Applied Linguistics 24/4: 492–518

Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing

LOURDES ORTEGA

Northern Arizona University

In this study I evaluate the cumulative evidence on the use of syntactic complexity measures as indices of college-level L2 writers' overall proficiency in the target language. Based on a synthesis of twenty-five studies, I arrive at several substantive findings. First, I conclude that the relationship between L2 proficiency and L2 writing syntactic complexity varied systematically across studies depending on whether a second or a foreign language learning context was investigated and whether proficiency was defined by programme level or by holistic rating. Second, aggregating available cross-sectional findings, I propose critical magnitudes for between-proficiency differences in syntactic complexity for four measures. Finally, I interpret the limited longitudinal evidence to suggest that an observation period of roughly a year of college-level instruction is probably needed for substantial changes in the syntactic complexity of L2 writing to be observed. I conclude the paper by discussing implications of these findings for future primary research.

Syntactic complexity (also called syntactic maturity or linguistic complexity) refers to the range of forms that surface in language production and the degree of sophistication of such forms. This construct is important in second language research because of the assumption that language development entails, among other processes, the growth of an L2 learner's syntactic repertoire and her or his ability to use that repertoire appropriately in a variety of situations. Length of production unit, amount of embedding, range of structural types, and sophistication of the particular structures deployed in production have all been the target of quantifications when characterizing syntactic complexity, resulting in a variety of global measures.

Measures of syntactic complexity are important research tools not only in the field of second language acquisition but in a variety of language-related disciplines, and although concrete definitions and applications of the measures vary across fields, the main purposes for use are very similar. In L2 writing research, specifically, syntactic complexity measures have been used to evaluate the effects of a pedagogical intervention on the development of grammar, writing ability, or both; to investigate task-related variation in L2 writing; and to assess differences in L2 texts written by learners across

proficiency levels and over time (see Polio 2001). This last use of syntactic complexity measures is my focus in the present research synthesis. Specifically, I examine the cumulative evidence yielded by primary studies that investigated the extent to which syntactic complexity measures derived from L2 writing samples can be useful indices of the writers' overall proficiency in the target language.

THE QUEST FOR GLOBAL INDICES OF L2 PROFICIENCY

Within SLA research, and inspired by the successful use of metrics of complexity in the study of child language acquisition, Larsen-Freeman (1978) argued for the need to identify or devise a measure that served as a yardstick by which to assign L2 learners to differing language development bands. Motivating these early research efforts was the question of whether global measures of L2 syntactic complexity could be used as anchors for characterizations of language development and language proficiency. Two decades after Larsen-Freeman's call, and motivated by the same research goal, Wolfe-Quintero *et al*. (1998) looked cumulatively at the strength of the relation between a number of these metrics and proficiency levels across thirty-nine L2 writing studies.[1] The researchers concluded that mean length of T-unit, mean length of clause, clauses per T-unit, and dependent clauses per clause were the most satisfactory measures, because they were associated linearly and consistently with programme, school, and holistic rating levels across the thirty-nine primary study reports. On the other hand, they cautioned that these very same metrics nevertheless discriminated poorly between adjacent levels of proficiency and that statistically significant relationships were only inconsistently found in studies where L2 proficiency level had been operationalized by means of holistic ratings. In their book-length treatment, Wolfe-Quintero *et al*. (1998) highlighted several problems in the cumulative state of knowledge offered by studies in this research domain, particularly the lack of information about the reliability of interlanguage codings and the inconsistent choice and definition of measures across studies. In spite of these weaknesses, they also concluded that '[c]learly, the potential uses for developmental measures in pedagogical, testing, and acquisition research make continued research on these measures worthwhile' (p. 127).

ON THE INTERPRETATION OF SYNTACTIC COMPLEXITY MEASURES

As demonstrated by their extended use across language-related fields, global measures of syntactic complexity are useful for a variety of research purposes, most of them related to their modest but practical value as tools for normative comparison across samples, populations, and contexts. Given their heuristic value, it is important that researchers accord interpretations that are

commensurate to the purposes for using these research tools in L2 writing in the first place.

Clearly, syntactic complexity metrics would be misapplied if they were to be used as absolute developmental indices or as direct indices of language ability. First, 'more complex' may mean 'more developed' in many different ways, and the nature of L2 development cannot be sufficiently investigated by means of these global measures alone. They can only provide a logistically convenient start for the analyst to search for evidence of steady developmental change that is comparable across individuals and groups. Second, 'more complex' does not necessarily mean 'better'. Progress in a learner's language ability for use may include syntactic complexification, but it also entails the development of discourse and sociolinguistic repertoires that the language user can adapt appropriately to particular communication demands.

In addition, it would be misguided to equate more linguistically complex writing with 'good' or 'expert' writing. Strong empirical evidence for this was found in a well-known meta-analysis of L1 writing by Hillock (1986), who concluded there was no consistent relationship between the syntactic complexity of written products and holistic ratings of 'good' writing. In L2 writing, the development of composing expertise is a complex phenomenon that goes well beyond establishing the linguistic abilities of L2 writers, as Shaw and Liu (1998) and Valdés *et al.* (1992), among many others, have discussed. Undoubtedly, however, any theory of L2 writing must deal with the question of how second language ability may affect the development of L2 writing expertise. In doing so, for instance, L2 writing theorists have identified L2 proficiency as the site of a fundamental difference between L1 and L2 writing (Silva 1993), and as a prerequisite both for the transfer of writing skills from the L1 to the L2 (Cumming 1989) and for the sustainment of productive writing behaviours (Roca de Larios *et al.* 2001). Thus, an improved understanding of the workings of syntactic complexity in L2 writing advances research programmes that investigate the role of L2 proficiency in the development of L2 writing expertise.

Accordingly, one goal in the present research synthesis is to help establish useful landmarks for the interpretation of syntactic complexity measures in future L2 writing research. Another goal is to improve current understandings of the relationship between syntactic complexity and L2 proficiency by paying closer attention to questions of magnitude and rate of grammatical development in L2 writing. I chose to summarize and interpret the accumulated empirical evidence in this research domain through the methodology of research synthesis (Cooper and Hedges 1994; Light and Pillemer 1984).

WHY RESEARCH SYNTHESIS?

Wolfe-Quintero *et al.* (1998: 117–18) identified three sources of information that reviewers can use to evaluate the construct validity of particular

measures as indices of L2 proficiency: predictive validity, concurrent validity, and repeated sampling reliability. Predictive validity would be found for a measure if associated known developmental stages for particular grammatical subsystems tended to characterize the production of L2 learners at similar mean values for that measure. This was, in a nutshell, Brown's (1973) initial motivation for advocating the use of mean length of utterance in child language development studies. Concurrent validity, on the other hand, would accrue for a measure that yields statistically significant between-proficiency differences in many studies, that correlates significantly and highly with proficiency levels in many studies, or both. Finally, repeated sampling reliability would be found when a measure exhibits a linear progression along independently determined proficiency levels, and this pattern is replicated consistently across studies. Noting in their review that the predictive validity of interlanguage measures has yet to be addressed by primary L2 researchers, Wolfe-Quintero *et al*. concentrated on assessing the concurrent validity and the repeated sampling reliability of global measures employed in the reviewed studies.

The examination of repeated sampling reliability is a welcome strategy that allows the reviewer to evaluate primary findings without only relying on outcomes of statistical significance testing. To put it simply, the accumulation of consistent values across enough studies amounts to evidence for significant differences (cf. Light and Pillemer 1984: 77). Underlying Wolfe-Quintero *et al*.'s use of concurrent validity, however, is the practice of taking a vote-count of statistically significant results across studies. The vote-count procedure is of limited usefulness and can result in error when aggregating information across studies because it 'ignores sample size, effect size, and research design' (Light and Pillemer 1984: 4). In brief, the problem is one of overreliance on statistical significance test results. As Norris and Ortega (2000: 493–5) have pointed out, this problem leads to two misguided practices in applied linguistics quantitative research: (a) employing statistical significance as the sole source of evidence for the presence or absence of a relationship, and (b) mistakenly using statistical significance to make interpretations about the magnitude or importance of a relationship.

