Research Reports

REPORT 69
AUGUST 2002

TEST OF ENGLISH AS A FOREIGN LANGUAGE™

The Performance of Native Speakers of English and ESL Speakers on the Computer-Based TOEFL and GRE General Test

Lawrence J. Stricker

The Performance of Native Speakers of English and ESL Speakers on the Computer-Based TOEFL and GRE General Test

Lawrence J. Strieker

Educational Testing Service
Princeton, New Jersey

Abstract

The purpose of this study was to replicate previous research on the construct validity of the paper-and-pencil version of the TOEFL test and extend it to the computer-based TOEFL. Two samples of GRE test takers were used: native speakers of English specially recruited to take the computer-based TOEFL, and ESL test takers who had routinely taken the computer-based TOEFL recently. Native speakers performed well on TOEFL, relative to ESL test takers and to the maximum possible scores on the test, and varied less in their test performance than did ESL test takers; TOEFL scores were highly but not perfectly correlated with the computer-based General Test scores for both groups of test takers; regressions of the General Test verbal scores on the TOEFL scores for ESL test takers were nonlinear, and the regressions of the other General Test scores were linear; and trends in the variances of the General Test verbal scores associated with TOEFL scores were also nonlinear, and the trends were either unsystematic or negatively linear for the other General Test scores. All of the findings are consistent with previous results with the paper-and-pencil TOEFL, support the construct validity of the computer-based TOEFL, and illuminate its interplay with ability tests for ESL test takers.

Key words: TOEFL, GRE General Test, ESL test takers, native speakers, construct validity, score interpretation

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service® (ETS®) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

❖   ❖   ❖

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Committee of Examiners. Its 13 members include representatives of the TOEFL Board, and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to oversee the review and approval of proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Committee of Examiners serve three-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Because the studies are specific to the TOEFL test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language and applied linguistics. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (2002-2003) members of the TOEFL Committee of Examiners are:

Lyle Bachman	University of California, Los Angeles
Deena Boraie	The American University of Cairo
Micheline Chalhoub-Deville *(Chair)*	University of Iowa
Jodi Crandall *(Ex Officio)*	University of Maryland, Baltimore
Catherine Elder	University of Auckland
Glenn Fulcher	University of Surrey
William Grabe	Northern Arizona University
Stan Jones	Carleton University
Keiko Koda	Carnegie Mellon University
Richard Luecht	University of North Carolina at Greensboro
Terry Santos	Humboldt State University
Merrill Swain	The University of Toronto
Richard Young	University of Wisconsin-Madison

To obtain more information about TOEFL programs and services, use one of the following:

Email: toefl@ets.org

Web site: http://www.toefl.org

Table of Contents

List of Tables

List of Figures

Introduction

The validity of the computer-based TOEFL test is of paramount importance. A great deal of validity evidence about the two paper-and-pencil versions of the test has accumulated over the years (e.g., Hale, Stansfield, & Duran, 1984). However, changes in the delivery format from paper-and-pencil to computer; in the test content, (e.g., restructuring the test from five sections in the original paper-and-pencil version to three sections in the later paper-and-pencil version and in the computer-based version, and adding an essay to the latter [TOEFL Programs and Services, 1997, 1998]); and in the test-taking population make the applicability of this evidence (much of it from studies more than two decades old) to the computer-based TOEFL uncertain.

Two related kinds of construct validity studies are of special interest. One kind involves the performance of native speakers of English on TOEFL. Native speakers of English are not the ultimate criterion group for an ESL test, because they vary in formal and informal education in English and in linguistic ability, even within the same English-speaking country (e.g., Alderson, 1980; Bachman, 1990; Hamilton, Lopes, McNamara, & Sherman, 1993; Lantolf & Frawley, 1985). Nonetheless, on average, native speakers of English should be reasonably proficient in English and perform well on an ESL test, both in absolute terms and relative to ESL speakers (Weir, 1988).

Indeed, three studies with the paper-and-pencil TOEFL found precisely that: Angoff and Sharon (1971) and Johnson (1977), employing college students at a western state college and at the University of Tennessee, respectively, and using the original five-section version of the test; and Clark (1977) employing college-bound high school students in New Jersey and using the later three-section version of the test. For example, in the Clark study the mean number of correct items (out of 150) on two test forms was about 135 (90%) for native speakers and about 89 (59%) for ESL speakers. However, Structure and Written Expression was somewhat more difficult than the two other sections for native speakers, with a mean of about 34 out of 40 (85%) versus about 48 out of 50 (96%) for Listening Comprehension and about 53 out of 60 (88%) for Reading Comprehension. Furthermore, though the total score and section scores were highly skewed for native speakers (see the distribution of the number of correct items in the Clark study in Figure 1), there was still substantial variation on certain section scores, with some test takers
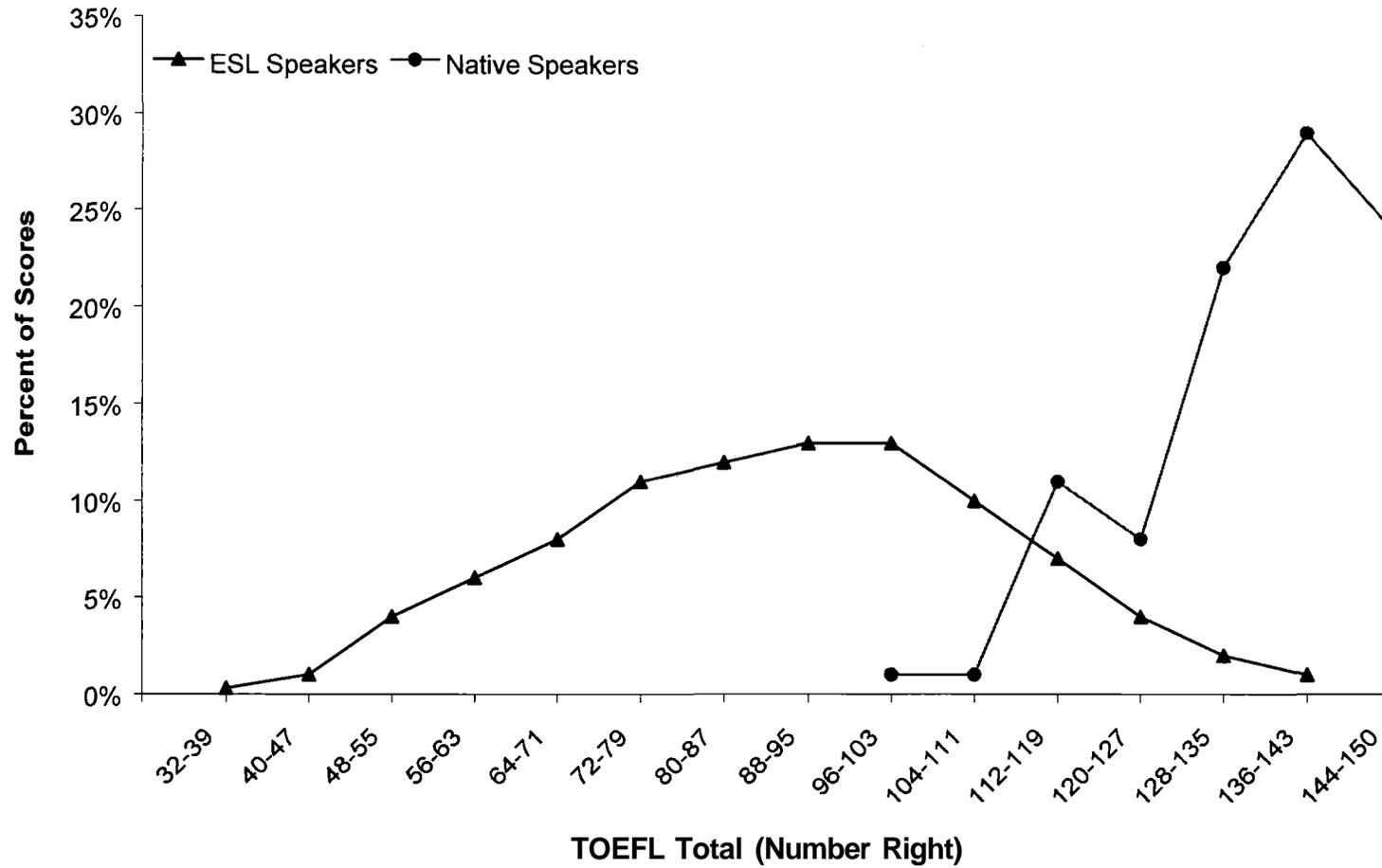
Figure 1. Distribution of TOEFL total scores for native speakers of English and ESL speakers. Adapted from Clark (1977); Form 1 and 2 data combined.

obtaining relatively low scores, though the variation was less than the variation for ESL speakers. For example, in the Johnson study, the standard deviation for Reading Comprehension was 5.8 for native speakers and 8.0 for ESL speakers, and the corresponding standard deviations for Listening Comprehension were 2.9 and 11.0, respectively.

The other kind of construct validity studies involve the relationship between TOEFL and English-language admissions tests, for native speakers of English as well as ESL speakers. An ESL test should correlate more highly with tests that have a high verbal load, such as verbal ability tests, than with other kinds of tests. In fact, several studies have reported high correlations between the five-section or later three-section versions of the paper-and-pencil TOEFL and verbal ability tests.

Studies of native speakers were conducted by Angoff and Sharon (1971) and Johnson (1977), both using the five-section version of TOEFL and the ACT. For example, in the Johnson study the TOEFL total score correlated .65 with the ACT composite score. Studies of ESL speakers, with the three-section version of the TOEFL, were done by Angelis, Swinton, and Cowell (1979) using the SAT and the GRE General Test, Powers (1980) using the Graduate Management Admission Test® (GMAT®), and Wilson (1982) using the GRE General Test and GMAT. For example, in the Wilson study, the TOEFL total score correlated .70 with the General Test's verbal score. Two of these studies also found higher correlations with a verbal ability test than with a quantitative ability test (Powers, 1980; Wilson, 1982). For example, in the Wilson study, the TOEFL total score correlated .21 with the General Test's quantitative score and .62 with its analytical score, compared to .70 with its verbal score. Note that these three studies of ESL speakers used the paper-and-pencil versions of the GRE General Test and GMAT, not the current computer-based versions, and the Wilson study used the original version of the General Test's analytical section, which had four item types, not the current version, which has two item types.

Furthermore, because an ESL test is easier than an English-language ability test and similar verbally loaded tests, and hence can discriminate among test takers at lower levels of ability, the ESL test should display unusual, curvilinear relations with English-language ability

tests that are useful in interpreting performance on these tests. The regression of the ability test on the ESL test ought to be concave: flat at the bottom of the ESL score range where there is little English-language ability and escalated at the higher end of the ESL score range. The variation in the ability test should follow the same trend: little variability at the bottom of the ESL score range and substantial variability at the higher end of the score range. The flatness in the regression and the minimal variability at the bottom of the ESL score range indicate that the ability test is not psychometrically viable at that range and its scores are not meaningful (Alderman, 1982; Angelis et al., 1979; Powers, 1980; Wilson, 1982).

The three TOEFL studies of ESL speakers discussed previously found such trends on verbal ability tests: Angelis et al. (1979) using the SAT and the GRE General Test, Powers (1980) using GMAT, and Wilson (1982) using the GRE General Test and GMAT. Powers failed to find such a trend for the GMAT quantitative score, but Wilson found such a trend for the General Test analytical score. As an illustration, Figure 2 shows the trends of the General Test verbal scores vis-a-vis the TOEFL total score in the Wilson study.

