 2001). The pseudowords for each level were generated from words at the same frequency level to ensure that the length of the pseudowords approximately mirrored that of the words.[4] The word items consisted of 18 words per the four frequency levels, 18 words from the Academic Word List. The items were presented in two blocks of 75 items each (36 target words, 30 pseudowords and 9 words from an Academic Word list). Presentation of items within each block was randomised for each individual. See Mochida & Harrington, 2006 for details.

Procedure. Test items were presented individually on a computer screen and participants asked to judge as quickly and as accurately whether they knew the word. They were told they would see items that were words and pseudowords, the latter possible words in English. They were warned that they might be tested on some of the words later, following the instruction set used by Eyckmans (2004, p 96). In fact they were not tested at the end. On each trial the participant first saw a blank screen with a star. After a 1500 msec interval a word or pseudo word appeared on the screen and the participant responded "Yes" or "No" by pressing the appropriate key on the keyboard. No feedback was given. The word appeared on the screen for only 5000 milliseconds. If the participant failed to answer in the allotted time a 'Not answered' was recorded as treated as a Miss in the subsequent data analysis. There were very few 'Not answered' responses, representing less than 1% of the total number of responses for any the groups. A practice set of five items was completed before the test.

Scoring and statistical analysis. The raw hit scores ('yes' responses to words) were corrected for guessing using the correction formula described in Huibregtse, Admiraal, & Meara (2001).[5] Test scores were calculated for each level (2K, 3K, 5K, and 10K) and for Overall performance. Items from the Academic Word List were not included in the analysis, as frequency of occurrence was the basis for defining item familiarity. The AWL is drawn from academic sources and is comprised of words drawn from different frequency levels (Coxhead, 2001). The RT data were screened for outliers. Responses 2.5 standard deviations from the individual's mean response time were replaced with values at 2.5 standard deviations. This adjustment affected less than 2% of the total number of responses. Two-way analyses of variance were then carried out for the Accuracy, RT and CV results for subjects and items. Where appropriate, follow up one-way analyses of variance with were done on the Accuracy and RT measures separately, with frequency level treated as the repeated measure.

Finally, correlations between the RT and CV scores for Group and Level were calculated. The statistical results are preliminary. Please do not cite without permission.

5. Results

Descriptive statistics: Accuracy. The accuracy results are presented in Table 2. Mean proportions and standard deviations for the false alarm rate (proportion of 'yes' responses to pseudowords), hits (proportion of 'yes' responses to words), and corrected scores are given by group and word frequency levels. Reliability measures (Cronbach's alpha) were calculated for the three groups: Intermediate ESL = .68; Advanced ESL = .74; and English L1 controls = .81. The reliability value for the Intermediate group is low, which makes the interpretation of results somewhat qualified.

The Intermediate ESL Group falsely recognised around 10% of the pseudowords as words. Both the Advanced ESL and the L1 English groups had mean false alarm rates of less than 5% overall, that is, incorrectly identifying less than three pseudowords out of the 60 presented.[6] The high standard deviations, especially for the Intermediate ESL group, indicate considerable individual variability in the responses.

The accuracy means discriminate performance by the three groups and across the word levels within groups. Corrected scores for the Intermediate ESL group ranged from .81 for the 2K level to .24 at the 10K level. The performance of the Advanced ESL group approached ceiling for the 2K and 3K words, with performance falling off sharply at the 5K and 10K levels. The English L1 group performed near ceiling for the 2K, 3K and 5K level, with a decline in performance at the 10K level.

Descriptive statistics: Response times and the coefficient of variance. The Response time data were screened for outliers. Responses more than 2.5 SDs beyond individual mean RTs were replaced with the value at the 2.5 SD point. This affected less than 3% of the data across the groups. The descriptive statistics for the Response time measures for correct word responses and the coefficients of variance (CV) are given in Table 3. Reliability measures for the three groups were, Intermediate ESL = .61; Advanced ESL = .86; and English L1 controls = .87. Again the reliability coefficient for the Intermediate ESL group was low, requiring caution in interpretation. Response times for the Intermediate ESL Group ranged from 1184 msec (CV = .35) at the 2K level to 1986 msec (.42) at the 10K level. See Table 2. For the Advanced ESL group, mean RTs went from 761 (.26) on the 2K items to 1205 (.39) msec on the 10K words, comparable to the findings in Segalowitz & Segalowitz (1993) for the Fast group and the Slow groups, respectively. In both those studies only a basic set of familiar vocabulary items was used. The English L1 controls had a mean RT of 733 (.24) which was at the upper range of L1 English subjects reported in Ratcliff, Gomez & McKoon (2004) and higher than responses on a similar task reported in Muljani, Koda, & Moates (1998).