In order to overcome these problems, an improved approach to evaluating cumulative evidence is to focus on magnitude of observed effects, employing systematic procedures to translate these effects into some equivalent way across studies so as to make their comparison and aggregation possible. Research synthesis offers just this improved focus on magnitude of effects and the systematic procedures needed to implement it.

METHOD

Studies included in the synthesis

The present research synthesis focused on studies that investigated the relationship between quantitative differences in the syntactic complexity of L2 written texts and proficiency differences among L2 learner groups in college-level second or foreign language instructional contexts. Initially, a potential body of relevant literature was retrieved from Wolfe-Quintero *et al*. (1998), through a computer search of the databases of ERIC, and by cross-checking reference sections from each retrieved study report. Inclusion in the synthesis was limited to studies that reported on samples of L2 learners enrolled in college-level second or foreign language classes. That is, the following were excluded: studies of high school L2 writers and studies whose participants were recruited from freshman composition courses or regular university courses.

Two relatively homogeneous bodies of relevant research were identified, as shown in Table 1. First, twenty-one cross-sectional studies examined measures of syntactic complexity across two or more L2 proficiency groups. Sixteen of these studies were reviewed by Wolfe-Quintero *et al*. (1998), and five additional studies were included in the present synthesis. The analyses focus on the six most frequently used syntactic complexity measures across these twenty-one studies. Three of the measures tap length of production at either the clausal or phrasal level (mean length of sentence [MLS], mean length of T-unit [MLTU], and mean length of clause [MLC]), one measure reflects amount of coordination (mean number of T-units per sentence [TU/S]), and two measures gauge the amount of subordination (mean number of clauses per T-unit [C/TU], and mean number of dependent clauses per clause [DC/C]).

In addition, I was able to identify a handful of longitudinal studies that investigated learners' linguistic development as indexed by changes in the syntactic complexity of L2 writing over time (see Table 1). Only results for MLTU were investigated in the longitudinal subsample because this was the only measure found to be employed by all six studies.

Research questions

The two sets of L2 writing studies were closely inspected in the search for answers to the following questions:

> *Research Question 1*: What (if any) is the impact of instructional setting and proficiency sampling criterion on the mean values and ranges observed for any given syntactic complexity measure across the twenty-one cross-sectional studies?

I chose to treat the instructional setting, or whether a study was conducted in a foreign or a second language context, as a variable in the synthesis because

Table 1: College-level L2 writing studies included in the research synthesis

	Study	N	Measures investigated
Cross-sectional			
ESL (k = 13)	Bardovi-Harlig and Bofman (1989)	30	C/TU
	Flahive and Snow (1980)	300	MLTU, C/TU
	Gaies (1976)	25	MLTU, MLC, C/TU
	Homburg (1984)	30	MLS, MLTU, TU/S, C/TU
	Ho-Peng (1983)	60	MLTU
	Kameen (1979)	50	MLS, MLTU, MLC, C/TU, DC/C
	Larsen-Freeman (1978)	212	MLTU
	Larsen-Freeman (1983-Study 2)	102	MLTU
	Larsen-Freeman and Strom (1977)	48	MLTU
	Perkins (1980)	29	MLTU, C/TU
	Perkins and Homburg (1980)	23	MLTU
	Sharma (1980)	60	MLTU, MLC, C/TU
	Tedick (1990)	105	MLTU
EFL	Hirano (1991)	158	C/TU, DC/TU
(k = 3)	Neff *et al.* (1998)	180	MLTU, MLC, C/TU
	Nihalani (1981)	29	MLTU
FL in the USA	Cooper (1976)	50	MLS, MLTU, MLC, TU/S, C/TU
(k = 5)	Dvorak (1987)	12	MLTU
	Henry (1996)	67	MLTU
	Kern and Schultz (1992)	73	MLTU
	Monroe (1975)	110	MLS, MLTU, MLC, TU/S, C/TU
Longitudinal			
ESL	Arthur (1992)	14	MLTU
(k = 2)	Larsen-Freeman (1983-Study 3)	23	MLTU
EFL	Arnaud (1992)	50	MLTU
(k = 3)	Casanave (1994)	28	MLTU
	Ishikawa (1995)	4	MLTU
FL in the USA (k = 1)	Kern and Schulz (1992)	73	MLTU

Note. All studies are identified in the reference list with a preceding asterisk.

of repeated observations by some L2 writing researchers that the FL learning context influences the nature of L2 writing development in special ways (e.g. Hirose and Sasaki 1994; Ishikawa 1995; Kubota 1998). A related argument is that L2 competence may proceed more slowly and might develop less fully in foreign language than in second language instructional settings.[2] I also treated the proficiency sampling criterion as a study variable of interest in the synthesis because Wolfe-Quintero *et al*. (1998: 98), on the basis of statistical significance results, concluded that proficiency is probably related to increases in syntactic complexity in L2 writing only when the former is defined by programme level, but not by holistic ratings. I averaged observed means across studies and inspected summary descriptive statistics of accumulated findings to address this research question.

> *Research Question 2*: What are the observed typical average differences between any two L2 proficiency levels for a given measure across studies, and at what magnitude are such between-proficiency differences found to be statistically significant?

This question pertained to the concurrent validity of the measures. However, I was careful to avoid the problem of overreliance on statistical significance for the interpretation of aggregated study outcomes. The technique of stem-and-leaf plotting (Greenhouse and Iyengar 1994) allowed me to interpret statistical significance results together with sample size across studies and to address the fundamental issue of 'how different is different enough' in terms of magnitudes expressed in readily interpretable units (see explanation of this use of stem-and-leaf plots in the Results section).

> *Research Question 3*: What is the observed amount of change when gains in syntactic complexity ratios relative to length of observation period are compared across longitudinal studies?

This question was ultimately related to the underexplored issue of rate in L2 development, including 'how long is long enough' to observe substantial L2 syntactic development, and how much change in syntactic complexity can be expected at different junctures over a given curricular length of college-level L2 instruction. I calculated pre-to-post effect sizes (Light and Pillemer 1984) in order to address this research question.

RESULTS

General study characteristics

Table 2 shows the cross-sectional set of twenty-one studies by instructional setting and proficiency sampling criterion. Of the twenty-one L2 writing studies, 62 per cent analysed L2 writing produced by ESL university students, whereas only 38 per cent were concerned with L2 writing produced by foreign language university students. Of those, three were EFL studies and the other five investigated various foreign languages in the USA. (Since there

were so few in each category, the two groups were collapsed as 'FL' in the analyses.) Programme level was used to define proficiency group differences in 52 per cent of the studies, whereas holistic ratings were used for the same purpose in 38 per cent of the studies. Only two studies used standardized test score bands as a basis for establishing proficiency level groupings.[3]

Writing tasks varied considerably across studies. A little less than one-third of the studies, all of them drawing on ESL college populations, analysed compositions written under examination conditions. An almost equally frequent prompt type consisted of argumentative or persuasive essays collected through in-class writing assignments under a time limit. Less frequent (only three studies) was the elicitation of controlled writing by asking students to rewrite a passage so as to combine sentences using coordination and subordination. Finally, Cooper (1976) collected miscellaneous writing from a range of classroom assignments; Dvorak (1987) elicited a picture narrative and an argumentative essay; Nihalani (1981) collected home compositions of unknown topic; and Henry's (1996) students wrote on the topic entitled 'Me' for 10 minutes.