Accordingly, two related studies were carried out to replicate previous research on these issues and extend it to the computer-based TOEFL, using test takers from the GRE General Test population. General Test examinees are particularly relevant because of the large number that also take TOEFL relative to SAT and GMAT examinees, and the greater heterogeneity in the background and ability of General Test examinees relative to GMAT takers.

Study 1

The purpose of this study was to assess the comparative level of performance and variation in performance of native speakers of English and ESL speakers on the computer-based TOEFL, and the absolute level of performance of native speakers relative to maximum possible scores on this test.
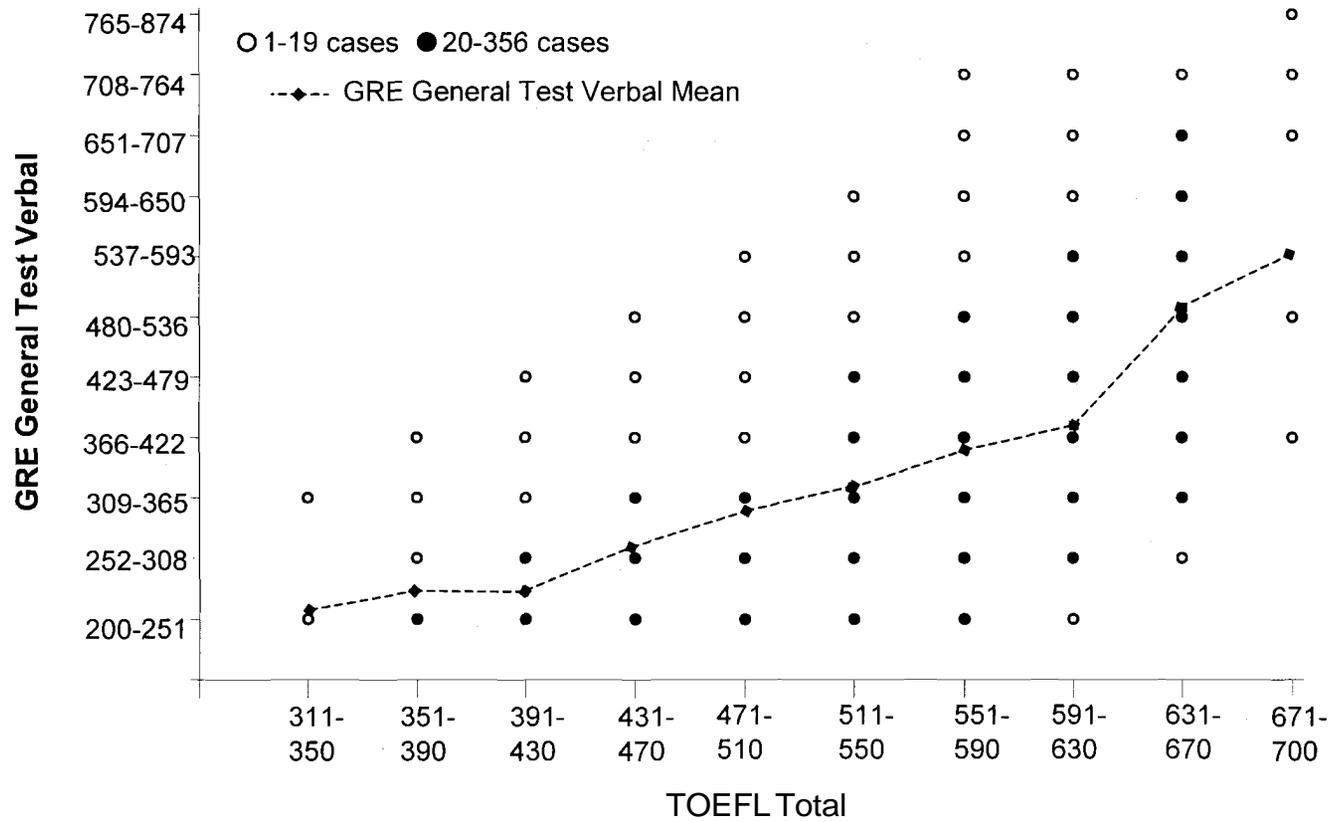
Figure 2. Regression of GRE General Test verbal scores on TOEFL total scores for ESL speakers. Adapted from Wilson (1982).

Method

Sample

Two samples of GRE General Test examinees were used:

1. A sample of 6,334 ESL speakers. These were all test takers who

    a. Took the computer-based General Test between September 1, 1999, and February 28, 2000; it was the first time they had taken the General Test, and they took it with no special accommodations for disabilities and without any irregularity (e.g., they did not request cancellation of their scores).

    b. Took the computer-based TOEFL between June 1, 1999, and May 30, 2000, it was the first time they had taken TOEFL, and they took it with no special accommodations and without any irregularity.

2. A total of 168 native speakers of English. These were all test takers who:

    a. Took the computer-based General Test between September 1, 1999, and February 28, 2000; it was the first time they had taken the General Test; they took it with no special accommodations; it was taken at one of five test centers (Case Western Reserve University, Cleveland OH; ETS-Bay Area, Oakland, CA; Miami-Dade Community college, Miami, FL; University of Minnesota, Minneapolis, MN; University of Missouri-Kansas City, Kansas City, MO), and they reported on the GRE background information form that they were U.S. citizens and that English was their best language ("Do you communicate better [or as well] in English than in any other language?"). ETS- and university-controlled testing centers in large cities were used to simplify data collection and increase the availability of test takers for subsequent testing with the computer-based TOEFL after the end of the school year.

    b. Took the computer-based TOEFL at the same testing center where they had taken the General Test, between March 23, 2000, and June 30, 2000, and without special accommodations; tested after being screened in a telephone interview to ensure that English was their native language ("Where were you born?", "What was the first language you learned?"); and received $75 for their participation, plus a bonus of $100 promised to the 10 of 200 scheduled test takers with the highest TOEFL total scores.

    These 168 test takers took both tests without any irregularity.

    The 168 test takers were among 189 students who had been recruited for the study. Two waves of mailings — March 10, 2000, and March 24, 2000 — each to 600 test takers, were sent

to randomly chosen test takers from the five testing centers who had taken the General Test without special accommodations during the specified period (September 1, 1999, to February 28, 2000) and reported on the background information form that they were U.S. citizens and that English was their best language. (A copy of the letter appears in the Appendix.) Participants were sent excerpts from the *1999-2000 TOEFL Information Bulletin for Computer-Based Testing* (Educational Testing Service, 1999) covering background information about the test, computer-based tutorials, and sample test items. Twenty-one test takers were excluded from the study because the General Test they had taken was not their first such test, it was not administered by computer, or they inadvertently canceled their TOEFL scores. Thirteen of the 189 test takers (all of whom were among the 168 in the study) received a bonus for earning the highest TOEFL total scores (300, the maximum possible score).

The characteristics of the ESL and native speakers of English samples are summarized in Table 1. The two samples were similar in age, but differed appreciably in their sex (more native speakers were women), their educational status (more native speakers were undergraduates or college graduates), and their General Test and TOEFL scores (all of the scores were lower for the ESL speakers, except for the General Test quantitative score).

Analysis

Differences between the mean TOEFL total score and the mean score on each of the three sections — Listening, Structure/Writing, and Reading — for the native speakers of English and the ESL speakers were evaluated by $\underline{t}$ tests of the two sample means. Corresponding differences between the variances of the scores for the two groups were evaluated by t tests based on the Brown-Forsythe procedure (1974).[1]

Differences between the mean TOEFL total scores and mean section scores for native speakers and maximum possible scores (300 for the total score, 30 for the section scores) were evaluated by one-sample t tests of the sample mean.

Both statistical and practical significance were considered in evaluating the results. For statistical significance, an .05 alpha level was used. For practical significance, a $\underline{d}$ of .20 and an $\underline{r}$

7

of .10 were used, representing Cohen's (1988) definition of a "small" effect size and accounting for 1% of the variance.

Table 1

Characteristics of Samples, Studies 1 and 2

| | Study 1 | | | | Study 2 | |
| | Native Speakers | | ESL Speakers | | ESL Speakers | |
Variable	N	Mean or Percent	AN	Mean or Percent	AN	Mean or Percent
Age (in years): Mean	168	26.24 (7.41)	6,332	25.93 (4.77)	3,487	25.78 (4.78)
Sex: Percent female	168	72.02	6,015	40.80	3,307	39.04
Educational Status (Percent):	168		5,895		3,240	
Undergraduate		38.69		23.94		24.85
College graduate		39.88		17.40		16.17
Some graduate work		13.69		38.91		39.17
Other		7.74		19.75		19.81
GRE General Test (Mean)						
Verbal	168	545.66 (108.48)	6,334	387.78 (107.43)	3,489	392.39 (109.71)
Quantitative	168	584.46 (132.45)	6,334	657.60 (128.80)	3,489	654.33 (131.51)
Analytical	168	624.05 (122.10)	6,334	550.81 (143.54)	3,489	551.50 (144.56)

(continued)

8

Table 1 (continued)

Variable	Study 1				Study 2	
	Native Speakers		ESL Speakers		ESL Speakers	
	N	Mean or Percent	N	Mean or Percent	N	Mean or Percent
TOEFL (Mean):						
Total Listening	168	281.68 (13.24)	6,334	244.70 (34.76)	3,489	246.98 (34.41)
Listening	168	28.08 (1.71)	6,334	23.93 (4.26)	3,489	24.22 (4.21)
Structure/Writing	168	28.70 (1.55)	6,334	24.35 (3.98)	3,489	24.52 (3.99)
Reading	168	27.73 (2.06)	6,334	25.14 (3.53)	3,489	25.36 (3.47)

<center>**Results**</center>

Differences Between Performance of Native Speakers and ESL Speakers

        The means and standard deviations for the TOEFL total and section scores for the native speakers of English and ESL speakers, t tests, and corresponding ds for mean differences and variance differences are reported in Table 2. The frequency distributions for the two groups are shown in Figures 3-6. Note that most of the TOEFL scores were moderately or highly correlated, with correlations ranging from .18 to .84 (all statistically, $p. < .05$, and practically, $r > .10$, significant).

        All of the differences in means were statistically ($p < .05$) and practically ($d > .20$) significant: native speakers consistently had higher means. The d for the total score was 1.08 (a large effect), and the ds for the section scores ranged from .74 for Reading to 1.11 for Structure/Writing (all medium or large effects).

        All of the differences in variances were also significant: native speakers had consistently smaller variances. The d for the total score was .68, and the ds for the section scores ranged from .44 for Reading to .66 for Structure/Writing (all medium effects).

        There was substantial skewness in the distributions of all TOEFL scores for the native speakers, with scores piled up at the top of the distributions, and minimal overlap between the two groups' distributions for all scores.

Differences Between Performance Levels of Native Speakers and Maximum Possible Scores

        The one-sample t tests and corresponding ds for the comparisons of the means for the TOEFL scores for the native speakers with maximum possible scores are also reported in Table 2. All of the mean differences were significant: native speakers had consistently lower means than the maximum scores. The d was 1.38 for the total score (a large effect), and the ds for the section scores ranged from .84 for Structure/Writing to 1.12 for Listening (all large effects).

