Part 2

Introduction

1.1 Overview

Part 1 introduced the lexical facility construct. Lexical facility combines size and processing skill as a unitary second language (L2) vocabulary skill construct. The challenges arising from combining the two was acknowledged, but the case was made for treating the two as a unitary construct, both due to the time-contingent nature of L2 vocabulary knowledge and the potential utility of combining knowledge and skill as a measurement tool to characterize individual and group differences in L2 proficiency and performance.

Part 2 provides empirical evidence for the account. It presents a set of studies that investigate the lexical facility measures (vocabulary knowledge, mean recognition time, and consistency) as reliable indices of individual differences in L2 vocabulary skill, separately and in combination (Chap. 6). The sensitivity of the measures to performance differences in selected domains of academic English performance is then examined. These domains consist of university entry standards (Chap. 7), performance on the International English Language Testing System (IELTS; Chap. 8), language program placement (Chap. 9), and general and academic English classroom performance (Chap. 10). A summary chapter that identifies the main findings is also included (Chap. 11). The data presented here are drawn from published and unpublished research by

the author and colleagues. In the final chapter (Chap. 12), the implications for L2 vocabulary teaching, learning, and testing are discussed, including the potential for incorporating time measures into models of L2 vocabulary acquisition and L2 theory more generally.

1.2 Aims of the Empirical Research

The research aims to establish lexical facility as a context-independent index of L2 vocabulary skill that is sensitive to performance differences in various academic English domains. The research program reported in Part 2;

1. *compares the three measures of lexical facility (VKsize, mnRT, and CV) as stable indices of L2 vocabulary skill;*
2. *evaluates the sensitivity of these measures individually and as composites to differences in a range of academic English domains; and, in doing so,*
3. *establishes the degree to which the composite measures combining the VKsize measure with the mnRT and CV measures provide a more sensitive measure of L2 proficiency differences than the VKsize measure alone.*

1.3 An Overview of Methods Used

The seven studies reported in Part 2 use the same approach to testing the lexical facility proposal, with the methodology adapted in minor ways to the particular study. These methods and their rationale are described here.

Each of the studies collected data for the lexical facility measures with the Timed Yes/No Test, described in Chap. 5. The test elicits a yes/no response on test items that yields an estimate of the individual's English vocabulary size and a measure of the speed and consistency with which these words are recognized. The tests items are both actual words and pseudowords—orthographically possible but nonexistent words interspersed randomly as a control for guessing. Word items are drawn from a range of frequency-of-occurrence bands, and the proportion of words recognized (hits) across the range provides an estimate of the individual's vocabulary size. The incorrect recognition of pseudowords as words (false

alarms) is used to adjust the final score. The computerized test format presents the test items individually in a randomized order for each test-taker. The test records test-takers' yes/no responses and the time they take to recognize each item. These responses are used to calculate individual and composite measures of lexical facility. These are described next.

Individual Measures

The test responses are used to calculate measures of vocabulary knowledge, mean recognition time, and recognition time consistency, as well as composites combining these measures. The vocabulary knowledge score (*VKsize*) is the proportion of hits ('yes' responses to words) minus the proportion of false alarms ('yes'; responses to pseudowords). The 'hit minus false alarm' score is an estimate of an individual's vocabulary size, or breadth. Two individuals with the same hit rate (an index of size) but different false-alarm rates will therefore also have different VKsize scores. The notation VKsize is used to denote the fact that the measure is not a direct estimate of the individual's vocabulary size, but rather a measure of vocabulary knowledge based on a frequency-indexed size measure that also takes guessing into account.

The second response measure is mean recognition speed (*mnRT*), based on the average of all the individual recognition times for the correct hits. The third measure is the coefficient of variation (*CV*), which is an index of the consistency of speed with which a set of words is recognized. It is a single value that reflects the relationship between the variability of the response times in a given set, as reflected in the standard deviation (SD), and the mean response time itself. The CV is not collected directly by the test but is derived from the mnRT and SD responses. It is the ratio of the SD of the mnRT to the mnRT itself (SD_{mnRT}/mnRT).

Hits	'Yes' responses to words
VKsize	Accuracy score providing an indirect measure of vocabulary size: proportion of hits minus false alarms ('yes' responses to pseudowords)
mnRT	Mean response time of correct hits
CV	Coefficient of variation: ratio of the standard deviation of the mean RT to the mean RT (SD_{mnRT}/mean RT)

Composite Measures

Two composite measures are also examined. The *VKsize_mnRT* measure combines the VKsize scores and mnRTs as a composite of size and speed. *The VKsize_mnRT_CV* measure combines all three measures in a single value. The third possibility *VKsize_CV*, which combines the VKsize scores and the CV, is not examined, as the CV is only interpretable in combination with the mnRT. It is possible for a test-taker to be very slow and very consistent.