Table 3 summarizes selected study characteristics for the cross-sectional set. In addition, to the noted differences in the writing tasks, two factors shown in Table 3 largely differed across studies: sample size and elicitation method

Table 2: Cross-sectional college L2 writing studies by instructional setting and proficiency sampling criterion

Proficiency criterion	ESL (k = 13)	FL (k = 8)
Program level (k = 11)	Flahive and Snow (1980), Gaies (1976), Ho-Peng (1983), Larsen-Freeman (1978), Larsen-Freeman (1983-Study 2), Tedick (1990)	Cooper (1976), Dvorak (1987), Henry (1996), Monroe (1975), Neff *et al.* (1998)
Holistic ratings (k = 8)	Bardovi-Harlig and Bofman (1989), Homburg (1984), Kameen (1979), Larsen-Freeman and Strom (1977), Perkins (1980), Perkins and Homburg (1980)	Kern and Schultz (1992), Nihalani (1981)
Score ranges (k = 2)	Sharma (1980)	Hirano (1991)

Note. k = number of studies.

Table 3: Study characteristics of cross-sectional college L2 writing studies

	k	mean	*SD*	min.	max.
N-size per study	21	84	74	16	300
N-size per cell	73	23	16	3	80
Time limit (minutes)	11[a]	38	13	10	50
Corpus length (words/writer)[b]	13[c]	234	110	70	500

Note. *k* = number of studies (and number of learner groups in second row of table).
[a] Writing time conditions were not reported in seven studies, and were unavailable for one study (Gaies, 1976, reported in Sharma, 1980); two studies elicited writing without time limit.
[b] Mean number of words per writer for each study was calculated on the basis of reported averages for various proficiency groups and should therefore be taken as a general indication of relative length rather than as an exact index.
[c] Corpus length was not reported in seven studies and was unavailable for one study (Gaies, 1976, reported in Sharma, 1980).

(including timing of the writing and resulting corpus length, which ranged across studies from 70 to 500 words per writer).

Given the small size of the longitudinal set, a detailed synthesis of study features was not pursued. Nevertheless, an inspection of the longitudinal studies suggests very similar study characteristics to those synthesized for the cross-sectional set.

Research Question 1: Instructional setting and proficiency sampling criterion

Instructional setting effects

The inspection of the mean values across L2 groups aggregated from all twenty-one cross-sectional studies (shown in Table 4) reveals that, for each of the six complexity measures, the mean of ESL groups combined was always higher than the mean of FL groups combined. This difference was only trustworthy for MLS (the mean of ESL groups was 7.83 words higher than that of the FL groups) and MLTU (the mean of ESL groups was 3.13 words higher than that of the FL groups), both comparisons yielding non-overlapping confidence intervals, which amounts to a statistically significant difference. For the rest of the measures, the observed differences across instructional settings were not trustworthy, as shown in largely overlapping confidence intervals (see Table 4). However, the confidence intervals for TU/S (with a mean difference of 0.20 T-units in favour of the ESL groups) and DC/C (with a mean difference of 0.14 dependent clauses in favour of the ESL groups) would shrink and become non-overlapping if another two groups

Table 4. Means for six syntactic complexity measures across twenty-one college L2 writing studies[a]

	k	Mean	*SD*	Min.	Max.	Range	95% confidence interval	
							Upper	Lower
MLS								
All groups	13	15.53	4.85	9.17	23.59	15.42	18.46	12.60
ESL groups	5	20.35	3.39	15.52	23.59	9.07	24.56	16.14
FL groups	8	12.52	2.62	9.17	16.90	8.73	14.71	10.33
MLTU								
All groups	65	12.18	3.03	5.10	18.40	14.30	12.92	11.44
ESL groups	41	13.34	2.51	7.00	18.40	12.40	14.13	12.55
FL groups	24	10.21	2.87	5.10	15.30	11.20	11.42	9.00
MLC								
All groups	17	7.64	1.43	5.64	10.83	6.19	8.38	6.90
ESL groups	7	7.86	1.59	6.25	10.83	5.58	9.33	6.39
FL groups	10	7.49	1.37	5.64	9.90	5.26	8.47	6.51
TU/S								
All groups	11	1.29	0.12	1.20	1.59	1.39	1.37	1.21
ESL groups	3	1.44	—	1.35	1.59	1.24	1.76	1.12
FL groups	8	1.24	0.06	1.20	1.37	1.17	1.29	1.19
C/TU[b]								
All groups	32	1.49	0.19	1.07	1.92	1.85	1.56	1.42
ESL groups	19	1.53	0.20	1.07	1.92	1.85	1.63	1.43
FL groups	13	1.44	0.17	1.20	1.78	1.58	1.54	1.34
DC/C								
All groups	5	0.31	0.09	0.18	0.40	1.22	0.42	0.20
ESL groups	2	0.39	—	0.37	0.40	1.03	0.57	0.21
FL groups	3	0.25	—	0.18	0.33	1.15	0.45	0.05

Note. *k* = number of learner groups.
[a] Native speaker groups were not included.
[b] Bardovi-Harlig and Bofman's (1989) groups were excluded here because of lack of comparability in clause definition.

were added to the ESL and FL categories for each measure. In other words, a small increase in the size of the sample available for TU/S and DC/C comparisons would lead us to reject the null hypothesis of no effect for instructional setting, in the hypothetical case of future studies contributing a similar distribution of group means.

A reasonable explanation for this finding is that FL studies simply tended to employ samples at lower proficiency levels than did ESL studies. One can speculate, further, that this sampling bias may reflect a characteristic of the larger populations, assuming that college-level ESL instruction by definition generally involves learners at higher levels of L2 proficiency than what is typically the case for FL instruction (e.g. Hirose and Sasaki 1994; Kasper 1997).

Effects of proficiency sampling criterion

The relationship of programme level and holistic rating with mean syntactic complexity values was examined by inspecting results for two measures which were sufficiently investigated across studies: MLTU (in eighteen studies) and C/TU (in eight of the same studies). Since in the total cross-sectional corpus the number of studies using a holistic rating criterion was skewed in favour of ESL studies (cf. Table 2), an interaction between instructional setting and proficiency sampling criterion might be expected. Consequently, accumulated means and ranges were also inspected for programme level versus holistic rating studies of ESL and FL samples separately.

Descriptive statistics for the data are shown in Table 5. It can be seen that studies that operationalized proficiency in terms of holistic ratings consistently yielded more homogeneous observations for MLTU and C/TU, as reflected in smaller standard deviations and narrower ranges. Further, no interaction was found between proficiency sampling criterion and instructional setting, the holistic rating studies consistently yielding more homogeneous observations than the programme level studies with both ESL and FL learner samples.

Apparently, researchers who established L2 proficiency differences among their participants by means of holistic ratings ended up with samples that represented a narrower range of proficiencies than did researchers who sampled participants at varying L2 proficiency bands based on their institutional membership into various programme levels. This may be because holistic rating studies typically targeted small proficiency differences within a single programme level (e.g. Kern and Schultz 1992) or within very narrow bands of test scores obtained from exams at key institutional junctures, such as placement or entrance exams (e.g. Bardovi-Harlig and Bofman 1989; Homburg 1984). In addition, it may be more difficult for raters to make fine-grained judgements that sufficiently discriminate individual performances within such truncated samples comprising fairly small proficiency differences.