<center>10</center>

Table 2

TOEFL Performance of Native and ESL Speakers

TOEFL Score	Native Speakers			ESL Speakers			Sample Differences				Native Speakers' Mean Differences From Maximum Score	
							Mean		Variance			
	Mean	SD	Variance	Mean	SD	Variance	t^a	3d	t^b			
Total	281.68	13.24	175.30	244.70	34.76	1207.98	13.76	1.08	8.65	.68	17.93	1.38
Listening	28.08	1.71	2.93	23.93	4.25	18.13	12.60	.99	7.92	.62	14.57	1.12
Structure/Writing	28.70	1.55	2.40	24.35	3.98	15.86	14.14	1.11	8.41	.66	10.82	.84
Reading	27.73	2.06	4.25	25.14	3.53	12.43	9.48	.74	5.59	.44	14.30	1.10

Note. N = 168 for native speakers and AN= 6,334 for ESL speakers. All ts are statistically significant (p <.05), and all CDs are practically significant (3d > .20)
[a] t test of two sample means. [b] t test of two sample variances based on the Brown-Forsythe (1974) procedure. [c] One-sample t test of the sample mean.
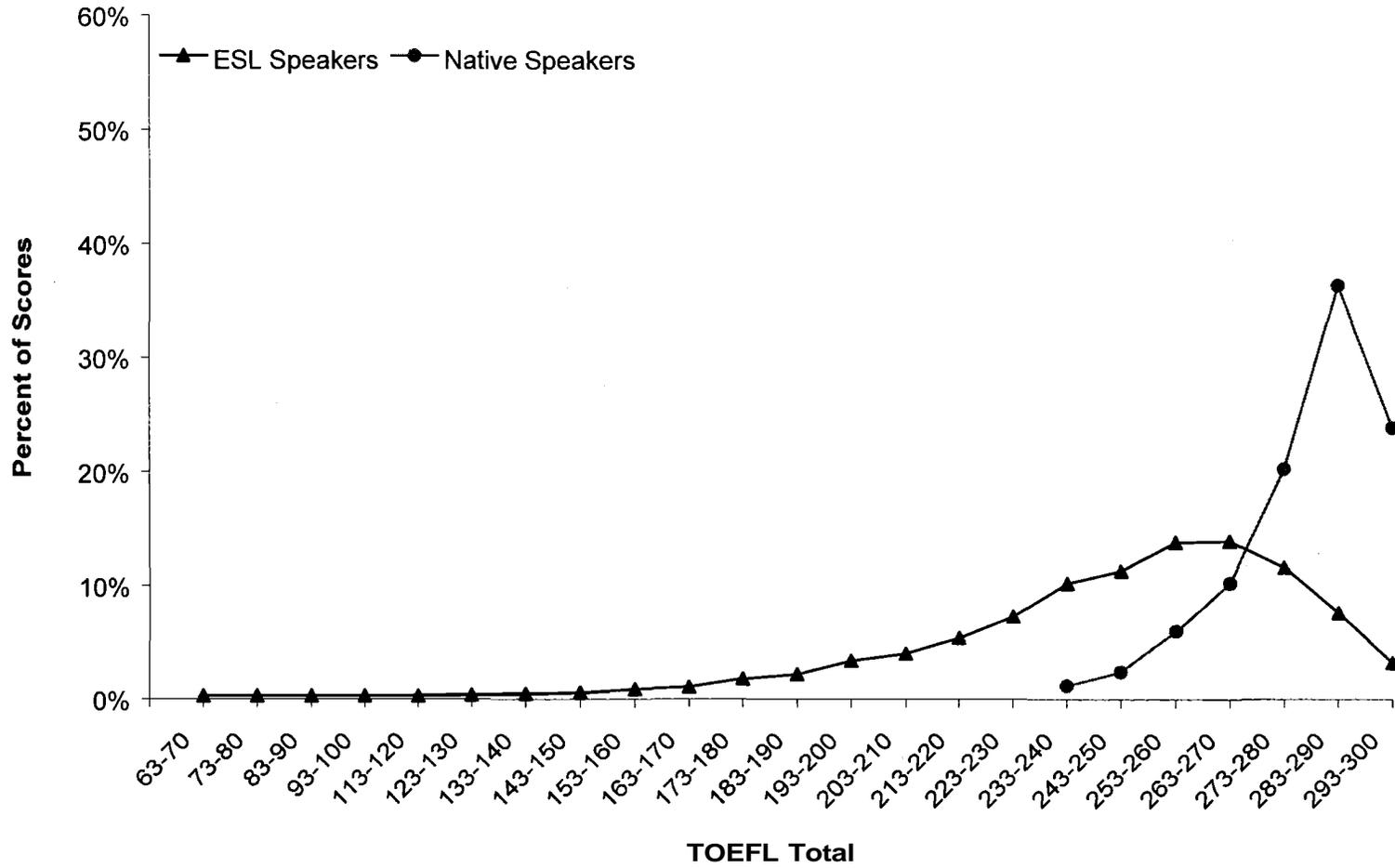
11

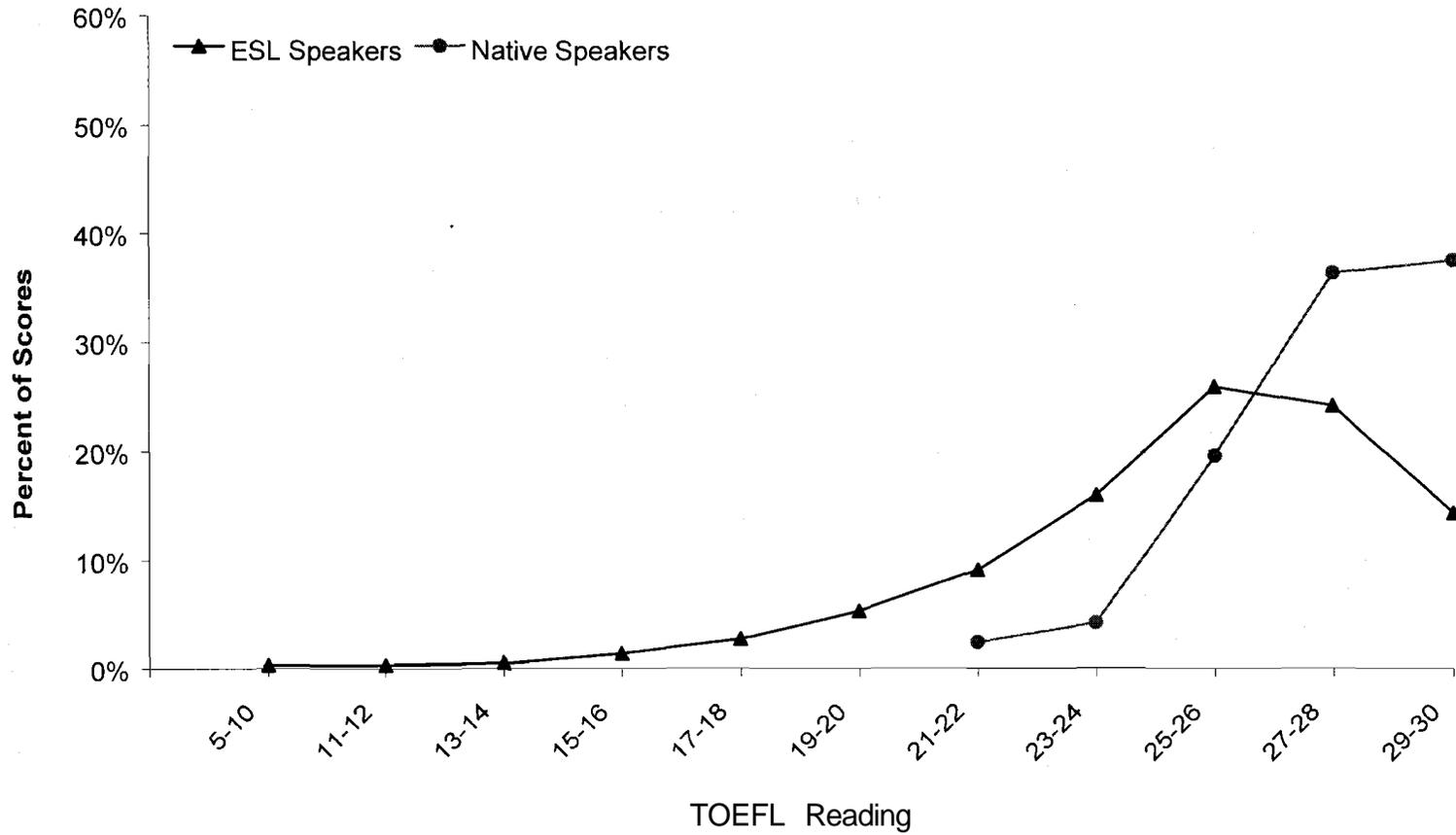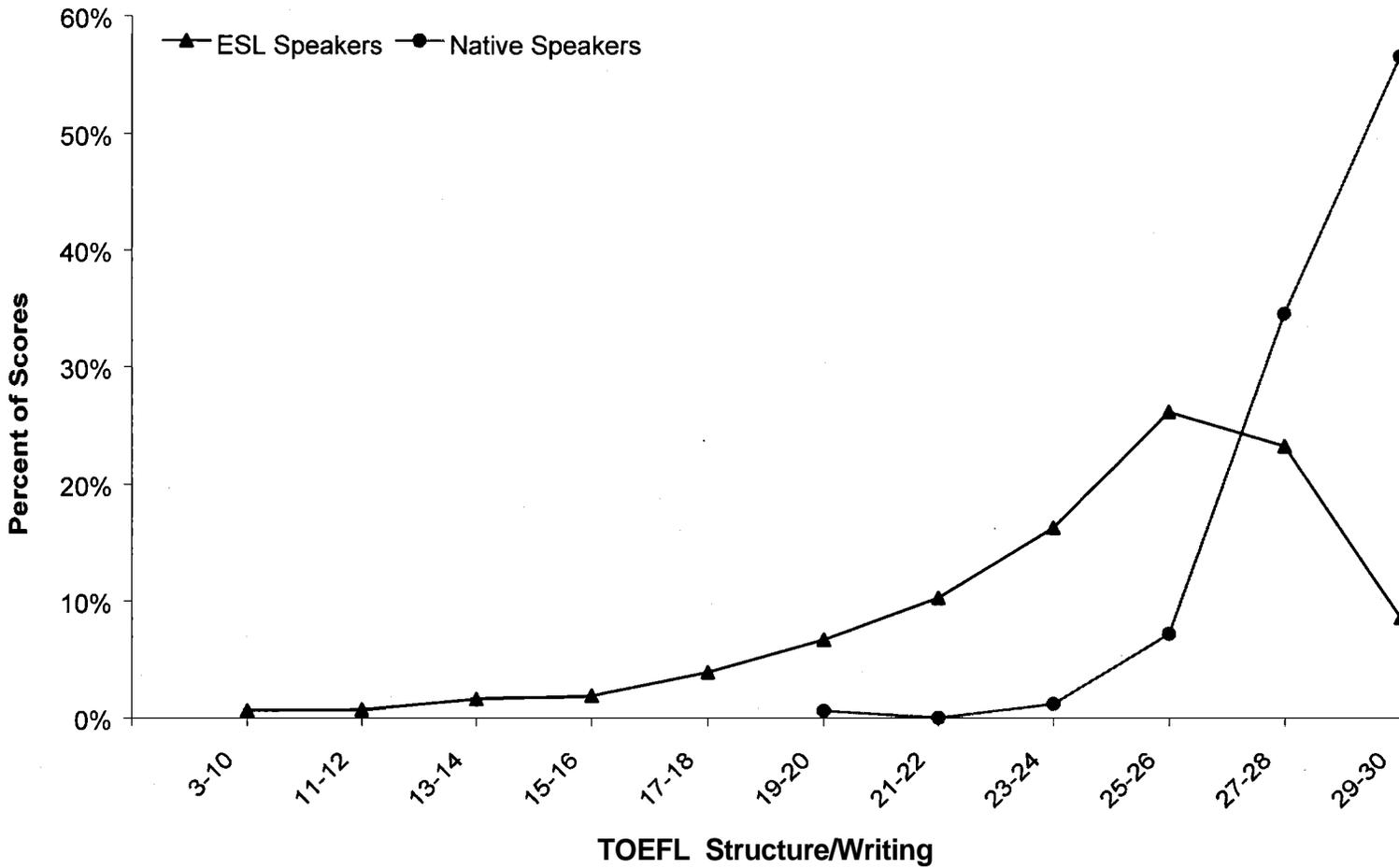*Figure 3.* Distribution of TOEFL total scores for native speakers of English and ESL speakers.