Relationship between speed and accuracy. The interpretation of RT differences is potentially compromised by a strategic trade off by individuals in the speed and accuracy of response. There was a significant negative correlation between accuracy and RT for responses across all levels (n =110): 2K, *r* = -.615; 3K, -583; 5K, -.575; 10K, -.621, all significant at p < .000. The inverse correlations indicate that response time and accuracy are measuring a similar underlying proficiency, with little discernable tradeoff between speed of response and accuracy evident.

Deleted: ¶

Deleted: reaction time

Deleted: reaction time

Analysis of variance Accuracy. The corrected scores were analysed in a mixed two-way analysis of variance for subjects and items. Group was the between subjects factor (Intermediate ESL x Advanced ESL x English L1) and Frequency Levels as the repeated measures factor (2K x 3K x 5K x 10K). As sphericity assumptions were not met for the three measures, all *p* results reflect the Greenhouse-Geiser correction. Uncorrected degrees of freedom are reported. For subjects both main effects were significant. For Group, *F1* (2,107) = 162, p < .000, partial η^2 = .751, and for Level, F_1 (3, 321) = 498.86, p < .000, partial η^2 = .823. Pairwise comparisons for Group and Level differences showed that the mean differences were significant at *p* < .05, with a Bonferroni adjustment made for multiple comparisons. There was also a significant Group x Level interaction, F_1 (6, 321) = 59.67, p < .000, partial η^2 = .485. The same pattern of results was evident in the item analysis. As items were uniquely assigned to levels, a two-way analysis of variance was carried out with Group and Level as between-subject factors. Group was significant at F_2(2,204) = 65.82, p < .000, partial η^2 = .391, and Level, F_2(3, 204) = 101.37, *p* < .000, partial η^2 = .598. The Group x Level interaction was also significant, F_2 (6,204) = 7.97, *p* < .000, partial η^2 = .190. As was the case in the subject analysis, the interaction was the result of no significant difference between 2K and 3K level means for the English L1 controls.

The significant Group x Level interaction was examined in separate one-way ANOVAs done on the each Group. Level was the within group, repeated measure. For the Intermediate ESL group, *F*(3,93) = 89.90, *p* < .000, partial η^2 = .744. Pairwise comparisons showed that all four mean differences were significant. The same pattern was observed for the Advanced ESL group, *F*(3,105) = 188.81, *p* < .000, partial η^2 = .844. Pairwise comparisons again showed all the level mean differences to be significant. The observed Group x Level interaction was evident in the English L1 data. The overall effect for Level was *F*(3,123) = 54.89, *p* < .000, partial η^2 = .572, but pairwise comparisons showed no significant difference between Accuracy at the 2K and 3K levels, where performance was at ceiling.

Analysis of variance: Response time. The RT results mirrored the Accuracy responses. Group was significant at F_1(2,107) = 191.80, *p* < .000, partial η^2 = .782, and Level, F_1(3, 321) = 120.56, *p* < .000, partial η^2 = .530. Pairwise comparisons indicated that the mean differences for Group and Level were significant. There was also a significant Group x Level interaction, F_1(6, 321) = 18.53, *p* < .000, partial η^2 = .257. Both main effects and the interaction were significant in the item analysis. For Group, F_2(2,204) = 169.26, *p* < .000, partial η^2 = .624, and Level, F_2(3, 204) = 39.13, *p* < .000, partial η^2 = .365, and Group x Level, F_2(6,204) = 7.97, *p* < .01, partial η^2 = .107.

The follow-up one-way ANOVAs yielded results similar to the Accuracy findings. For the Intermediate ESL group, *F*(3,93) = 42.58 *p* < .000, partial η^2 = .579, and for the Advanced ESL group, *F*(3,105) = 43.64, *p* < .000, partial η^2 = .555. Pairwise comparisons showed the four levels were significantly different for both groups. For the English L1 group, *F*(3,123) = 76.3 , *p* < .000, partial η^2 = .65. Levels 2K and 3K were not significantly different but the other differences were.