Table 5: MLTU and C/TU results for subsample of eighteen studies by instructional setting and proficiency sampling criterion

	k	Mean	St. Error	*SD*	Min.	Max.	Range
MLTU							
Program levels	40	11.88	0.55	3.42	5.10	17.21	13.11
Holistic ratings	22	13.06	0.46	2.13	9.17	18.40	10.23
FL program levels	18	10.01	0.78	3.21	5.10	15.30	11.20
ESL program levels	22	13.41	0.62	2.82	7.00	17.21	11.21
FL holistic ratings	6	10.78	0.66	1.48	9.17	12.37	4.20
ESL holistic ratings	16	13.91	0.43	1.67	10.99	18.40	8.41
C/TU							
Program levels	18	1.48	0.05	0.22	1.07	1.92	1.85
Holistic ratings (only ESL)[a]	8	1.59	0.04	0.11	1.38	1.74	1.36
FL program levels	10	1.46	0.06	0.17	1.20	1.78	1.58
ESL program levels	8	1.51	0.11	0.28	1.07	1.92	1.85
All ESL groups	16	1.55	0.05	0.21	1.07	1.92	1.85

Note. *k* = number of learner groups.[a] All holistic rating groups were ESL groups, as no FL study employed C/TU with holistic rating levels.

Research Question 2: Between-proficiency level differences in syntactic complexity

Research Question 2 reconceptualized concurrent validity in terms of magnitude rather than statistical significance alone: when inspecting the evidence for concurrent validity across the twenty-one cross-sectional L2 writing studies, (a) how large a difference in syntactic complexity between groups is large enough to be statistically significant, and (b) how large a difference can be considered typical of what can be observed between any two given proficiency levels in college L2 writing? I was guided by the caution that sample size may have critically affected some of the significance test outcomes observed by individual researchers, because cell and study *N*-sizes within the cross-sectional sample varied widely (cf. Table 3). This analytical approach was enabled by the use of stem-and-leaf plots.

On stem-and-leaf plots

A detailed explanation of how to read a stem-and-leaf plot follows, using Figure 1 as the case in point (see also Greenhouse and Iyengar 1994: 386–7).

A stem-and-leaf plot has a vertical line and multiple horizontal lines. The numbers aligned vertically parallel to the left of the line (the stem) represent the first digit of an observed mean difference. For instance, it can be seen in Figure 1 that the largest difference in MLS observed between two groups in a given study reached eight words and, in the opposite direction, the largest observed negative difference reached two words. The numbers to the right of the line (the leaves) represent study comparisons; each number reflects the rounded one-decimal value to be added to the left-hand numbers horizontally. Thus, in Figure 1, there are 14 numbers to the right of the vertical line, which correspond to 14 between-group comparisons, contributed in this particular plot by four studies. So, for instance, a between-proficiency MLS difference of one word was found in three group comparisons, and the specific one-word differences observed in the three comparisons were 1.0, 1.5, and 1.7, respectively. Gaps to the right-hand side of the stem-and-leaf plot indicate that no studies observed a difference of that size. Thus, for instance, Figure 1 shows that no study observed a difference of three or seven words in MLS between any two proficiency groups. An asterisk next to a number indicates that the observed between-groups difference was reported to be statistically significant in the original study. Numbers in parentheses (see, for example, Figure 2) indicate an observed mean difference that was not tested statistically in the original study. It should be noted that the mean differences represented in the stem-and-leaf figures include both adjacent proficiency differences and maximal proficiency differences (i.e. difference between the lowest and highest proficiency level groups in any given study contributing multiple group comparisons).

Magnitude of between-proficiency differences in MLS

Figure 1 presents a stem-and-leaf summary of the observed mean differences in MLS contributed by fourteen groups sampled in four studies. As can be seen, the observations varied widely. However, it can also be seen that differences of four and a half words per sentence or above were always statistically significant, whereas differences of up to roughly two and a half words per sentence were not. All comparisons involved similar sample sizes (mean cell *N*-size was 16, cell *N*-size range = 10–25). From inspection of Figure 1, it can be tentatively concluded that a difference in MLS means of 4.5 or higher is likely to be statistically significant for cell samples of ten or more participants. This prediction is only tentative and may change if data from new studies employing MLS become available (notice that there are many gaps in the stem-and-leaf plot, possibly because the number of studies contributing data is so low).

–2	3
–1	
–0	
0	8
1	0 5 7
2	0 3 6
3	
4	5* 7*
5	8*
6	1* 6*
7	
8	1*

Figure 1: Stem-and-leaf plot for MLS between-proficiency level differences across four studies

Magnitude of between-proficiency differences in MLTU

Figure 2 shows the mean differences obtained across nineteen studies that employed MLTU. Of the 68 comparisons, 49 involved adjacent levels of proficiency, 16 were comparisons of maximal proficiency differences within studies, and three comparisons contrasted an advanced level and a native-speaking group. The stem-and-leaf figure clearly shows that many comparisons clustered around small differences of about half a word to slightly over two words per T-unit (overall, the mean difference was 1.4 words for adjacent-level comparisons and three words for maximal difference comparisons). It can also be seen that differences of two words per T-unit or above tended to be statistically significant in the original studies (no inferential tests were available for twelve comparisons, all from three studies).

–1	5 2
–0	6 4 3 2 1
0	0 0 1 1 (2) 2 2 2 3 3 (4) (4) 4 4 4 4 6 6 7 (9) 9* 9
1	0 0 1* 2 (4) 4 (5) 5 6 6 6 (7) (7) 8*
2	(2) 2* 2* 2* 2 3* 8* 9* 9
3	0* 1* 1* 3* 4* 5* (7)
4	0* 1* 2* 4*
5	3* 7*
6	(2) (6)
7	
8	9*

Figure 2: Stem-and-leaf plot for MLTU between-proficiency level differences across 19 studies

The pattern for statistical significance of MLTU comparisons at the critical level of roughly two words of difference seems to be disconfirmed by four cases (0.9*, 1.1*, 2.2, and 2.9), which will be discussed below. Before discussing disconfirmatory evidence, however, a brief consideration of the relationship between sample size, effect size, and statistical significance is warranted (Kramer and Rosenthal 1999). Further, this relationship ought to be examined taking instructional setting and proficiency sampling criterion into account, since both factors led to differences in study outcomes, as shown earlier. Table 6 summarizes all this information.

As shown in Table 6, the mean of observed differences that turned out to be statistically significant for the group of ESL studies using a programme level criterion, which had a substantially higher average sample size than the rest of the studies, was roughly three words. By comparison, for all other study groups, only somewhat larger differences were found to be statistically significant. This is only what would naturally be expected of the relationship between effect size and sample size; that is, the larger the sample size, the lower the critical value necessary for a statistically significant observation. More interesting, for the purpose of establishing critical magnitude values, is to look at the ranges of observed mean differences that were found to be statistically significant and those that were not (this is indicated in Table 6 under minimum and maximum values associated with each mean). This examination will also allow us to return to the issue of disconfirmatory evidence found in Figure 2 and interpret it in the light of the data presented in Table 6.

For the ESL studies using a programme level criterion (with a mean cell *N*-size of 37 participants), the minimum mean difference that was statistically significant was 1.8 words, without exceptions. Nevertheless, caution must be exerted in taking this critical value as definitive, since the total number of observations is limited to nine and several mean group comparisons below 1.8 were not tested inferentially in three studies (see values in parentheses in Figure 2).