Figure 4. Distribution of TOEFL Reading scores for native speakers of English and ESL speakers.

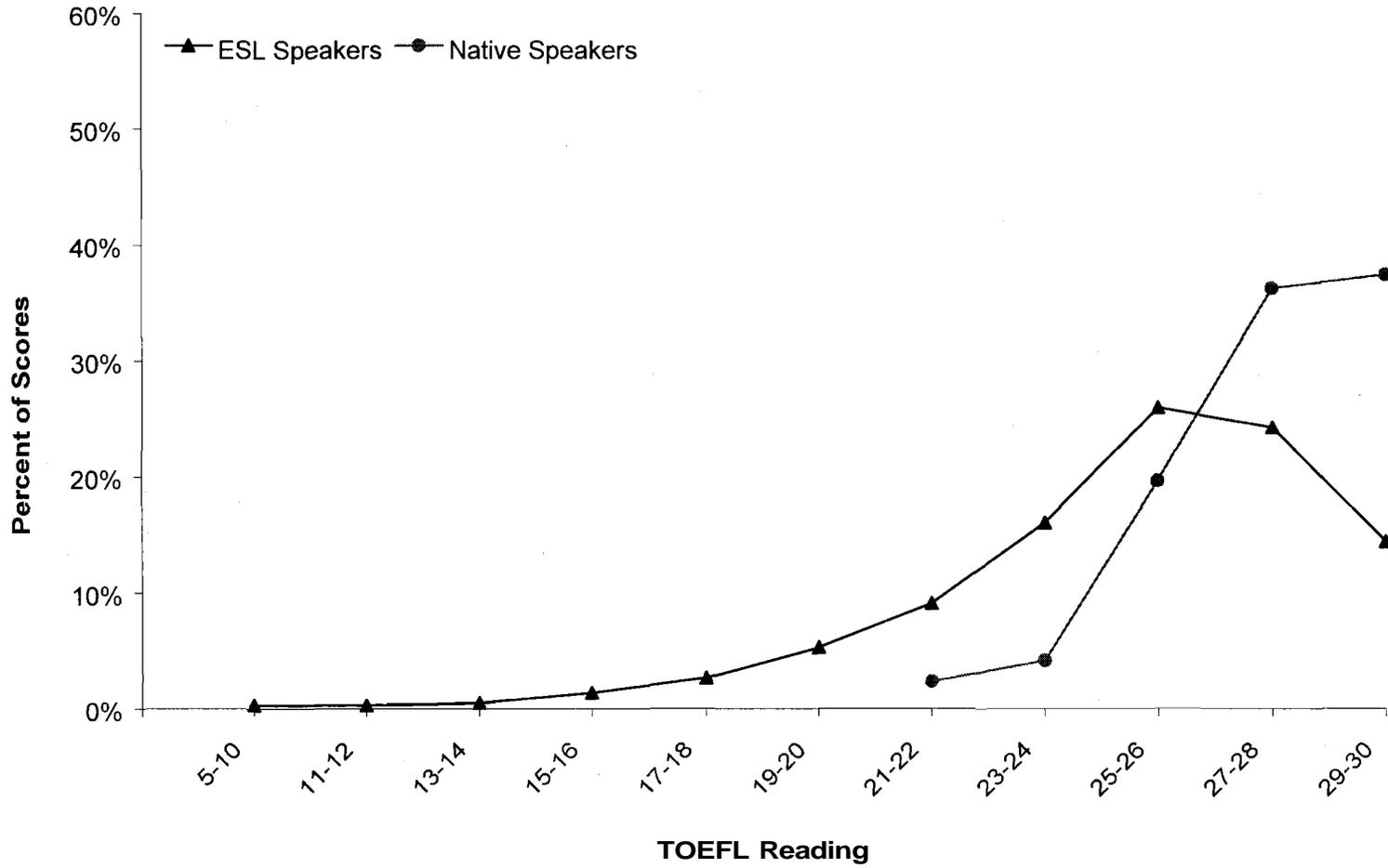*Figure 5.* Distribution of TOEFL Structure/Writing scores for native speakers of English and ESL speakers.

Figure 6. Distribution of TOEFL Reading scores for native speakers of English and ESL speakers.

Discussion

The present findings are broadly congruent with previous studies (Angoff & Sharon, 1971; Clark, 1977; Johnson, 1977) that found native speakers of English performed well on TOEFL, both in relative terms vis-a-vis ESL speakers and in absolute terms vis-a-vis the maximum possible scores on the test, and varied less in their test performance than did ESL speakers. However, there were some substantial differences from these studies in the pattern of results for the comparisons of native and ESL speakers. For instance, in the Johnson study, differences in mean performance for the native and ESL speakers were appreciably greater (e.g., the $\underline{d}$s for the corresponding sections of the test, Listening Comprehension in the Johnson study and Listening in the present study, were 1.82 and .99, respectively), as were differences in their variances (e.g., the ratios of native speaker variance to ESL speaker variance for this same section of the test were .07 in the Johnson study and .16 in the present study).

These and other divergent results are probably attributable in large part to differences in the populations of native and ESL speakers being compared. All of the previous studies contrasted native speakers (college-bound high school seniors [Clark, 1977] and college students [Angoff & Sharon, 1971; Johnson, 1977]) with all TOEFL test takers, whereas the present study contrasted native speakers with TOEFL test takers who were from the same GRE General Test population, making the two groups comparable in educational level, verbal ability, and the like.

One feature of the results deserves emphasis. Although the native speakers did well on TOEFL, they did not perform perfectly, and the variation in their test performance, though considerably less than that of ESL speakers, was nontrivial. These outcomes reinforce the point made previously (e.g., Alderson, 1980; Bachman, 1990; Hamilton et al., 1993; Lantolf & Frawley, 1985) that native speakers may vary in their education in English and linguistic ability.

Study 2

The purpose of this study was to evaluate the relationships between performance on the computer-based TOEFL and the GRE General Test for native speakers of English and ESL speakers.

Method

Sample

Two samples of General Test takers were used:

1. The sample of 168 native born speakers of English used in Study 1.

2. A sample of 3,489 ESL speakers from the larger sample of 6,334 ESL test takers in Study 1, limited to those who took the computer-based TOEFL and General Test within 15 days of each other. (This restriction was imposed to minimize discrepancies in performance on the two tests produced by English acquisition during the time between the administration of the tests.)

The characteristics of the two samples are summarized in Table 1. Paralleling the comparisons of native and ESL speakers in Study 1, the two samples were similar in age but differed appreciably in their sex, educational status, and General Test and TOEFL scores.

Analysis

The product-moment correlations of the TOEFL total and section scores with the General Test verbal, quantitative, and analytical scores were computed separately for the native and ESL speakers.

The regression of each General Test score on each TOEFL score for the ESL speakers was evaluated by analysis of variance, assessing both linearity and nonlinearity (Hays, 1994). In these analyses, TOEFL scores were grouped into intervals: nine for the total score, nine for Listening, eight for Structure/Writing, and seven for Reading.[2]

Heterogeneity in variance of each GRE score across the score range for each TOEFL score for the ESL speakers was also evaluated by analysis of variance, using the Brown-Forsythe (1974) procedure, and assessing both linear and nonlinear trends in the variances. TOEFL scores were grouped into the same intervals used in the regression analyses.

Both statistical and practical significance were again considered in evaluating the results, using the same guidelines as Study 1. For practical significance, an $\underline{r}$ or η of .10, a difference in correlations of approximately .10 (a difference of .10 in Z-transformed correlations), and a difference of .02 in variance proportions for η^2 and r^2 in assessing nonlinearity of regression or trends in variances were used. These values represent Cohen's (1988) definition of a small effect size and account for 1% of the variance (2% for the difference in variance proportions).

Results

Correlations of TOEFL With General Test

The correlations of the TOEFL scores with the General Test scores are reported in Table 3 for the native speakers of English and ESL speakers. All of the correlations of the TOEFL scores with the General Test scores were statistically ($p < .05$) and practically ($\underline{r} > .10$) significant for native speakers. For this group, the TOEFL scores' correlations with the General Test's verbal scores were consistently highest, with moderate or high correlations ranging from .33 to .61.

Paralleling the results for the native speakers, all of the TOEFL scores' correlations with the General Test scores for ESL speakers were significant, and the correlations with the General Test's verbal scores were consistently highest, with high correlations ranging from .54 to .64.

Most of the correlations were higher for ESL speakers than for native speakers, including four pairs of correlations that were significantly different (statistically, $2 < .05$, and practically, $|r_1 - r_2| > .10$) for the two groups. Three of the four pairs involved the General Test analytical score — TOEFL total, Structure/Writing, Reading — and one involved the General Test verbal score — TOEFL Structure/Writing. In contrast, one correlation was significantly lower for ESL speakers: General Test quantitative with TOEFL total.

Note that all of the TOEFL scores were highly correlated, ranging from .66 to .89, and most of the General Test scores were moderately to highly correlated, ranging from .27 to .66 (all significant).

18

Table 3

Correlations of TOEFL With GRE General Test for Native and ESL Speakers

TOEFL Score	Native Speakers			ESL Speakers		
	General Test			General		
	Verbal	Quantitative	Analytical	Verbal	Quantitative	Analytical
Total	.61	.44	.39	.64	.34	.53
Listening	.44	.34	.33	.54	.25	.44
Structure/Writing	.33	.19	.17	.56	.28	.44
Reading	.58	.42	.35	.62	.40	.53

Note. $\underline{N}$ = 168 for native speakers and $\underline{N}$ = 3489 for ESL: speakers. All correlations are statistically ($\underline{p}$ < .05) and practically ($\underline{r}$ > .10) significant.

19

Regressions of General Test Scores on TOEFL Scores

The analyses of variance of the regressions of the General Test scores on the TOEFL scores for ESL speakers are summarized in Table 4. All four of the regressions of the General Test verbal scores on the TOEFL scores were significantly nonlinear (statistically, $p < .05$, and practically, $\eta^2 - r^2 > .02$): concave in shape, flat at the bottom of the TOEFL score range, and escalating at the high end of the range. (These four regressions are shown in Figures 7-10.) All of the remaining eight regressions of the General Test quantitative and analytical scores on the TOEFL scores were significantly and positively linear. (The figures for these eight regressions are in the Appendix.)