Coefficient of variance. The CV measure was less sensitive than the accuracy and RT results. Group was significant at F_1(2,107) = 53.17, *p* < .000, partial η^2 = .498, and Level, F_1(3, 321) = 12.73, *p* < .000, partial η^2 = .106. Pairwise comparisons showed that the mean differences for Group and Level were significant. The Group x Level interaction was significant, F_1(6, 321) = 2.32 *p* < .05, partial η^2 = .042. For items, Group was significant at F_2(2,204) = 38.51.26, *p* < .000, partial η^2 = .274, and Level, F_2(3, 204) = 54.42.13, *p* < .000, partial η^2 = .145. With the exception of the 2K and 3K levels, pairwise differences were significant. The Group x Level interaction was also significant, F_2(6,204) = 5.7, *p* < .000, partial η^2 = .145. This was again due to the Advanced ESL and English L1 data, where there were differences between the 2K and 10K results.

The follow-up one-way ANOVAs showed marked differences among the groups. The Intermediate ESL group showed no significant difference between the level means, *F*(3,93) = 2.18, *n.s.* In contrast, the CV discriminated between levels for the Advanced ESL group, F(3,105) = 9.127, p < .000, partial η^2 = .8207. This effect was due to the CV for the 2K level being significantly different from all the other levels. There were no reliable differences between the other level means. There was also a small but significant effect for Level in the English L1 data, $F(3,123) = 4.87$, $p < .005$, partial $\eta^2 = .106$, due to a significant difference between the 2K and 10K levels alone.

The $r_{CV\text{-}RT}$. The correlations between RT and CV for each group and level combination are given in Table 4. The Intermediate ESL group only showed a positive correlation for the 2K items, and the most difficult level, the 10K actually yielded a negative correlation. For the Advanced ESL group there was a significant positive correlation for 2K, 3K and 5K, but not the 10K level. The English L1 controls showed a significant correlation at all levels although the 10K level just met significance at $p < .05$. There was no evidence for a weakening correlation between RT and CV as performance improved. Performance on the 2K items, where responses should be fastest and the least variable, was similar to that at the other levels.

Summary of results

Question (1) Accuracy and RT performance. Accuracy and RTs performance consistently improved with increasing proficiency and item familiarity. The Intermediate ESL group was less accurate and slower than the Advanced ESL group, who in turn were less accurate and slower than the English L1 controls, all the mean differences were statistically significant. There were interactions between group and item familiarity levels for the accuracy and RT responses, due to ceiling performance by the English L1 controls on the high frequency items. Most importantly, there was little discernable trade off in speed and accuracy meaning that that RT results can be considered an independent dimension of L2 development. The reliability of the RT responses by the Intermediate ESL group was low.

Question (2). The CV as an index of L2 development. The CV measure discriminated between the three proficiency levels and to a lesser extent between frequency levels. The L1 group had the lowest CVs and there was a significant difference in CV between the 2K (CV = .21) and 10K (.26) levels. The mean RT was a more sensitive discriminator of proficiency and item familiarity level than the CV. This is most evident for the Intermediate L2 group, where mean RTs were significantly different for each frequency level, while there was no difference between frequency levels for the CV. For the Advanced L2 group as well, the CV separated performance on the 2K from that on the other levels, while RTs discriminated between all four levels. For the English L1 group the only significant difference was between 2K and 10K, for the CV values identified two levels of performance, 2K/3K/5K and 10K, while the RT discriminated between three levels, 2K, 3K/5K, and 10K.

Question (3) A $r_{CV\text{-}RT}$ continuum for lexical processing. There was no u-shaped curve in the RT-CV correlation for either proficiency or item difficulty levels. Significant RT-CV correlations were only evident at the 2K level for the Intermediate ESL, while for they were present for the 2K, 3K and 5K levels for the Advanced ESL group. The English L1 controls had significant RT-CV correlations at all four levels of frequency. Thus the presence or absence of the correlation served as better discriminators of performance between levels than the CV measure alone. The results are in contrast to Akamatsu (2001) who reported no RT-CV correlation for L2 subjects on the high frequency words used in that study. One oddity was the significant *negative* correlation for Intermediate ESL learners on the 10K items. This is probably due to the low overall accuracy on these items (30%). Although only correct responses were used for the RT analysis, the relative difficulty of the 10K items likely affected the pattern of correlations. The interpretation of Response time data in tasks with high error rates is taken up below.

Overall the $r_{\text{CV-RT}}$ pattern was consistent with earlier findings. Significant positive correlations were evident only for the Advanced L2 and English L1 subjects, and not evident in the former for the least familiar items.