For FL studies using a programme level criterion (with a mean cell *N*-size of 17 participants), the minimum mean difference in MLTU found to be statistically significant (0.9 words) overlaps with the maximum mean difference found to be statistically not significant (2.2 words). This overlap is visually located in the values 0.9*, 1.1*, and 2.2 in Figure 2. The difference of 2.2 words found by Cooper (1976) in a comparison between junior and senior German students' MLTU was statistically not significant but large enough to be so, judging from similar-size, statistically significant between-proficiency level differences contributed by three other studies (all shown as 2.2* in Figure 2). Conversely, Henry's (1996) comparisons between fourth- and sixth-semester Russian students (with a mean difference of 1.1*) and between second- and fourth-semester students (with a mean difference of 0.9*) were small enough not to have been statistically significant, judging from similar-size differences obtained in two other studies, shown in the

Table 6: The relationship between sample size, magnitude of observed difference, and statistical significance in MLTU mean comparisons across nineteen studies

	ESL program level groups (k = 9)	FL program level groups (k = 23)	ESL holistic rating groups (k = 15)	Other groups[a] (k = 9)
Mean cell *N*-size	36.5	16.92	11.6	18
Statistically significant differences				
k	6	13	3	0
mean	2.93	3.48	3.80	—
minimum/maximum	1.8/3.5	0.9/8.9	3.3/4.1	—
Statistically non-significant differences				
k	3	10	12	9
mean	0.49	1.03	0.28	0.26
minimum/maximum	0.1/1.2	–0.4/2.2	–1.5/2.9	–0.58/0.95

Note. k = number of group comparisons.
[a] Groups contributed by Sharma (1980) and Hirano (1991)

values without asterisks on the 0-word and 1-word difference leaves. A close examination of these individual studies revealed that sample size does not explain the results. Thus, the observed overlaps should be considered related to other (unknown) characteristics of the studies involved (for example, differences in types of significance tests employed or differences in variance in scores). Consequently, more data from future FL studies using a programme level criterion will be necessary to determine a critical effect size for significant between-proficiency level differences in MLTU in FL programme level studies. For now, a conservative provisional guess appears to be 2.2 words, even with relatively small cell *N*-sizes.

For ESL studies using a holistic rating criterion of proficiency (with a mean cell *N*-size of 12 participants), the minimum mean difference in MLTU found to be statistically significant was 3.3 words, contributed by a comparison in Homburg (1984) and the maximum mean difference found to be statistically not significant was 2.9 words, contributed by a comparison in Larsen-Freeman and Strom (1977). Although within the ESL holistic rating group of studies these results are non-overlapping, a difference of 2.9 was large enough that it should have been statistically significant against the backdrop of the findings gleaned from ESL and FL programme level groups (cf. Figure 2). Perhaps Larsen-Freeman and Strom's (1977) lack of statistically significant

results in their comparison between 'poor' and 'excellent' ESL writers can be explained as a consequence of a lack of power in the performed ANOVA analysis because of the small size of the sample and the unequal size of the cells, or because of depressed within- or between-group variances (a problem that appears to be frequent in holistic rating studies, as already shown). If this study is considered an outlier, the critical value for significant between-proficiency level differences in MLTU for ESL holistic rating studies would be established at 2.9 words or higher for typical studies of this sort.

Summarizing from the results presented in Figure 2 and in Table 6, a mean difference in MLTU between two cross-sectional samples of differing proficiency levels can be expected to be statistically significant if it is roughly two words or higher. Further, the size of the samples involved will moderate this estimate. For larger samples (of mostly 20 to 50 participants per cell, as in the ESL studies using a programme level criterion) 1.8 words of difference may be enough for statistically significant results; conversely, for smaller samples (of mostly 10 participants per cell, as in the FL studies using a programme level criterion and the ESL studies using a holistic rating criterion) a mean difference of over two words (2.2–2.9) may be required. Finally, mean MLTU between-proficiency level differences of one word or less are likely to be statistically not significant.

Magnitude of between-proficiency differences in MLC

Figure 3 presents a stem-and-leaf plot of the observed mean differences in MLC across six studies, including seventeen adjacent proficiency level differences and six maximal proficiency differences between the highest and lowest proficiency level group in each study. Most groups across comparisons comprised twenty participants or more (average sample size per cell was 20, $SD = 8$). It can be seen in Figure 3 that mean difference values of 1.1 words per clause or higher tended to be statistically significant, with the exception of two observed group differences: 1.1 (contributed by Cooper 1976) and 1.3 (contributed by Neff *et al.* 1998).

These two exceptions to the pattern seem to be attributable to sample size. Specifically, the observed difference of 1.1 was statistically significant in

–0								
0	1	2	4	5	5	5	5	8
1	0	0	1*	1*	1	3	4*	6*
2	(3)	3*	6*					
3								
4	6*							

Figure 3: Stem-and-leaf plot for MLC between-proficiency level differences across six studies

studies with a cell sample of 22 (Monroe 1975) and 30 (Neff *et al*. 1998), but not significant when cell sample size was only 10 (Cooper 1976). Similarly, a mean difference of 1.3 between unequal cell sizes (30 L1 English college writers and 15 L1 English professional writers) was not significant in Neff *et al*. (1998) but a smaller size difference of 1.1 obtained in the same study from a comparison involving equal cell sizes of 30 participants (30 fourth-year EFL college writers and 30 L1 English college writers) turned out to be statistically significant. It can, thus, be concluded tentatively, and until more cross-sectional L2 writing studies employ MLC, that between-proficiency level differences below one word per clause will tend to be statistically not significant, and that differences of slightly over a word (1.1–1.3) may be expected to be statistically significant, provided the samples involved are large enough (probably twenty participants or larger).

Magnitude of between-proficiency differences in C/TU

Due to an insufficient number of observations, critical values for TU/S and DC/C could not be inspected. Fortunately, the data available for C/TU, which taps amount of subordination, were sufficient to examine the question of magnitude for this measure. Figure 4 shows a stem-and-leaf summary of the group mean differences in C/TU yielded by thirty-six group comparisons across ten studies, including twenty-six adjacent level L2 group comparisons, seven comparisons involving maximal proficiency differences, and three comparisons between an advanced proficiency level group and a native speaker baseline group. As shown visually in Figure 4, all differences of .20 or greater, whether positive or negative, were statistically significant. However, many of the C/TU comparisons in the ten studies were not tested inferentially (those in parentheses) and, therefore, results could look different if more observation points were available.

–0.2	5*								
–0.1	(8)	5	4	0					
–0.0	2	1							
0.0	0	3	(4)	6	6	(6)	8	8	9
0.1	0	2	2	3	(6)	7	7		
0.2	0*	0*	(1)	(2)	5*	(8)	9*		
0.3	0*	1*	(2)	(6)					
0.4									
0.5									
0.6	7*								

Figure 4: Stem-and-leaf plot for C/TU between-proficiency level differences across ten studies

Summary of findings for Research Question 2

To summarize, the following magnitudes of observed between-proficiency level differences can be tentatively expected to be statistically significant in cross-sectional studies given samples of a medium size (of about ten participants or more per cell): 4.5 or more words per sentence (MLS), two or more words per T-unit (MLTU), slightly over one or more words per clause (MLC), and at least .20 positive or negative differences in number of clauses per T-unit (C/TU). These differences can be viewed as important because they are linked to real observed criterial differences in proficiency level for the particular measures involved, and across studies of similar characteristics (e.g. target populations, sample sizes, and designs). On the other hand, they also ought to be regarded as provisional until more research is available because of gaps in accumulated observations.

Research Question 3: Rate of longitudinal change

The final issue addressed in the present research synthesis was: is it possible to infer from existing longitudinal studies an estimate of what could be considered a 'normal' or 'typical' rate of syntactic development in college-level L2 writing? As mentioned, only MLTU was employed by all six longitudinal studies.

Table 7 shows a summary of the observed MLTU means, and a pre-to-post effect size associated with the observed means in each study, calculated by means of Cohen's *d* (see Norris and Ortega 2000). Studies in Table 7 are listed from shortest to longest observation period. Arthur (1979) provides insufficient report for the calculation of effect sizes. Therefore, inferences are being made here on the basis of seven comparisons contributed by five studies. Given the very limited data available, answers to Research Question 3 are purely exploratory.