The composite measures are calculated by first switching the sign on the raw mnRT and CV measures so that higher values reflect better performance. This makes the scores consistent with the VKsize values and permits standardized (z) scores for the three measures to be added together and averaged. Because the use of standardized scores usually results in some negative scores, a value of 5 is added to each score to make all the results positive. The composite scores are compared with the individual scores for how reliably they discriminate among levels and groups, and how large an effect they have on the criterion differences. Of particular interest is whether the combination of VKsize and mnRT/CV is more sensitive to criterion differences than the VKsize measure alone.

Establishing Response Reliability

In all the studies, the responses are initially examined for factors that may affect the outcomes, independent of the research variables of interest. These potentially compromising factors are both general to quantitative measurement research and specific to the use of the Timed Yes/No Test format. The raw findings are examined for instrument reliability, excessive false-alarm rates, the occurrence of outliers, and a potential speed–accuracy trade-off in responses.

Instrument Reliability Instrument reliability reflects the degree to which a test consistently measures what it is intended to measure; that is, it gives the same results every time it is used (in a hypothetically similar situation). The reliability of the Timed Yes/No Test is assessed by Cronbach's

alpha coefficient, a widely used measure of the consistency of responses across the items. Test instruments are assumed to be minimally reliable if the coefficient exceeds .7 out of a total 1. However, values above .8 are preferred for cognitive tests (Field 2009). In the studies reported later, the values range from mid-.8 to mid-.9.

Excessive False-Alarm Rates A potentially compromising factor unique to the Yes/No Test format is the presence of high false-alarm rates. An individual's false-alarm rate is subtracted from the hit rate (i.e., the number of actual words they correctly recognize) to yield the VKsize score. Higher false-alarm rates result in lower VKsize scores, and excessively high false-alarm rates call into question the viability of the score as a valid measure of underlying vocabulary knowledge. High individual false-alarm rates may result when the test-taker genuinely confuses pseudowords with known words, or does not pay close attention during testing, or fails to understand the task demands. Extremely high or low false-alarm and hit rates together are indicative of a pronounced tendency to respond to all items with either a 'yes' (high rates) or a 'no' (low rates). These cases need to be removed from the analysis. It is difficult to specify what constitutes a reasonable maximum false-alarm rate. In the studies reported in Part 2, the mean group false-alarm rates range from around 5% to 20%, the range reflecting decreases in group proficiency. Within the group, means are individual false-alarm rates that can be very high, with cases removed from the study only when the rate exceeded 45%. This is a very high level of guessing, but in most of the studies, a lower threshold would result in a substantial loss of data. If a number of test-takers must regularly be discarded because of excessive false-alarm rates, the instrument is of limited application in either research or assessment. Still, the high rate of guessing in some studies and groups within studies does raise questions about the validity of the measure. To address this concern, in several studies, a follow-up statistical analysis is performed in which the false-alarm rate threshold is lowered, scores exceeding the threshold are removed, and the data analysis is run again.

Outliers A common problem encountered in response time data analysis is the occurrence of outlier values. These are item response times that are

either too fast or too slow to reflect the cognitive process of interest. Random finger presses, lapses of attention, or external distractions can all contribute to responses that do not reflect the word recognition process. Outliers are identified here using an absolute value approach in which response times faster than 300 milliseconds and slower than 5000 milliseconds are the low and high cut-off values (Jiang 2013). The high cut-off value is the item time-out value for the test program and any response beyond this time is automatically discarded. The time-out value is set at 5000 milliseconds to accommodate the lower-proficiency test-takers in several of the studies. The data points that fell below the low cut-off of 300 milliseconds are simply removed. These involved only a handful of data points in any given study, well below 1% of the data.

Speed–Accuracy Trade-Off A potential confounding factor in the Timed Yes/No Test format is the possibility of a trade-off between speed and accuracy in individual test performance. Test-takers are instructed to respond as quickly *and* accurately as possible on every trial. Given the dual dimension of the task, it is possible that individuals might differ in how much emphasis they give to the respective dimensions. Some may work very quickly at the expense of accuracy, while others very slowly but very carefully. In the studies reported here, the VKsize scores are correlated with the inverse of the mnRT and CV scores to avoid the presence of minus signs in the reporting and discussion of results. If greater size (VKsize) and higher speed (mnRT or CV) are both elements of greater lexical facility, a positive correlation should exist overall between the two. Evidence of a systematic speed–accuracy trade-off is a significant negative correlation.