6. Discussion

The CV and the $r_{\text{CV-RT}}$ were examined here as potential indices of lexical processing. First the $r_{\text{CV-RT}}$ results will be discussed and then the CV as measure of L2 lexical processing will be addressed. A three-phase account of skill development was tested in which the decrease in RTs and CVs were predicted to vary in a u-shaped manner as a function of proficiency and item familiarity. In the first phase no significant $r_{\text{CV-RT}}$ is evident, as faster RTs are not accompanied by decreasing CVs. The correlation does appear in the second phase as performance improved. However, contrary to expectations, there was no evidence for the third phase, in that there was still a significant correlation for the fastest performance. If full automaticity were to emerge it was most likely to occur in the English L1 responses on the 2K items, but even here the $r_{\text{CV-RT}}$ was moderately strong (r = .47) and significant. The results here contrast with Akamatsu (2001) who reported evidence of ceiling performance by Japanese university students on the high familiarity items

Although the three-phase account received little support, the $r_{\text{CV-RT}}$ results are consistent with the earlier findings as a significant $r_{\text{CV-RT}}$ did emerge as proficiency and item familiarity increased (Segalowitz & Segalowitz, 1993; Segalowitz et al 1998). However, neither set of results requires the original automaticity/restructuring account. The emergence of a significant $r_{\text{CV-RT}}$ may or may not be the result of the modification of the underlying processes such that the more controlled, resource-demanding components are either removed or their contribution to response variability greatly reduced (Segalowitz et al 1998, p54-55). In order to evaluate this claim the mechanisms modified in the course of improving lexical decision performance need to be specified.

Improved word recognition in the English lexical decision task may be the result of a qualitative shift in processing from dependence on both orthographic and phonological codes, to direct access to semantic knowledge through orthographic information alone (Coltheart, Rastle, Perry, Langdon & Ziegler, 2001). Here the phonological component would the element that “drops out” (Segalowitz & Segalowitz, 1993, p374), with an accompanying reduction in variability. Alternatively, the significant $r_{\text{CV-RT}}$ may be due to faster items being recognised more quickly as the result of the development of stronger underlying knowledge representations. In this case the increasing strength of visual and phonological information interact and result in better performance (Muliani, Koda & Moates, 1998; Harm & Seidenberg, 2004). If the automatization of the lexical decision task is the result of a shift from phonological to direct access of visual information, a concurrent task paradigm in which the lexical decision task is performed while a concurrent phonological processing load is imposed should have relatively greater effect on the unautomatized low frequency items.

Alternatively, the $r_{\text{CV-RT}}$ may be better understood as a reflection of relative task complexity as it interacts with individual skill development than as an indicator of the emergence of automaticity per se. As the task in question starts to resemble natural language processes, that is, incorporates a larger number of elements that vary in familiarity, then the observed correlation may be evident even when the individual approaches optimal performance. Response times and variability decrease may decrease in a given task, but that decrease varies within and across individuals. The $r_{\text{CV-RT}}$ characterises this improvement but is neutral to the underlying factors responsible.

A final issue with the $r_{\text{CV-RT}}$ is statistical. The coefficient is a correlation between the RT and a ratio (SD/RT) that includes the RT as one of its terms. Segalowitz et al (1998) note this problem and argue that the potentially circular nature of the account is avoided by showing that the correlation is not a natural consequence of lower RTs and SDs, that is, that a speed-up in RTs and SDs can occur without a significant $r_{\text{CV-RT}}$, as in stimulus identification task in Segalowitz & Segalowitz (1993). As suggested above, the $r_{\text{CV-RT}}$ may simply be a quantitative measure of improvement in complex task performance that is neutral to underlying causes

The CV as an index of processing .The automaticity/restructuring account attempts to give theoretical significance to a measure that usually has a descriptive function, that is, as a measure of response variability that controls for absolute Response time. A number of questions have been raised concerning the usefulness of $r_{\text{CV-RT}}$ as an index of qualitative change in the underlying skill development. However, regardless of how these issues are resolved, the $r_{\text{CV-RT}}$ and CV may still have promise as measures of L2 processing proficiency.

On-line processes are becoming an increasingly important element in SLA theory (Carroll, 2001; Piennemann, 2002; Sharwood-Smith & Truscott, 2005). As processing becomes more important the need for reliable and valid performance measures will also grow (Marinis, 2003). RT difference in individual and task performance will continue to be primary data, but as on-line measures of L2 processing become more sophisticated (e.g. greater use of ERP, eye tracking), differences in response variability should also become increasingly informative for theory and model building (Felser, 2005).

Decreasing CV values (with faster RTs) provides evidence of increased fluency that is not directly available from a comparison of mean RT and SD values alone. The CV provides a means to compare the development of processing skill across tasks and domains that vary in complexity, and hence the time they require to complete. A trial in a timed grammaticality judgments or sentence completion task takes longer than a lexical decision and meaningful comparisons of relative processing efficiency are not possible with mean RTs and SDs. The ability to compare performance across tasks, e.g., lexical decision and sentence processing, may provide a window on how the development of processing efficiency in one domain interacts with that in another (Fender, 2002).