In general, however, it is clear that the largest effect sizes, of over one standard deviation unit, were found for comparisons of L2 writing over the longest periods of time available: nine months (Kern and Schultz 1992) and three semesters (Casanave 1994). For periods of time of three months or less, effect sizes ranged from a negligible $d = 0.12$ by a group of EFL students (one of two intact classes studied in Ishikawa 1995), to a medium change of half a standard deviation over two and a half months by a group of ESL learners (Larsen-Freeman 1983, Study 3).

The longitudinal findings reported in Kern and Schultz's (1992) study of second-year French FL students are suggestive of different rates of syntactic complexification at different points over a year of instruction. Specifically, a repeated-measures comparison of MLTU means on argumentative essays written three months apart during the Fall semester yielded a small change ($d = .24$) that was statistically not significant (95 per cent confidence intervals largely overlapped), whereas a comparison of the same initial essays (in

September) with essays written by the same group nine months later (in May) resulted in a large and statistically significant change of one and a half standard deviations. Casanave's (1994) journal writing samples by four Japanese EFL students of an intermediate level were collected over an extended period of three semesters. No statistical analyses were performed in the original study, but a calculation of means and effect sizes for four students (based on individual scores reported there) showed a change similar in magnitude to that observed in Kern and Schultz (1992) over a similar period of time. In spite of the large magnitude of the difference, and unlike Kern and Schultz's results, the observed means in Casanave (1994) produced overlapping confidence intervals owing to the small sample size.

Overall, the results presented in Table 7 suggest that two to three months of university-level instruction may result in a negligible to small-sized change (up to .37 standard deviation units) in MLTU for FL samples, whereas for the same short period of time any change in MLTU may be expected to be of medium size (.50 standard deviation units) for ESL samples. That is, the rate

Table 7: Effect sizes for MLTU in six longitudinal studies of L2 writing

	Sample characteristics	Observation period	Time 1 MLTU mean (*SD*)	Time 2 MLTU mean (*SD*)	Effect size (*d*)
Arnaud (1992)	50 EFL students, freshman	2 months	14.61 (2.76)	15.59 (2.97)	0.34
Arthur (1979)[a]	14 ESL students, unknown proficiency	2 months	10.50 (n.r.)	9.99 (n.r.)	—
Larsen-Freeman (1983)	23 ESL students, level 4 (advanced)	2.5 months	13.77 (3.02)	15.37 (3.39)	0.50
Ishikawa (1995)	28 EFL students, low intermediate	3 months	7.97 (2.07)	8.20 (1.87)	0.12
	29 EFL students, low intermediate	3 months	7.15 (1.4)	7.80 (2.12)	0.37
Kern and Schultz (1992)	73 French FL students, second year	3 months	9.52 (2.34)	10.07 (2.30)	0.24
	(same students as above)	9 months	9.52 (2.34)	14.08 (3.72)	1.5
Casanave (1994)	4 EFL students, TOEFL 420–470	3 semesters	9.35 (1.59)	11.13 (1.21)	1.3

[a] Effect size could not be calculated due to insufficient reporting. n.r. = Not reported.

of change over time may be greater for ESL than for FL learners. When longitudinal development is investigated over a longer period of roughly a year of instruction, changes in MLTU may be small in the first few months, but substantial by the end of the observation period. As cautioned already, these inferences are purely exploratory and may be useful only as guiding hypotheses for future research.

DISCUSSION AND CONCLUSION

The present research synthesis yielded several substantive findings. First, ESL learners in the studies synthesized tended to produce writing of higher syntactic complexity than did FL learners. Second, studies that established proficiency group differences on the basis of holistic ratings tended to yield narrower ranges of observed complexity values and more homogeneous results across compared groups. Third, the following critical magnitudes were proposed for between-group differences in syntactic complexity for college-level writing in a second or foreign language: 4.5 or more words per sentence (MLS), 2 or more words per T-unit (MLTU), slightly over 1 word per clause (MLC), and at least a 0.20 positive or negative difference in number of clauses per T-unit (C/TU). Finally, it was suggested that two to three months of university-level instruction may result in a small MLTU change in ESL samples and an even smaller change in FL samples, whereas by the end of a one-year observation period, changes in syntactic complexity (at least as reflected in MLTU) may be substantial.

Researchers interested in using syntactic complexity measures as global indices of L2 proficiency may refer to these findings as interpretive landmarks for aiding study design and interpretation of study outcomes in future college-level L2 writing research. In this final section, I expand on some of the implications of these findings and finish the discussion with a cautionary note acknowledging the limitations of the study.

The importance of instructional setting and proficiency sampling criterion

What should be made of the finding that syntactic complexity values varied systematically depending on instructional setting and proficiency sampling criterion? The uncovered influence of instructional setting is important not only because it may reveal a sampling bias in the particular studies synthesized but also because it may reflect general characteristics of ESL and FL contexts as comprising distinct L2 populations. Namely, by comparison to second language instruction contexts, foreign language contexts are likely to involve students who start at generally lower levels of L2 proficiency and who typically do not study for as long as ESL learners. Further, learners in FL contexts are likely to undergo a generally slower pace of development and to achieve overall lower levels of ultimate attainment. Researchers who

investigate L2 writing in EFL (Hirose and Sasaki 1994; Ishikawa 1995; Kubota 1998) and FL settings (Reichelt 2001) have regularly pointed out the different nature of FL writing in comparison to ESL writing (see also Weigle 2002: 98, 133). The cumulative findings examined here lend some empirical support for the importance of this context variable.

The finding that primary investigations using a holistic rating proficiency criterion exhibited reduced variance helps to explain the preponderance of statistically non-significant observations for this kind of study previously noted by Wolfe-Quintero *et al.* (1998).[4] Namely, reasonable within- and between-group variance is a condition for the meaningful use of correlation-based inferential statistics (cf. Tabachnik and Fidell 1996). This amounts to saying that a methodological problem (too narrow or truncated sampling) with psychometric consequences (lack of variance) may prevent researchers from detecting in their data a positive relationship between syntactic complexity and L2 proficiency, even when this relationship might in fact be warranted at the descriptive and theoretical level. Consequently, L2 writing researchers interested in employing measures of syntactic complexity as global indices of L2 proficiency ought to invest efforts in devising study sampling strategies, as well as strategies for establishing proficiency categories within a study, that surmount the problem of homogeneous or truncated samples and ensure reasonable within- and between-group variance.[5]

Critical values for magnitude and rate in L2 syntactic complexity changes

Present applied linguistics quantitative research programmes appear to be driven less by developmental questions of magnitude and rate essential to the characterization of the range of normal variation in language development for specific populations and contexts, and more by the concern to establish statistical significance in the data. In light of the over-reliance on inferential statistical significance testing in quantitative investigations in our field, I feel it is important to restate that the critical values distilled in the present synthesis matter not necessarily because they were associated with statistically significant between-proficiency level differences (something that to a great extent is dependent on sample size), but most crucially because they were linked to consistent criterial differences in proficiency levels for the particular measures involved across studies of similar characteristics (e.g. target populations and designs). In other words, the importance of the proposed critical magnitudes and effect sizes resides in the fact that they can be regarded (albeit provisionally, until more data are available) as 'typical' of between-proficiency differences and differences over time in L2 syntactic complexity for these contexts. Moreover, these benchmarks can aid in the interpretation of results obtained in future college-level L2 writing studies with small

samples sizes and case studies based on a few individuals, where the use of inferential statistics would be inappropriate.