Evaluating the Sensitivity of the Lexical Facility Measures

The focus of the empirical research program is establishing the sensitivity of the lexical facility measures to the various performance criteria. Sensitivity is reflected in the degree to which the measures discriminate between levels in a given criterion and the effect size of these differences.

In each study, the descriptive results are first presented, followed by the inferential statistics used to test the sensitivity of the measures.

Descriptive Statistics The means, SDs, and confidence intervals (CIs) for the lexical facility measures are reported in all studies. The value of the CI as a statistical measure for both descriptive and inferential statistics is being increasingly recognized in L2 research (Larson-Hall and Herrington 2010; Larson-Hall and Plonsky 2015). The CI is a range of values that is likely to include the observed mean value for the sample. A bootstrapped (see below) 95% CI value is reported in all the studies, meaning that there is a 95% chance that the observed mean is contained in the interval between the lower- and upper-bound values. A lack of overlap in the CIs of two mean values indicates a statistically significant difference between them.

Discriminating Between Performance Levels Statistical tests are used to establish the sensitivity of the lexical facility measures to criterion performance differences, both individually and in combination. The sensitivity of each measure is based on the statistical significance of the test and the accompanying effect size. The alpha value for all the studies is .05, and specific adjustments are made where appropriate for multiple comparisons.

Group mean differences are tested using *t*-tests for comparisons involving two groups and an analysis of variance (ANOVA) for comparing more than two groups when the relevant assumptions are met. The *t*-tests and ANOVA assume that data which are normally distributed exhibit homogeneity of variance; that is, the SDs of the samples are approximately equal. The data were tested for the key assumptions of normality and equality of variance, which were generally, but not always, met. Where variance assumptions are not met for the standard ANOVA, Welch's ANOVA is used for the omnibus test and the Games–Howell test for any follow-up pairwise comparisons (Tabachnick and Fidell 2013). The studies here use bootstrapping for calculating mean CIs. Bootstrapping provides a more robust way to deal with non-normally distributed data than the use of nonparametric tests, particularly for smaller sample sizes (Larson-Hall and Herrington 2010).

The strength of association between the predictor lexical facility measures and criterion performance variables is measured the using Pearson's product moment correlation for bivariate correlations. In instances where the potential statistical significance of the difference between two bivariate correlations is of interest, the Fisher *r*-to-*z* transformation developed by Richard Lowry and available at http://vassarstats.net/tabs_rz.html is used. The data were analyzed using SPSS: statistical package for the social sciences. The program reports significance levels up to $p = .000$. In instances where these values are obtained, they are reported as $p < .001$.

Bootstrapped CIs for the group means are reported throughout. This permits all the results to be presented in a uniform manner, regardless of the size of the sample or whether a particular data set met all the assumptions for the use of the parametric measures.

Interpreting the Effect Size The other element of sensitivity is the effect size, which is the strength of the measure as a discriminator of criterion-level differences. In correlation and regression analyses, the effect size is calculated directly in the *r*-value. This value squared, R^2 (also called the coefficient of determination), represents the amount of variance in differences in the criterion variable, for example, the proficiency levels, attributable to differences in the predictor variable.

For the tests of group mean differences, standardized effect sizes are calculated separately. The *t*-test uses Cohen's *d*, and ANOVA the η^2 (eta-squared) test (Fritz et al. 2011). The relative importance of the observed effect sizes is interpreted using a recently introduced scale for interpreting the *r*- and *d*-values in L2 research (Plonsky and Oswald 2014, p. 899). The scale revises upward the widely used benchmarks suggested in Cohen (1988). Benchmark values for the interpretation of *d* are small ($d = .40$), medium ($d = .70$), and large ($d = 1.00$). Plonsky and Oswald (2014) note that these values pertain to between-group contrasts, with pre-post and within-group contrasts requiring larger effect sizes. The benchmarks for these contrasts are small ($d = .60$), medium ($d = 1.00$), and large ($d = 1.40$). Between-group contrasts involving proficiency levels are of primary interest in the studies presented in the following chapters, but within-group contrasts will also be relevant when comparing performance

over item frequency bands. The relative impact of the *r* effects are small (*r* = .25), medium (*r* = .40), and large (*r* = .60).

The following presents the empirical evidence for the lexical facility proposal. Chapters 6, 7, 8, 9, 10, and 11 report on a series of empirical studies, and the final chapter, Chap. 12, discusses the implications and way forward for the lexical facility account.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2011). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. doi:10.1037/a0024338.

Jiang, N. (2013). *Conducting reaction time research in second language studies*. London/New York: Routledge.

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368–390.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(S1), 127–159.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912. doi:10.1111/lang. 12079.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.