Another question for further research concerns whether quantitative thresholds can be fixed for CV values for the development of L2 processing efficiency. Empirical findings to date, the present study included, suggest that fluent performance (L1 or advanced L2) is characterised by a CV of around .20 -.22. To the extent that relatively fixed thresholds can be established for specific processing tasks, the CV may be incorporated in models of L2 proficiency. The findings from the current study indicate that the presence of a significant $r_{\text{CV-RT}}$ was a better discriminator of performance between levels within the groups than the CV measure alone, and more needs to be known about how the CV and correlation values interact in development.

The L2 testing literature has traditionally characterised proficiency in terms of what the learner knows, as reflected in response accuracy.[7] The incorporation of the temporal dimension, particularly variability, in the assessment of proficiency may have far-reaching implications for measurement in testing and research. The focus on variability itself is new to SLA research, but there is evidence elsewhere that response time variability can be an isolable and informative aspect of skill development - or loss. Research in aging suggests that aging effects in the speed and efficiency of cognitive processing may be expressed primarily through RT variability rather than RT means (MacDonald, Hultsch & Dixon, 2003). The implications of this line of research for SLA remain to be explored.

Measuring processing time. The measurement of processing time and variability in SLA also presents methodological challenges. The application of Response time methodology to SLA processes must accommodate the fact that L2 performance is typically more variable than the behaviour of the individuals (usually normal L1 subjects) on which the methodology was developed. Response time performance is usually examined in contexts where error rates are minimised in order to control possible effects that knowledge differences might have on RT performance (Sternberg, 1998). Whether RT differences are informative in contexts where knowledge is developing – the results here suggest they are –

need to be examined more carefully. At issue is whether RT differences are informative only for performance on material already learned (i.e., practice effects), or can they characterize the development of that knowledge. When response times are examined in task with higher error rates, as in the data here, it is important to ensure that systematic trade-offs in speed and accuracy are not occurring. Even if no obvious trade-offs are evident, high error rates still present problems. The Intermediate ESL subjects here answered only 30% of the 10K items correctly, resulting in the RT analysis being done on only a small subset of the original stimuli. This poses problems both for statistical power and for the representativeness of the correct items as members of the category.

The variability in response times can also be due to factors unrelated to the L2 processes of interest. Individual differences in base information processing rates (Faust, Balota, Spieler, & Ferraro, 1999) and L1-specific processing proficiency may also obscure experimental effects. Existing differences in L1 processing efficiency can be particularly important and need to be identified and controlled (as, for example in Segalowitz & Frenkiel-Fishman, 2005).

Fluency in an L2 depends both on what the individual knows and how efficiently that knowledge can be accessed in real time. The results here show that direct…..As the later comes to play an increasingly central role in SLA theory, measures of processing efficiency like the CV will become increasingly valuable.

References

Akamatsu N 2001 'Effects of training in word recognition on automatization of word recognition processing of EFL learners' A paper presented at the annual meeting of the American Association of Applied Linguistics St Louis

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

Carroll S E 2001 *Input and evidence: the raw material of second language acquisition* John Benjamins: Amsterdam/New York

Coltheart M Rastle K Perry C Langdon R & Ziegler J 2001 'The DRC model: A model of visual word recognition and reading aloud' *Psychological Review* 108 204 - 258

Coxhead A 2000 'A new academic word list' *TESOL Quarterly 34* 213-238

De Groot A M B D & Keijzer R 2000 ,What is hard to learn is easy to forget: The roles of word concreteness cognate status and word frequency in foreign-language vocabulary learning and forgetting' *Language Learning 50* 1-56

DeKeyser R.M. 2001 'Automaticity and automatization.' in P. Robinson (ed.) *Cognition and second language instruction*. Cambridge University Press Cambridge. p. 225-251.