Trends in General Test Score Variances Associated With TOEFL Scores

The analyses of variance based on the Brown-Forsythe (1974) procedure of trends in variances of General Test scores associated with TOEFL scores for ESL speakers are summarized in Table 5. In all 12 analyses, the variances were significantly heterogeneous (statistically, $p < .05$, and practically, n > .10) In three analyses, the trends were significantly nonlinear: General Test verbal score and analytical score variances associated with the TOEFL total score, and General test verbal score associated with the TOEFL Reading score. (These three trends are shown in Figures 11-13.) The two nonlinear trends for the General Test verbal score resembled the nonlinear regressions of this score. The nonlinear trend for the General Test analytical score, in contrast, had an inverted U shape, low at both the bottom and top of the TOEFL score range and high at the middle of the range.

In all but three of the nine remaining analyses, the variances were significantly linear (the exceptions were the General Test analytical score variances associated with all three TOEFL section scores). (The figures for these nine trends are in the Appendix.) The two linear trends for the variances of the General Test verbal score associated with the TOEFL Listening and Structure/Writing scores were positive, but all four of the linear trends for the variances of the General Test quantitative score associated with the TOEFL section scores were negative.

Table 4

Regressions of GRE General Test Scores on TOEFL Scores for ESL Speakers

Regression	η^a	$\underline{r}^b$	$\eta^2\text{-}\underline{r}^{2c}$
General Test Verbal on TOEFL Total	.69	.63	.08
General Test Quantitative on TOEFL Total	.35	.34	.01
General Test Analytical on TOEFL Total	.52	.52	.01
General Test Verbal on TOEFL Listening	.58	.52	.06
General Test Quantitative on TOEFL Listening	.25	.25	.00
General Test Analytical on TOEFL Listening	.44	.43	.01
General Test Verbal on TOEFL Structure/Writing	.59	.55	.04
General Test Quantitative on TOEFL Structure/Writing	.30	.28	.01
General Test Analytical on TOEFL Structure/Writing	.44	.44	.00
General Test Verbal on TOEFL Reading	.64	.60	.04
General Test Quantitative on TOEFL Reading	.40	.39	.01
General Test Analytical on TOEFL Reading	.53	.52	.01

Note. $\underline{N} = 3,489$. η, $\underline{r}$, and $\eta^2\text{-}\underline{r}^2$ differences that are statistically ($\underline{p} < .05$) and practically (η or $\underline{r} > .10$, $\eta^2\text{-}\underline{r}^2 > .02$) significant are underlined.
[a] η is the overall correlation between the General Test and TOEFL scores, corresponding to the $\underline{F}$ ratio in the one-way analysis of variance. [b] $\underline{r}$ is the linear correlation between the two scores. [c] $\eta^2\text{-}\underline{r}^2$ is the departure of the overall correlation from linearity.

Trend in Variances	η [a]	$\underline{r}$ [b]	$\eta^2\text{-}\underline{r}^2$ [c]
General Test Verbal on TOEFL Total	.38	.35	.02
General Test Quantitative on TOEFL Total	.18	-.16	.01
General Test Analytical on TOEFL Total	.14	.03	.02
General Test Verbal on TOEFL Listening	.33	.31	.01
General Test Quantitative on TOEFL Listening	.18	-.17	.00
General Test Analytical on TOEFL Listening	.11	-.01	.01
General Test Verbal on TOEFL Structure/Writing	.33	.32	.01
General Test Quantitative on TOEFL Structure/Writing	.12	-.11	.00
General Test Analytical on TOEFL Structure/Writing	.10	.04	.01
General Test Verbal on TOEFL Reading	.40	.38	.02
General Test Quantitative on TOEFL Reading	.20	-.18	.01
General Test Analytical on TOEFL Reading	.12	.05	.01