Non-linear relationships between syntactic complexity and L2 proficiency: broadening the scope of the research domain

After careful examination of the accumulated data, and of the themes discussed by primary researchers in the individual studies synthesized, I would like to suggest that the scope of the research domain needs to be broadened to accommodate two theoretically motivated predictions for cases in which statistical (or otherwise important) group differences are not to be found: a developmental prediction and a prediction of cross-rhetorical transfer.

The developmental prediction was proposed by Cooper (1976), Monroe (1975), Sharma (1980), and more recently Wolfe-Quintero *et al.* (1998: 73). It presents an argument for non-linear complexification as far as subordination is concerned, on the grounds that advanced proficiency groups should be expected to produce writing that capitalizes on complexification at the phrasal, rather than at the clausal, level. This developmental proposal finds independent theoretical support in the Hallidayan notions of grammatical metaphor and 'synoptic' and 'dynamic' styles.[6] Empirically, the prediction would be supported if future studies consistently show a curvilinear relationship between amount of subordination (C/TU) and phrasal length (MLC).

The cross-rhetorical transfer prediction is suggested here on the basis of findings reported in Neff *et al.* (1998). Essentially, Neff *et al.* (1998) concluded that the thirty fourth-year Spanish EFL writers in their study may have transferred from their L1 a pronounced preference for subordination over phrasal means of elaboration in formal writing styles, and this resulted in levels of C/TU which were unexpectedly high by comparison with those of L1 English writers, but very similar to C/TU levels found in a Spanish L1 baseline of newspaper writers. If these results can be replicated in future research, this would mean that whatever theoretically predicted relationships are proposed between proficiency and syntactic complexity (including the curvilinear model mentioned above), they will need to be modified to take into account cross-rhetorical transfer.

In sum, of particular importance to future research of syntactic complexity in L2 writing will be the interpretation of observed mean values in light of possible developmental and cross-rhetorical influences. This will require the combined interpretation, within single studies, of multiple measures tapping different sources of syntactic complexification, such as phrasal and clausal. Ideally, researchers interested in this research programme will also opt for rich designs similar to that of Neff *et al.* (1998) and will include within-subject comparisons of L1 and L2 writing by the same learners (see Kubota 1998) as well as comparisons with L1 baseline groups. Such designs would enable

researchers to better tease out the relative contribution of developmental and cross-rhetorical transfer influences and to explore complex interactions between them.

A word of caution: limitations of this research synthesis

This synthesis comprised a small number of studies with relatively varied study characteristics (cf. Table 3). There is some evidence indicating that a number of task-related factors affect the syntactic complexity levels of L2 writing (see review in Weigle 2002). Therefore, it is important to acknowledge that elicitation characteristics of particular writing tasks, as well as sample size, timing of the writing, corpus length, and the target languages investigated, all differed considerably across studies and that this variability probably introduced unidentified sources of error (i.e. construct-irrelevant variance that is unintended and/or unsystematic) in the study findings synthesized here. A further problem in the present synthesis is that, across individual studies, there was insufficient overlap in the specific measures investigated and not all measures reported were submitted to statistical analyses, which created gaps in the data available for synthesis.

An important additional consideration is the issue of unknown error within the primary studies investigated, given that the reliability of the interlanguage codings was generally not reported. This means that the amount and sources of error in the analytical process of quantifying syntactic complexity could not be estimated in the synthesis, and this error goes undetected and clouds the interpretation of effect sizes (see Thompson 1996). It is also important to note that each primary study contributed multiple and unequal observations to the synthesis (i.e. ten of the cross-sectional studies contributed means from three different proficiency groups, whereas three studies contributed only two, and the remaining eight studies contributed means from four to seven different proficiency groups). This means that unknown error contributed by certain studies may have disproportionately influenced the findings in the synthesis.

In order to enable better accumulation of knowledge in this research domain, sufficient numbers of primary studies will need to investigate consistently a core set of syntactic complexity measures and to report fully the extent and sources of error associated with these measures. Hopefully, the issues raised and discussed in this synthesis have offered insights for improved research practices.

The strength of adopting the methodology of research synthesis is that, rather than summarizing previous single-study observations in an additive fashion via narrative accounts or vote-counting techniques, a research synthesis 'can provide information that is *not available in any single study*' (Light and Pillemer 1984: 43, emphasis in original). I hope that the findings yielded by this synthesis will offer conceptually clear landmarks for interpretation of future study outcomes. I believe a fruitful line of research for future primary studies will be to engage in a more systematic examination

of magnitude and rate of grammatical development in L2 writing and of developmental and cross-rhetorical influences on L2 writers' syntactic repertoires. The accumulated evidence indicates that, if employed appropriately and interpreted meaningfully, global metrics of syntactic complexity can be valuable tools in such a research programme.

(Revised version received January 2003)

ACKNOWLEDGEMENT

This synthesis was part of my dissertation research, which was generously supported by a doctoral Mellon Fellowship at the National Foreign Language Center. My thanks go to my dissertation committee, particularly Mike Long and Craig Chaudron; and to Kate Wolfe-Quintero, Bill Grabe, Claire Kramsch, and three anonymous reviewers. I am most indebted to John Norris for suggesting to me the idea of using stem-and-leaf plots in this synthesis.

NOTES

1 Wolfe-Quintero *et al*. classified length-based indices (such as mean length of T-unit) as measures of fluency. I adhere to the more traditional position that length-based measures indeed tap syntactic complexity.

2 I am indebted to Gabriele Kasper for helping me make this connection for the synthesis.

3 In a comprehensive review of operationalizations of L2 proficiency in applied linguistics research, Thomas (1994) found that programme level, which she calls 'institutional status', was by far the most frequently employed criterion for establishing the proficiency of samples (40 per cent, or 63 of 157 studies). Standardized test scores were the second most used criterion, although much less frequent than programme level (only 22 per cent or 35 studies). It is unclear whether Thomas included some studies using holistic scores under her category 'in-house assessment instrument' (employed by only 14 per cent of her corpus). Thus, the use of holistic ratings to establish sample proficiency differences appears to be a practice distinctly favoured by L2 writing researchers.

4 Another reason is the small sample size typical of holistic rating studies, as shown in Table 6.

5 One anonymous reviewer expressed reservations about the usefulness of using programme level as a criterion to establish proficiency. While the use of programme levels for sampling and for establishing proficiency is not without problems (see Thomas 1994), for research where quantitative analyses are pursued, the basic problem of adequate within- and between-group variance is ameliorated when programme level is the strategy of choice. Further, programme levels across higher education institutions in the USA are defined in broadly similar ways, such as the traditional four years of foreign language or the typical five levels of many intensive English programmes. It should also be remembered that institutional status is used as a benchmark of developing abilities in much educational research, including, of course, the studies of school-aged children's language development, where syntactic complexity measures were first used in connection to writing.

6 I am grateful to Bill Grabe for drawing my attention to this affinity.

REFERENCES

*__Arnaud, P. J. L.__ 1992. 'Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests' in P. J. L. Arnaud and H. Bejoint (eds): *Vocabulary and Applied Linguistics*. London: Macmillan. pp. 133–45.

*__Arthur, B.__ 1979. 'Short-term changes in EFL composition skills' in C. Yorio, K. Perkins, and J. Schachter (eds): *On TESOL '79: The Learner in Focus*. Washington, DC: TESOL. pp. 330–342.

*__Bardovi-Harlig, K.__ and __T. Bofman.__ 1989. 'Attainment of syntactic and morphological accuracy by advanced language learners.' *Studies in Second Language Acquisition* 11: 17–34.

__Brown, R.__ 1973. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

*__Casanave, C.__ 1994. 'Language development in students' journals.' *Journal of Second Language Writing* 3: 179–201.