Dijksta T. 2005. 'Bilingual visual word recognition and lexical access' in F. F. Kroll and AMB De Groot (ed.) pp 179-201 *Handbook of bilingualism: Psycholinguistics approaches* Oxford: Oxford University Press

Eyckmans J 2004: *Measuring receptive vocabulary size* Utrecht: LOT

Faust M E Balota D A Spieler D H & Ferraro F R 1999 Individual differences in information-processing rate and amount: Implications for group differences in response latency *Psychological Bulletin 125* 777-799

Felser C 2005 Experimental psycholinguistic approaches to second language acquisition *Second Language Research 21* 95-97

Fender M J 2001 A review of L1 and L2/ESL word integration development involved in lower-level text processing *Language Learning 51* 319-396

Fukkink R G Hulstijn J and Simis A 2005 Does training in second-language word recognition affect reading comprehension? An experimental study *Modern Language Journal 89* 54-75

Harm M & Seidenberg MS 2004 Computing the meanings of words in reading: Cooperative division of labour between visual and phonological processes *Psychological Review 111* 662–720

Huibregtse I Admiraal W and Meara P 2002 Scores on a yes-no vocabulary test: correction for guessing and response style *Language Testing 19* 227-245

Kempe V & MacWhinney B 1996 The crosslinguistic assessment of foreign language vocabulary learning *Applied Psycholinguistics 17* 149-183

Koda K 1996: L2 word recognition research: A critical review *Modern Language Journal 80* 450-460

Kroll J F & Tokowicz N 2001 The development of conceptual representation for words in a second language In J L Nicol & T Langendoen Eds *Language processing in bilinguals* Blackwell Cambridge MA 49-71

Logan G 1988 Toward an instance theory of automatization *Psychological Review 95* 492-527

MacDonald S W S Hultsch DF & Dixon RA 2003 Performance variability is related to changes in cognition: Evidence from the Victoria longitudinal study *Psychology and Aging 18* 510-523

Marinis T 2003 Psycholinguistic techniques in second language acquisition research *Second Language Research 19* 144-161

Mochida A & Harrington M 2006 The Yes-No test as a measure of receptive vocabulary knowledge To appear in *Language Testing*

Muljani M Koda K & Moates D 1998 Development of word recognition in a second language *Applied Psycholinguistics 19* 99-113

Phillips N A Segalowitz N O'Brien I & Yamasaki N 2004 Semantic priming in a first and second language: evidence from Response time variability and event-related brain potentials *Journal of Neurolinguistics 17* 237-262

Pienemann M 1998 *Language processing and second language development: Processability theory* Amsterdam: John Benjamins

Pienemann M 2002 Issues in second language acquisition and processing *Second Language Research 18* 189-192

Ratcliff R Gomez P & McKoon G 2004 A diffusion model account of the lexical decision task *Psychological Review 111* 159-182

Schmitt N Schmitt D and Clapham C 2001 Developing and exploring the behaviour of two new versions of the vocabulary levels test *Language Testing 18* 55-88

Schoonen R van Gelderen A de Glopper K Hulstijn J Simis A Snelling P Stevenson M 2003 First and second language writing: The role of linguistic knowledge speed of processing and metacognitive knowledge *Language Learning 53* 165-202

Segalowitz N 2003 Automaticity and second languages In C Doughty & MH Long Eds *The handbook of second language acquisition* Blackwell: Oxford 382-408

Segalowitz N & Frenkiel-Fishman N 2005 Attention control and ability in complex cognitive skill: Attention-shifting and second language proficiency *Memory & Cognition* xx

Segalowitz N & and Freed B 2004 Context contact and cognition in oral fluency acquisition: learning Spanish in "At Home" and "Study Abroad" contexts *Studies in Second Language Acquisition*

Segalowitz N and Hulstijn J 2005 Automaticity in bilingualism and second language learning In F F Kroll and AMB De Groot Ed p 371-388 *Handbook of bilingualism: Psycholinguistics approaches* Oxford: Oxford University Press

Segalowitz N Poulsen C & Segalowitz S 1999 RT coefficient of variation is differentially sensitive to executive control involvement in an attention-switching task *Brain & Cognition 38* 255-258

Segalowitz N and Segalowitz S J 1993 Skilled performance practice and differentiation of speed-up from automatization effects: evidence from second language word recognition *Applied Psycholinguistics 13* 3 369-385

Segalowitz N Segalowitz S J & Wood A G 1998 Assessing the development of automaticity in second language word recognition *Applied Psycholinguistics 13* 3 369-385

Segalowitz N Watson V and Segalowitz SJ 1995 Vocabulary skill: single case assessment of automaticity of word recognition in a timed lexical decision task *Second Language Research 11* 2 121-136

Seidenberg M & McClelland J 1989 A distributed developmental model of word recognition and naming *Psychological Bulletin 114* 510-532

Sharwood-Smith M & Truscott J 2005 Phases or continua in second language acquisition: A MOGUL solution *Applied Linguistics 26* 219-240