Note. $\underline{N}$ = 3,489. η, $\underline{r}$, and $\eta^2\text{-}\underline{r}^2$ differences that are statistically ($\underline{p} < .05$) and practically (η or $\underline{r} >$ .10, $\eta^2\text{-}\underline{r}^2 > .02$) significant are underlined.
[a] η is the overall correlation between the General Test deviation score and the TOEFL score, corresponding to the $\underline{F}$ ratio in the one-way analysis of variance (the Brown-Forsythe [1974] procedure). In this situation, it is an index of the heterogeneity of variance of the General Test scores across the score range for the TOEFL scores. [b] $\underline{r}$ is the linear correlation between the General Test deviation score and the TOEFL score. [c] $\eta^2\text{-}\underline{r}^2$ is the departure of the overall correlation from linearity.
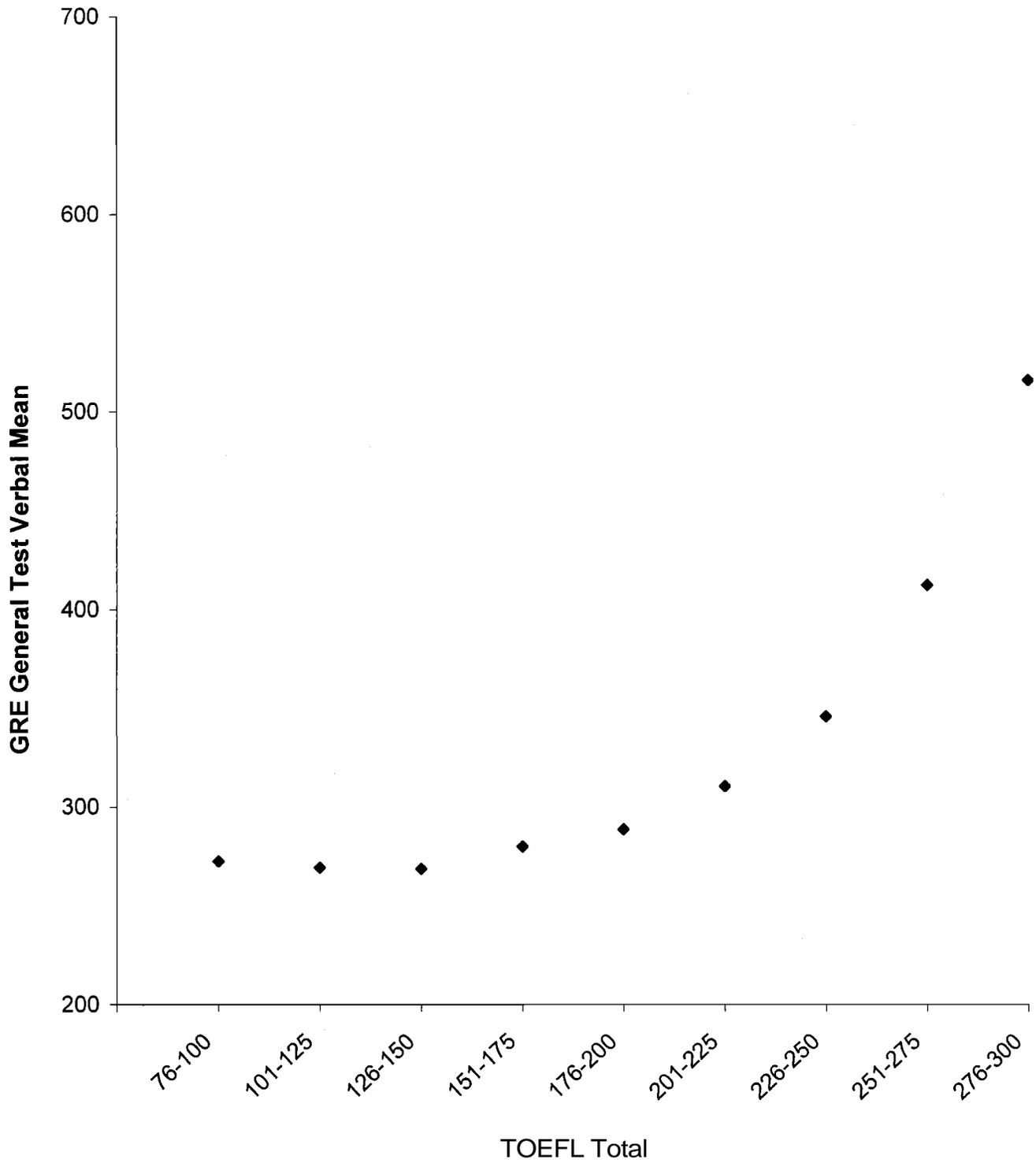
22
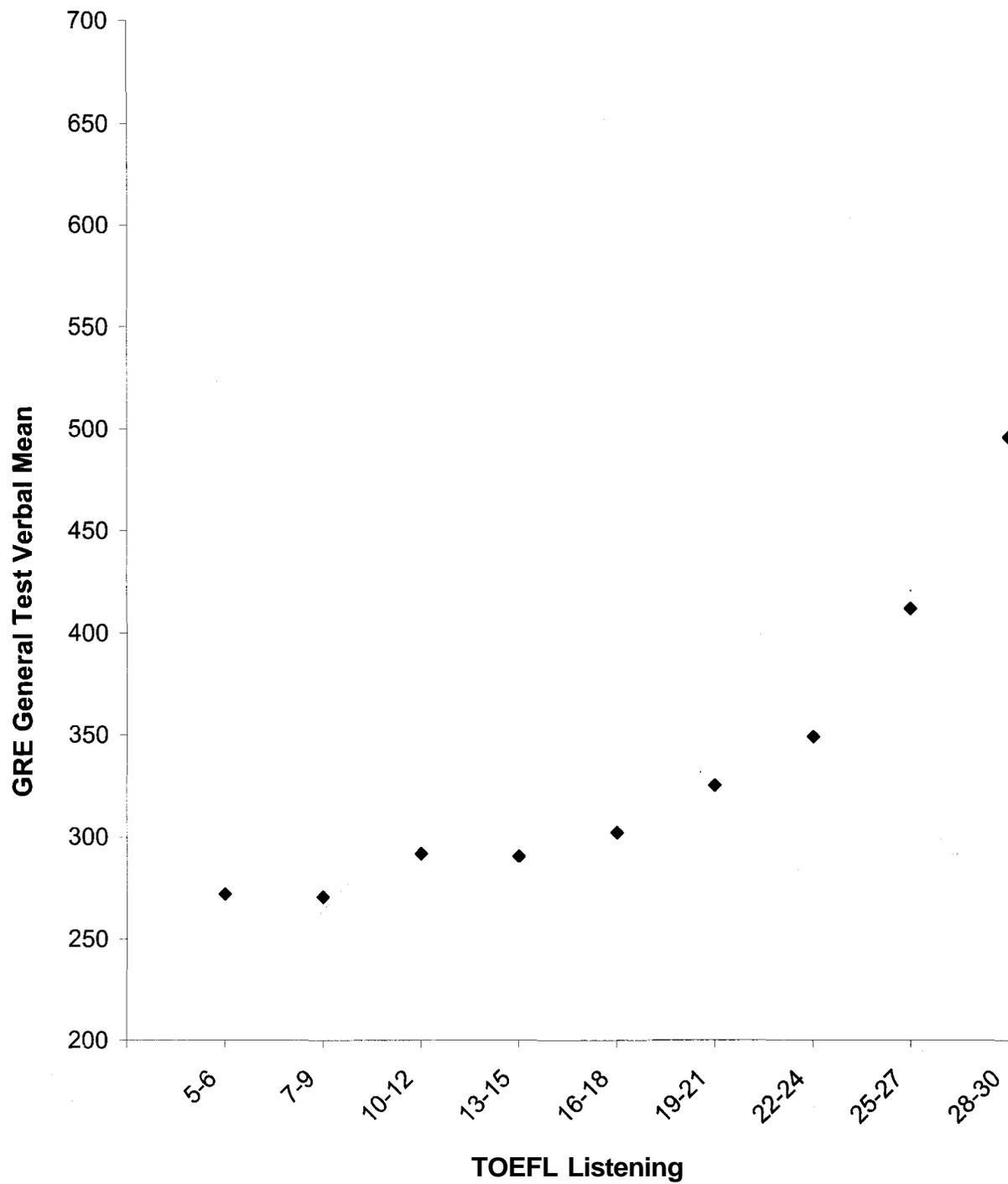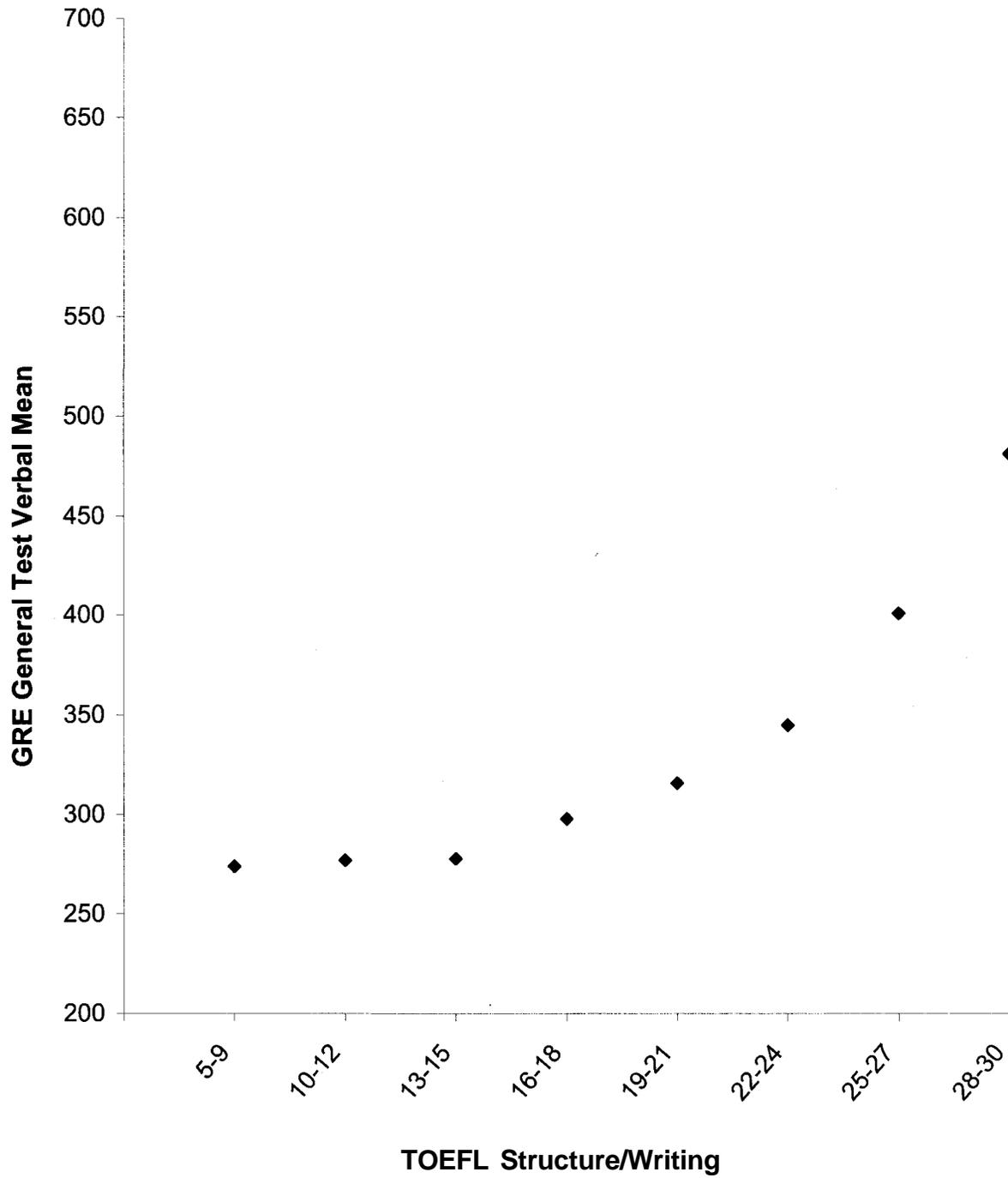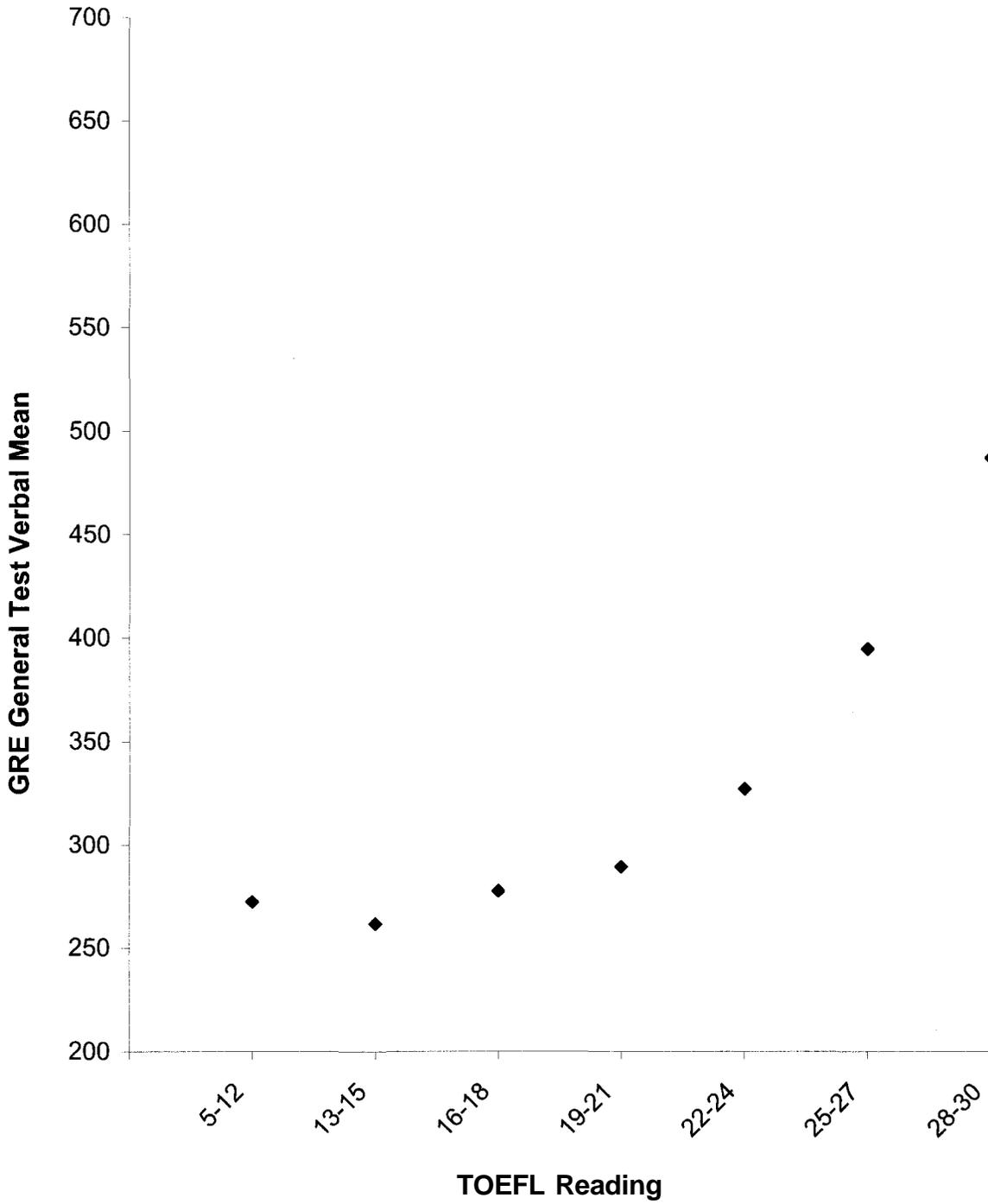
Figure 7. Regression of GRE General Test Verbal scores on TOEFL total scores for ESL speakers.

Figure 8. Regression of GRE General Test verbal scores on TOEFL Listening scores for ESL speakers.

Figure 9. Regression of GRE General Test verbal scores on TOEFL Structure/Writing scores for ESL speakers.