__Cooper, H.__ and __L. V. Hedges.__ (eds) 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

*__Cooper, T. C.__ 1976. 'Measuring written syntactic patterns of second language learners of German.' *Journal of Educational Research* 69: 176–83.

__Cumming, A.__ 1989. 'Writing expertise and second language proficiency.' *Language Learning* 39: 81–141.

*__Dvorak, T. R.__ 1987. 'Is written FL like oral FL?' in B. VanPatten, T. R. Dvorak, and J. F. Lee (eds): *Foreign Language Learning: A Research Perspective*. Rowley, MA: Newbury House. pp. 79–91.

*__Flahive, D. E.__ and __B. G. Snow.__ 1980. 'Measures of syntactic complexity in evaluating ESL compositions' in J. W. Oller and K. Perkins (eds): *Research in Language Testing*. Rowley, MA: Newbury House. pp. 171–6.

*__Gaies, S. J.__ 1976. Sentence-combining: A technique for assessing proficiency in a second language. Paper delivered at the Conference on Perspectives on Language, University of Louisville, Kentucky.

__Greenhouse, J. B.__ and __S. Iyengar.__ 1994. 'Sensitivity analysis and diagnostics' in H. Cooper and L. V. Hedges (eds): *The Handbook of Research Synthesis*. New York: Russell Sage Foundation. pp. 383–98.

*__Henry, K.__ 1996. 'Early L2 writing development: A study of autobiographical essays by university-level students of Russian.' *The Modern Language Journal* 80: 309–26.

__Hillocks, G. J.__ 1986. *Research on Written Composition: New Directions for Teaching*. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills/National Conference on Research in English.

*__Hirano, K.__ 1991. 'The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students.' *Annual Review of English Language Education in Japan* 2: 21–30.

__Hirose, K.__ and __M. Sasaki.__ 1994. 'Explanatory variables for Japanese students' expository writing in English: An exploratory study.' *Journal of Second Language Writing* 3: 203–29.

*__Homburg, T. J.__ 1984. 'Holistic evaluation of ESL compositions: Can it be validated objectively?' *TESOL Quarterly* 18: 87–107.

*__Ho-Peng, L.__ 1983. 'Using T-unit measures to assess writing proficiency of university ESL students.' *RELC Journal* 14: 35–43.

*__Ishikawa, S.__ 1995. 'Objective measurement of low-proficiency EFL narrative writing.' *Journal of Second Language Writing* 4: 51–69.

*__Kameen, P.__ 1979. 'Syntactic skill and ESL writing quality' in C. Yorio, K. Perkins, and J. Schachter (eds): *On TESOL '79: The Learner in Focus*. Washington, DC: TESOL. pp. 343–64.

__Kasper, G.__ 1997. 'The role of pragmatics in language teacher education' in K. Bardovi-Harlig and B. S. Hartford (eds): *Beyond Methods: Components of Language Teacher Education*. New York: McGraw-Hill. pp. 113–36.

*__Kern, R. G.__ and __J. M. Schultz.__ 1992. 'The effects of composition instruction on intermediate level French students' writing performance: Some preliminary findings.' *The Modern Language Journal* 76: 1–13. [Cross-sectional and longitudinal data.]

__Kramer, S. H.__ and __R. Rosenthal.__ 1999. 'Effect sizes and significance levels in small-sample research' in R. H. Hoyle (ed.): *Statistical Strategies for Small Sample Research*. Thousand Oaks, CA: Sage. pp. 59–79.

__Kubota, R.__ 1998. 'An investigation of L1–L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric.' *Journal of Second Language Writing* 7: 69–100.

__Larsen-Freeman, D.__ 1976. 'Evidence of the need for a second language acquisition index of development' in W. C. Ritchie (ed.): *Principles of Second Language Learning and Teaching*. New York: Academic Press.

*__Larsen-Freeman, D.__ 1978. 'An ESL index of development.' *TESOL Quarterly* 12: 439–48.

*__Larsen-Freeman, D.__ 1983. 'Assessing global second language proficiency' in H. W. Seliger and M. H. Long (eds): *Classroom-Oriented Research in Second Language Acquisition*. Rowley, MA: Newbury House. pp. 287–304. [Study 2, cross-sectional data; and Study 3, longitudinal data.]

*__Larsen-Freeman, D.__ and __V. Strom.__ 1977. 'The construction of a second language acquisition index of development.' *Language Learning* 27: 123–34.

__Light, R. J.__ and __D. Pillemer.__ 1984. *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.

*__Monroe, J. H.__ 1975. 'Measuring and enhancing syntactic fluency in French.' *The French Review* 48: 1023–31.

*__Neff, J., E. Dafouz, M. Díez, R. Prieto,__ and __C. Chaudron.__ 1998. Contrastive discourse analysis: Argumentative text in English and Spanish. Paper presented at the Twenty-Fourth Linguistics Symposium: Discourse across Languages and Cultures, University of Wisconsin-Milwaukee, September.

*__Nihalani, N. K.__ 1981. 'The quest for the L2 index of development.' *RELC Journal* 12: 50–6.

__Norris, J. M.__ and __L. Ortega.__ 2000. 'Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis.' *Language Learning* 50: 417–528.

*__Perkins, K.__ 1980. 'Using objective methods of attained writing proficiency to discriminate among holistic evaluations.' *TESOL Quarterly*, 14: 61–7.

*__Perkins, K.__ and __T. J. Homburg.__ 1980. 'Three different statistical analyses of objective measures of attained writing proficiency' in R. Silverstein (ed.): *Occasional Papers in Linguistics, No. 6*. Carbondale, IL: Southern Illinois University. pp. 326–37.

__Polio, C.__ 2001. 'Research methodology in second language writing research: The case of text-based studies' in T. Silva and P. K. Matsuda (eds): *On Second Language Writing*. Mahwah, NJ: Lawrence Erlbaum. pp. 91–115.

__Reichelt, M.__ 2001. 'A critical review of foreign language writing research on pedagogical approaches.' *The Modern Language Journal* 85: 578–98.

__Roca de Larios, J., J. Marín,__ and __L. Murphy.__ 2001. 'A temporal analysis of formulation processes in L1 and L2 writing.' *Language Learning* 51: 497–538.

*__Sharma, A.__ 1980. 'Syntactic maturity: Assessing writing proficiency in a second language' in R. Silverstein (ed.): *Occasional Papers in Linguistics, No. 6* . Carbondale, IL: Southern Illinois University. pp. 318–25.

__Shaw, P.__ and __E. T.-K. Liu.__ 1998. 'What develops in the development of second-language writing?' *Applied Linguistics* 19: 225–54.

__Silva, T.__ 1993. 'Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications.' *TESOL Quarterly* 27: 657–77.

__Tabachnick, B. G.__ and __L. S. Fidell.__ 1996. *Using Multivariate Statistics*. New York: HarperCollins.

*__Tedick, D. J.__ 1990. 'ESL writing assessment: Subject-matter knowledge and its impact on performance.' *English for Specific Purposes* 9: 123–43.

__Thomas, M.__ 1994. 'Assessment of L2 proficiency in second language acquisition research.' *Language Learning* 44: 307–36.

__Thompson, B.__ 1996. 'AERA editorial policies regarding statistical significance testing: Three suggested reforms.' *Educational Researcher* 25/2: 26–30.

__Valdés, G., P. Haro,__ and __M. P. Echevarriarza.__ 1992. 'The development of writing abilities in a foreign language: Contributions toward a general theory of L2 writing.' *The Modern Language Journal* 76: 333–52.

__Weigle, S. C.__ 2002. *Assessing Writing*. New York: Cambridge University Press.

__Wolfe-Quintero, K., S. Inagaki,__ and __H.-Y. Kim.__ 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Technical Report No. 17. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.