Siakaluk PD Sears CR and Lupker SJ 2002 Orthographic neighborhood effects in lexical decision: The effects of nonword orthographic neighborhood size *Journal of Experimental Psychology: Human Perception and Performance 28* 661-681

Snelling P Van Gelderen A De Glopper K 2002 Lexical retrieval: An aspect of fluent second language production that can be enhanced *Language Learning 52* 723-754

van Gelderen A Schoonen R De Glopper K Hulstijn J Simis A Snelling P & Stevenson M 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 Heuven W J B Dijkstra T and Grainger J 1998 Orthographic neighborhood effects in bilingual word recognition *Journal of Memory and Language 39* 458–483

Sternberg S 1998 Inferring mental operations from Response time data: How we compare objects In D Scarborough & S Sternberg Eds *An invitation to cognitive science Volume 4: Methods models and conceptual issues* pp 436-440 Cambridge MA: MIT Press

Tokowicz N & MacWhinney B 2005 Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation *Studies in Second Language Acquisition 27*2 173-204

Wang M & Koda K 2005 Commonalities and differences in word identification skills among learners of English as a second language *Language Learning 55*1 71-98

Slower response time More variability No RT-CV correlation	Faster response times Less variability Emerging RT-CV correlation	Fastest response times Minimal variability No RT-CV correlation
Controlled →	Automatizing →	Automaticity

Figure 1. Phases of development and the correlation between Response time and the coefficient of variance.

Table 1. Mean Response Time, Coefficient of Variance, Correlation of Response Time and Coefficient of Variance in Selected Studies

			Mean RT	Mean SD	Mean CV	RT-CV
			(msec)	(msec)	(SD/RT)	correlation
Segalowitz & Segalowitz (1993)		Stimulus detection task				
	(n=66)		264	54	.20	.19.
Lexical decision task						
Fastest	(n=22)	Overall	745	nr	.23	.55*
		Initial/Final	nr/nr	nr/nr	nr/nr	.67*/.67*
Slowest	(n=22)	Overall	1203	nr	.42	.20
		Initial/Final	nr/nr	nr/nr	nr/nr	.17/.51*
Entire Group	(n=64)	Overall	948	324	.32	.72*
		Initial/Final	880/750	nr/nr	nr/nr	.43*/.67*
Segalowitz, Segalowitz, & Wood (1998)		Lexical decision task				
Fastest	(n=50/39)	Initial/Final	660/660	nr/nr	.28/.28	.53*/.62*
Slowest	(n=40/32)	Initial/Final	846/765	nr/nr	.34/.30	.21/.65*
Entire Group	(n=90/71)	Initial/Final	742/707	nr/nr	.31/.29	nr/nr
Akamatsu (2001)		Lexical decision task				
High frequency	(n=49)	Pre/Posttest	871/785	113/85	nr/nr	.21/-.18
Low frequency		Pre/Posttest	1008/853	159/117	nr/nr	.52*/.42*
Fukkink, Hulstijn & Simis 2005		Lexical decision task (Experiment #1)				
Trained items		Pretest/Posttest	885/719	167/121	.33/.30	.40/.63*
Control items		Pretest/Posttest	931/839	163/123	.39/.19	.11/19
Phillips, Segalowitz, O'Brien & Yamasaki (2004)		Bilingual animacy judgment task				
L1 baseline Low proficiency			620	nr	.21	nr
L2 baseline Low proficiency			711	nr	.23	.75*[a]
Segalowitz & Frenkiel-Fishman (2005)		Bilingual animacy judgment task				
L1			748	nr	.34	nr
L2			940	nr	.44	.61*

nr = not reported in the study

* = *p* < .05

a = RT-CV correlation for all stimuli types

Table 2. False alarms, Hits and Corrected Accuracy Scores by Frequency Level and Group

		False alarms		Hits		Corrected score	
		M	*SD*	*M*	*SD*	*M*	*SD*
2000	Intermediate ESL	.10	.12	.89	.13	.80	.14
	Advanced ESL	.04	.06	.99	.04	.95	.06
	L1 English	.05	.09	.99	.02	.96	.10
3000	Intermediate ESL	.11	.11	.76	.15	.65	.15
	Advanced ESL	.06	.06	.94	.09	.88	.13
	L1 English	.04	.08	.98	.03	.94	.08
5000	Intermediate ESL	.12	.18	.44	.16	.43	.15
	Advanced ESL	.06	.07	.78	.18	.75	.17
	L1 English	.04	.08	.95	.07	.92	.07
10000	Intermediate ESL	.08	.14	.24	.16	.31	.13
	Advanced ESL	.02	.03	.42	.21	.52	.15
	L1 English	.03	.07	.81	.12	.80	.17
Overall	Intermediate ESL	.10	.15	.63	.15	.55	.14
	Advanced ESL	.04	.06	.82	.11	.78	.13
	L1 English	.04	.08	.95	.06	.91	.1