Figure 10. Regression of GRE General Test verbal scores on TOEFL Reading scores for ESL speakers.
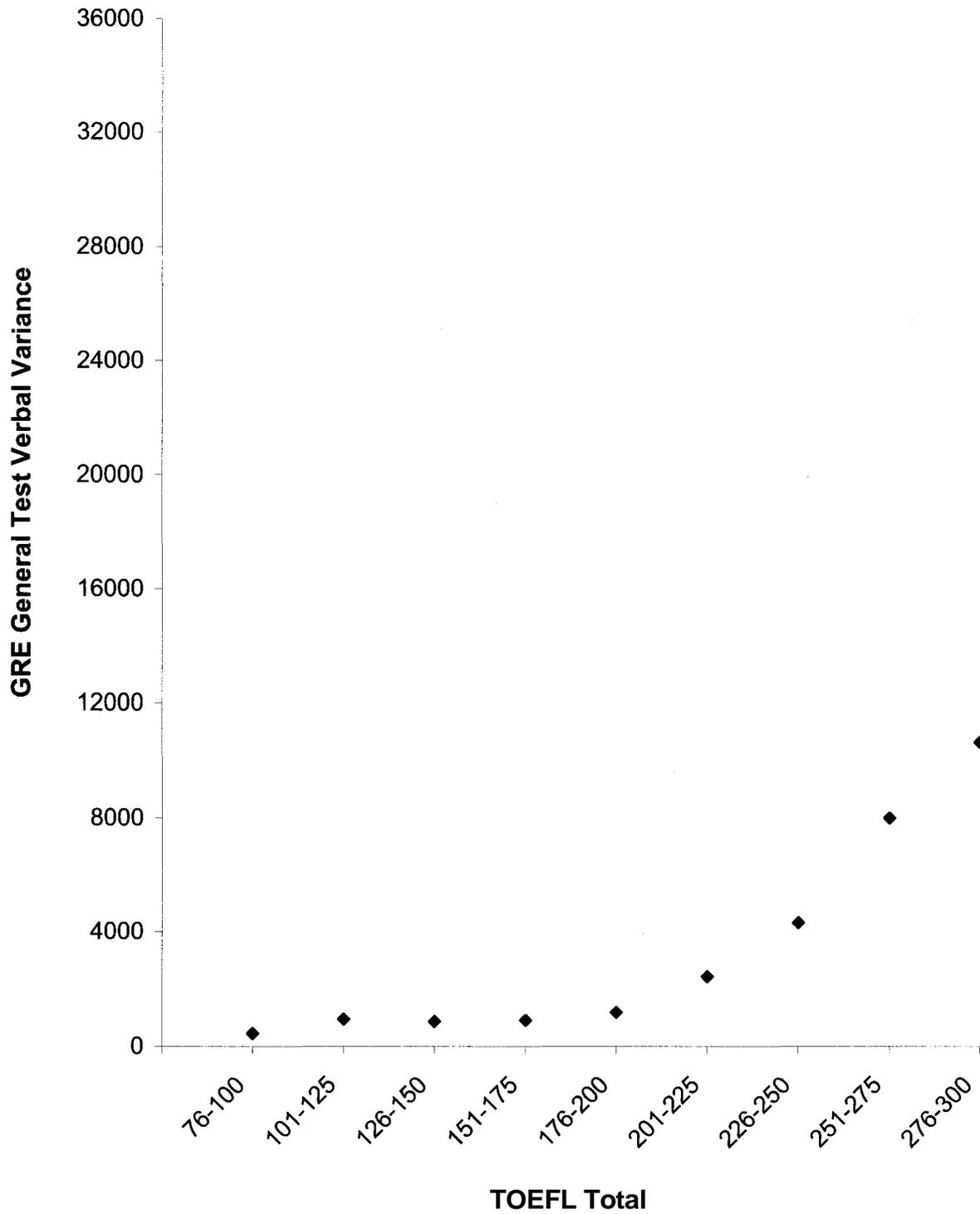
Figure 11. Trends in GRE General Test verbal score variances associated with TOEFL total scores for ESL speakers.
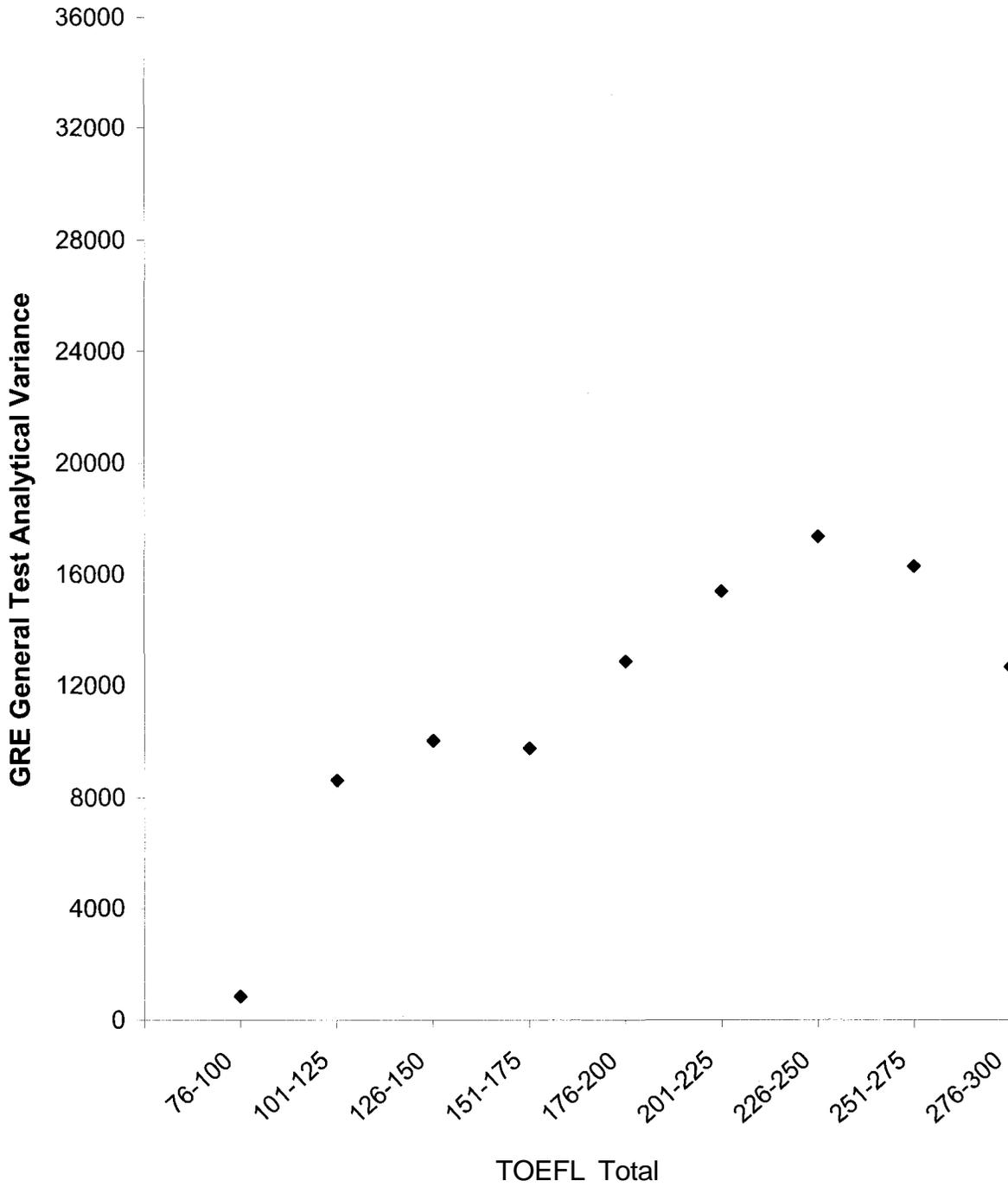
Figure 12. Trends in GRE General Test analytical score variances associated with TOEFL total scores for ESL speakers.
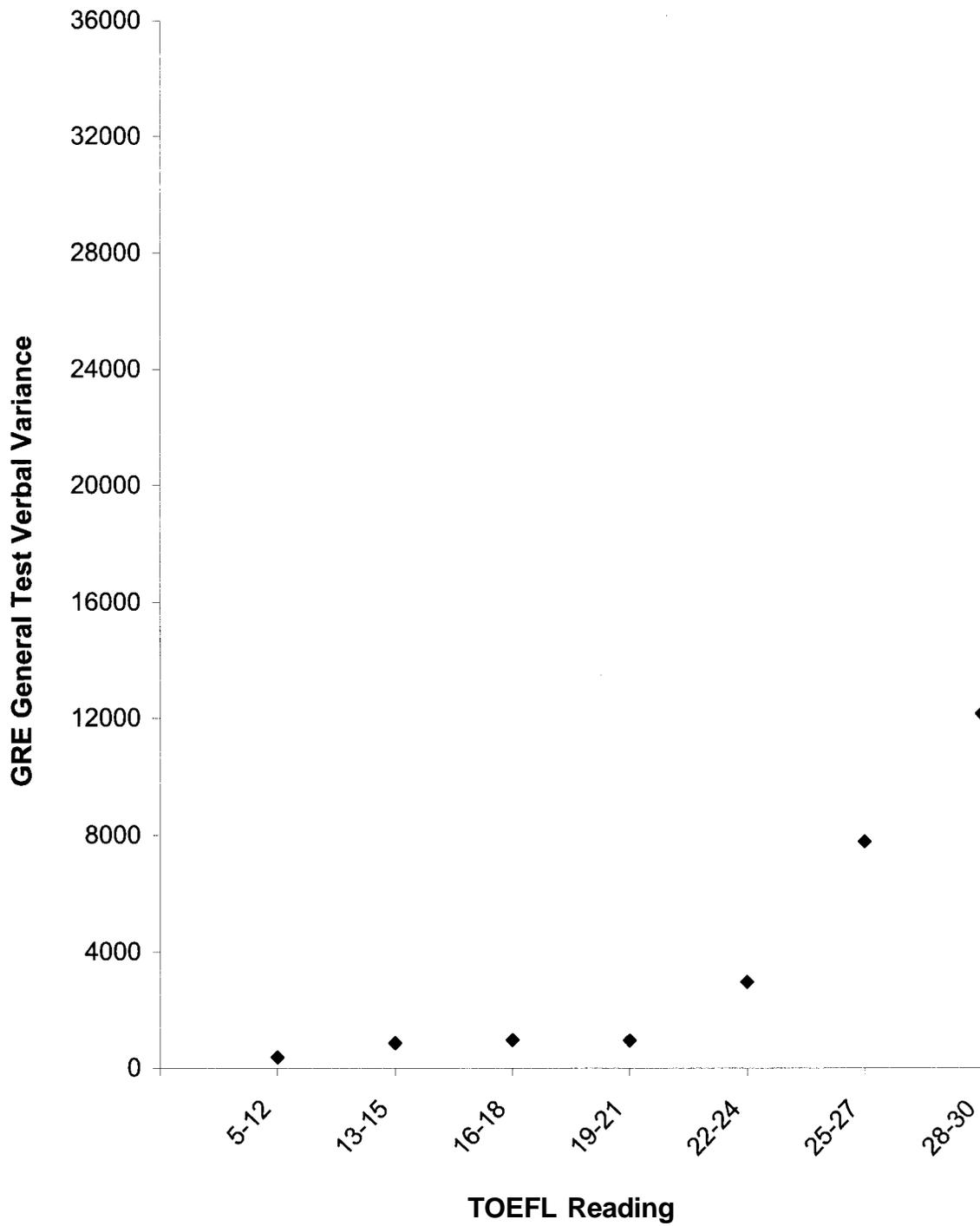
Figure 13. Trends in GRE General Test verbal score variances associated with TOEFL Reading scores for ESL speakers.

Discussion

The high correlations of the TOEFL total score with the GRE General Test verbal and analytical scores and its moderate correlation with the General Test quantitative score for ESL speakers are very similar to those reported previously (Wilson, 1982). The TOEFL section scores had the same pattern of correlations in this study. (Wilson did not report correlations for section scores.)

Although native speakers of English had the same differential pattern as ESL speakers of correlations of TOEFL scores with General Test verbal and quantitative scores, the correlations with the General Test analytical score were considerably lower for the native speakers than for the ESL speakers, suggesting that the General Test analytical score has a substantial verbal load for ESL speakers. These results for native speakers cannot be directly compared with previous findings, for there are no other studies with both TOEFL and General Test data for such test takers. However, Angoff and Sharon (1971) and Johnson (1977) found high correlations of the TOEFL total score with ACT scores for native speakers.

The correlational findings for the native and ESL speakers indicate that TOEFL has discriminant validity, given its higher correlations with the more verbally loaded sections of the General Test, as previously noted (Wilson, 1982). These correlations actually underestimate the associations with the verbal-loaded sections, in view of the skewness in the TOEFL scores for native speakers and the nonlinearity in the regressions of these General Test section scores observed in this study.[3]

The regression analyses uncovered the same nonlinearity for General Test verbal scores observed previously (Angelis et al., 1979; Wilson, 1982). In contrast to the present findings for the verbal score, the regressions were linear for the General Test quantitative and analytical scores. In the only other study that evaluated the regressions for either of these scores (Wilson, 1982), the analytical score had a nonlinear regression. A likely explanation for this discrepancy between the present study and the earlier one is the difference in the analytical section in the two studies: the present study used the current two-item type version, and the previous study used the original four-item type version.

The nonlinear regressions for the General Test verbal score reflect the unreliable, chance-level scores, as pointed out previously (Angelis et al., 1979; Wilson, 1982): the verbal score means were between 200 and 300 (on the GRE score range of 200 to 800) for test takers with low TOEFL scores (below 200 for the TOEFL total score). In contrast, most General Test quantitative and analytical scores were considerably higher, even for test takers with low TOEFL scores. The elevated General Test scores reflect the quantitative content of these sections of the test and the presumably good quantitative background of the ESL speakers (judging from international comparisons of mathematics achievement [e.g., LaPointe, Mead, & Askew, 1992; Medrich & Griffith, 1992; National Center for Education Statistics, 1998]), which enables them to achieve relatively high scores on these sections of the General Test despite their limited English proficiency.