Table 3. Means and Standard Deviations for Response Time and Coefficient of Variance by Group and Frequency Level

		Response time (msec)		Coefficient of Variance	
		M	*SD*	*M*	*SD*
2000	Intermediate ESL	1184	176	.35	.11
	Advanced ESL	761	108	.25	.10
	L1 English	703	84	.20	.09
3000	Intermediate ESL	1450	226	.41	.08
	Advanced ESL	880	197	.34	.14
	L1 English	736	79	.23	.09
5000	Intermediate ESL	1679	415	.37	.13
	Advanced ESL	1003	238	.39	.14
	L1 English	757	76	.22	.09
10000	Intermediate ESL	1986	486	.42	.17
	Advanced ESL	1206	372	.39	.15
	L1 English	871	107	.26	.09
Overall	Intermediate ESL	1574	326	.39	.12
	Advanced ESL	962	229	.34	.13
	L1 English	769	109	.23	.09

Table 4. Pearson's Product Moment Correlation of RT and CV by Group and Level

Level	2K		3K		5K		10K	
Group	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Intermediate ESL (n = 32)	.47	.007	-.012	.949	.152	.405	-.49	.004
Advanced ESL (n = 36)	.42	.008	.76	.000	.46	.006	.15	.397
English L1 (n = 42)	.47	.002	.56	.000	.58	.000	.31	.048

[1] Phase 3 is expected to emerge in L1 performance on familiar material. However, the studies that did collect L1 and L2 data (i.e. Phillips et al, 2004; Segalowitz & Frenkiel-Fishman, 2005) used L1 performance to adjust for the L2 RT and CV measures, but did not give the relevant L1 RTs and CVs themselves. Segalowitz & Frenkiel-Fishman, (2005) reported a significant correlation between L1 and L2 RTs (r = .673, p = .004), but none for the L1 and L2 CVs, r = -.03, *n.s*. Given the subsequent L2 RT-CV correlation reported, (residualized on L1 measures) r = .608, p = .012, it appears there was no RT-CV for the L1s in this study.

[2] The $r_{\text{CV-RT}}$ is central for the automaticity/restructuring account, but its absence has two sources. In the case of the stimulus detection task there is no correlation due to a lack of variability in responses. According to the data presented in Figure 1 in Segalowitz & Segalowitz (1993, p. 377) the RT means ranged from 200 – 350 msec (excluding two pronounced outliers), compared with 550 – 15000 msec range reported for responses in the lexical decision task in the study. In the lexical decision task the lack of a significant $r_{\text{CV-RT}}$ in Phase 1 is due, as was the case for Slow responders in that study, to excessive variability resulting from slower RTS and larger SDs (due to the dependence by slow responders on multiple, resource-demanding steps in executing the task). The lack of absence of a correlation in Phase 3 is due to fast performances and lack of variability, the same as in the stimulus detection task.

[3] It would also suggest that the Akamatsu (2001) findings are due to factors (items or treatment) specific to the study.

[4] Mean (and standard deviation) letter length for the word levels: 2K = 6.3(1.7); 3K = 5.7(1.6); 5K = 6.9(1.7); 10K = 6.9(1.6); AWL = 8.4(2.0).

[5] ISDT = $1-\dfrac{4h(1-f)-2(h-f)(1+h-f)}{4h(1-f)-(h-f)(1+h-f)}$

[6] The error rate here compares with overall rates from the L1 literature of 5.3% for adult English subjects (Ziegler and Perry, 1998: B57); 12% for Dutch children (van Bon, Hoevenaars and Jongeneelen, 2004: 65), and 15% for English children and 21% for German children reported in Goswami, Ziegler, Dalton and Schnieder (2001: 654). Van Heuven, Dijkstra and Grainger (1998) reported a 6.2 error rate for their advanced Dutch EFL subjects, while the strict condition in Eyckmans (2004) yielded a rate of 8.7. The false alarm rate here contrasts with Beeckmans et al. (2001), Cameron (2001), and the minimal instruction condition in Eyckmans (2004), all of whom reported much larger false alarm rates, as did research discussed in Meara (1996).

[7] A recent review of L2 testing doesn’t even mention response time as a dimension of L2 proficiency (Norris & Ortega, 2003).