The trends in the variances of the General Test scores are broadly congruent with the regression analyses. The nonlinear trends in the variances for the verbal score mirror the nonlinear regressions for this score. The unsystematic trends for the analytical score are consistent with the linear regression for this score. The negative linear trends for the quantitative score appears to reflect a ceiling effect: the quantitative score means were between 600 and 700 for test takers with high TOEFL scores (above 200 for the TOEFL total score). The same positive trends in General Test verbal and analytical score variances are apparent from visual inspection of the data reported previously (Angelis et al., 1979; Wilson, 1982); no quantitative analyses of these trends were made.

The nonlinearity in the regressions of the General Test verbal scores and in the trends of the variances of these scores suggests that verbal scores for test takers with low TOEFL scores cannot be meaningfully interpreted. The linearity in the regressions and in the trends of the variances of the General Test's quantitative and analytical scores, on the other hand, implies that these scores can be interpreted regardless of the test takers' TOEFL scores. This latter observation is obviously counterintuitive because both sections have verbal loads in their instructions and in the content of their items (Powers, 1980).

Judging from the inflection point in the nonlinear regressions and in the trends in variances for the General Test verbal score, it appears that a minimum TOEFL total score of approximately 200 is needed before the General Test verbal score can be interpreted. In a previous study (Wilson, 1982), a total score of 430 on the three-section version of TOEFL (using a different scoring metric from that used for the computer-based version) was identified in the same way from a regression analysis of the General Test verbal score. Precise comparisons of the two TOEFL total scores cannot be made because of differences in the scoring metrics, tests, and test-taking populations, but the present score seems to reflect a higher level of English proficiency. This score of 200 is 1.37 standard deviations below the mean of the sample and cuts off the bottom 10% of the distribution; the other score, of 430, is 2.08 standard deviations below the mean and cuts off the bottom 4%.

General Conclusion

The findings from these two studies are remarkably consistent with previous results, despite important changes in the delivery format and content of TOEFL and the General Test and in the test-taking population. Taken together, the findings support the construct validity of TOEFL and illuminate its interplay with ability tests for ESL test takers.

Although these results were based on the General Test population, it seems likely that comparable outcomes would be obtained with other test takers and other tests, given the general resemblance between these findings and earlier results with ACT, SAT, and GMAT test takers (Angoff & Sharon, 1971; Angelis et al., 1979; Johnson, 1977; Powers, 1980; Wilson, 1982).

References

Alderman, D. L. (1982). Language proficiency as moderator variables in testing academic
    aptitude. *Journal of Educational Psychology, 74,* 580-587.

Alderson, J. C. (1980). Native and nonnative speaker performance on cloze tests. *Language
    Learning, 30,* 59-76.

Angelis, P. J., Swinton, S. S., & Cowell, W. R. (1979). *The performance of non-native speakers
    of English on TOEFL and verbal aptitude tests* (TOEFL Report 3; ETS RR 79-7).
    Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Sharon, A. T. (1971). A comparison of scores earned on the Test of English as
    a Foreign Language by native American college students and foreign applicants to U.S.
    colleges. *TESOL Quarterly, 5,* 129-136.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England:
    Oxford University Press.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the
    American Statistical Association, 69,* 364-367.

Clark, J. L. D. (1977). *The performance of native speakers of English on the Test of English as a
    Foreign Language* (TOEFL Report 1). Princeton, NJ: Educational Testing Service.

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

Educational Testing Service. (1999). *1999-2000 TOEFL information bulletin for computer-based
    testing.* Princeton, NJ: Author.

Hale, G. A., Stansfield, C. W., & Duran, R. P. (1984). A comprehensive TOEFL bibliography,
    1963-82. *Modern Language Journal, 68,* 45-51.

Hamilton, J., Lopes, M., McNamara, T., & Sherman, E. (1993). Rating scales and native speaker
    performance on a communicatively oriented EAP test. *Language Testing, 10,* 337-353.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Johnson, D. C. (1977). The TOEFL and domestic students: Conclusively inappropriate. *TESOL
    Quarterly, 11,* 79-86.

Keppel, G. (1991). *Design and Analysis—A researcher's handbook* (3rd ed.). Upper Saddle
    River, NJ: Prentice Hall.

Lantolf, J. P., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *Modern Language Journal, 69,* 337-345.

LaPointe, A.E., Mead, N.A., & Askew, J.M. (1992). *Learning mathematics* (Report 22-CAEP—01). Princeton, NJ: Educational Testing Service, International Assessment of Educational Progress.

Levene, H. (1960). Robust tests for the equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278-292). Palo Alto, CA: Stanford University Press.

Medrich, E.A., & Griffith, J.E. (1992). *International mathematics and science assessments: What have we learned?* (NCES 92-011). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

National Center for Education Statistics (1998). *Pursuing excellence —A study of U.S. twelfth-grade mathematics and science achievement in international context.* (NCES 98-049). Washington, DC: U.S. Government Printing Office.

Powers, D. E. (1980). *The relationship between scores on the Graduate Management Admission Test and the Test of English as a Foreign Language* (TOEFL Report 5; ETS RR 80-31). Princeton, NJ: Educational Testing Service.

TOEFL Programs and Services. (1997). *TOEFL test and score manual, 1997 edition.* Princeton, NJ: Educational Testing Service.

TOEFL Programs and Services. (1998). *Computer-based TOEFL score user guide.* Princeton, NJ: Educational Testing Service.

Weir, C. (1988). The specification, realization, and validation of an English language proficiency test. In A. Hughes (Ed.), *Testing English for university study* (pp. 45-110). London: Modern English Publications and British Council.

Wilson, K. M. (1982). *GMAT and GRE Aptitude Test performance in relation to primary language and scores on TOEFL* (TOEFL Report 1; ETS RR 82-28). Princeton, NJ: Educational Testing Service.

Notes

'The Brown-Forsythe procedure is based on an earlier and more well-known test for differences among variances developed by Levene (1960). The Brown-Forsythe procedure analyzes, with conventional t tests in the case where there are two samples or with one-way analyses of variance in the case where there are several samples, the absolute difference between each score and the sample median. A significant I or FA indicates that the variances are heterogeneous. The Brown-Forsythe procedure is more powerful and less sensitive to departures from normality than are other tests used previously to evaluate heterogeneity of variance (Keppel, 1991).

[2] Intervals were equal, except for the lowest interval for Structure/Writing and for Reading, which combined adjacent intervals with small frequencies.

[3] The GRE verbal scores of ESL speakers are also substantially skewed. The mean is 392.39, the median is 360.09, and the skewness index, 3 (mean-median)/standard deviation, is .88.

Appendix

Educational Testing Service (ETS)
March 10, 2000

Dear GRE Test Taker:

We need your help and we'll pay you for it! We're doing research on the Test of English as a Foreign Language (TOEFL). This is a test of English proficiency for people whose native language is not English, widely used in colleges and universities in this country. We're interested in how well American-born speakers of English do on the test. We're inviting you to take part in this research because you recently took the GRE General Test at one of the six testing centers in the study.

To take part in the study, (1) you must have been born in the 50 states, (2) English must be the first language you learned, and (3) you must take the TOEFL (free of charge) before July in the same testing center where you took the GRE. The TOEFL is computer-administered, covers listening, reading, and writing an essay, and takes about four hours to complete. We'll send you a check for $75 as soon as we get your TOEFL scores. At the end of the study, we'll send the ten test takers with the highest total scores (in a group of about 200 test takers) an additional check for $100.

We hope that you will take part in the study if you are eligible. If you'd like to participate:

1. Call Peggy Redman at 609-734-5785 as soon as possible. She will give you a "CBT" voucher number to pay for the test.

2. Call your testing center (listed on back) and make an appointment to take the TOEFL at your convenience, using the voucher number. Ms. Redman will send you a letter confirming that you are in the study and some material to familiarize you with the test and the testing procedures (they are similar to those for the GRE). You don't need to study for the test.

3. Go to the testing center and take the test. Bring the letter and the same kind of photo-bearing ID you used when you took the GRE. (Acceptable ID are a current driver's license, student or employee ID card, or current passport; an expired driver's license or passport, Social Security card, and credit card are not acceptable).

Your TOEFL scores will be kept confidential and won't be available to colleges or universities. Please call Ms. Redman if you have any questions about the study.

Sincerely,
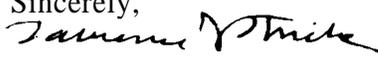
Lawrence J. Strieker, Ph.D.
Principal Research Scientist

Testing Centers in TOEFL Research Study

Cleveland, OH
Case Western Reserve University
Educational Support Services
Kelvin Smith Library
11055 Euclid Avenue
Cleveland, OH 44106
(216) 368-1030

Columbia, SC
University of South Carolina
Orientation and Testing Office
345 Russell House
Columbia, SC 29208
(803) 777-2782

Kansas City, MO
University of Missouri
Counseling and Testing Center, Suite 206
4825 Troost Avenue
Kansas City, MO 64110
(816) 235-1635

Miami, FL
Miami Dade Community College-North
11380 N.W. 27th Avenue
Miami, FL 33167
(305) 237-1015

Minneapolis, MN
University of Minnesota
9 Eddy Hall
192 Pillsbury Drive, SE
Minneapolis, MN 55455
(612) 624-5520

Oakland, CA
Educational Testing Service
1000 Broadway, Suite 310
Oakland, CA 94607
(510) 873-8100

Educational Testing Service **ETS**
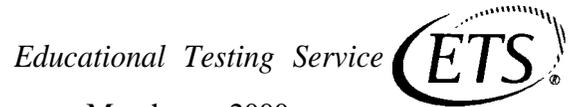
March _, 2000

Dear_____:

Thanks for agreeing to participate in our TOEFL research. Your "CBT" voucher number to pay for the TOEFL is_____. I have enclosed some material to familiarize you with the test and the testing procedures. You don't need to study for the test.

If you haven't already done so, please call the testing center where you took the GRE and make an appointment to take the TOEFL. Keep in mind that you must take the test before July. As soon as you get your appointment number from the testing center, please fill out the enclosed forms and return them to me in the accompanying envelope.

When you go to the testing center for the test, show the test center administrator this letter and the same kind of photo-bearing ID you used when you took the GRE (current driver's license, student or employee ID card, or current passport). Then take the test and follow all of the testing procedures, exactly like regular TOEFL test takers. However, do not cancel your scores or designate any institution to receive reports of the scores.

We'll send your check for $75 when we get your scores (about three to six weeks after the test). Remember that you'll get an extra $100 if you're one of the top ten scorers.

Please call me (609-734-5785) if you have any questions. And please call me as soon as possible if you decide, for any reason, not to take part in the study so that we can get a replacement.

Sincerely,
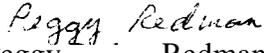
Peggy Redman

Peggy   .   Redman
Research Associate

Enc.

Dear Test Center Administrator:

The bearer is participating in the TOEFL research study about which ETS informed you. Please submit an Electronic Irregularity Report (EIR) #19 in which you report using this letter and the ID required for the GRE to admit him or her to the test center instead of the more extensive ID required for the TOEFL. Please call me if you have any questions or problems with this request (609-734-5785).

Sincerely,

Peggy Redman

Peggy   .   Redman
Research Associate
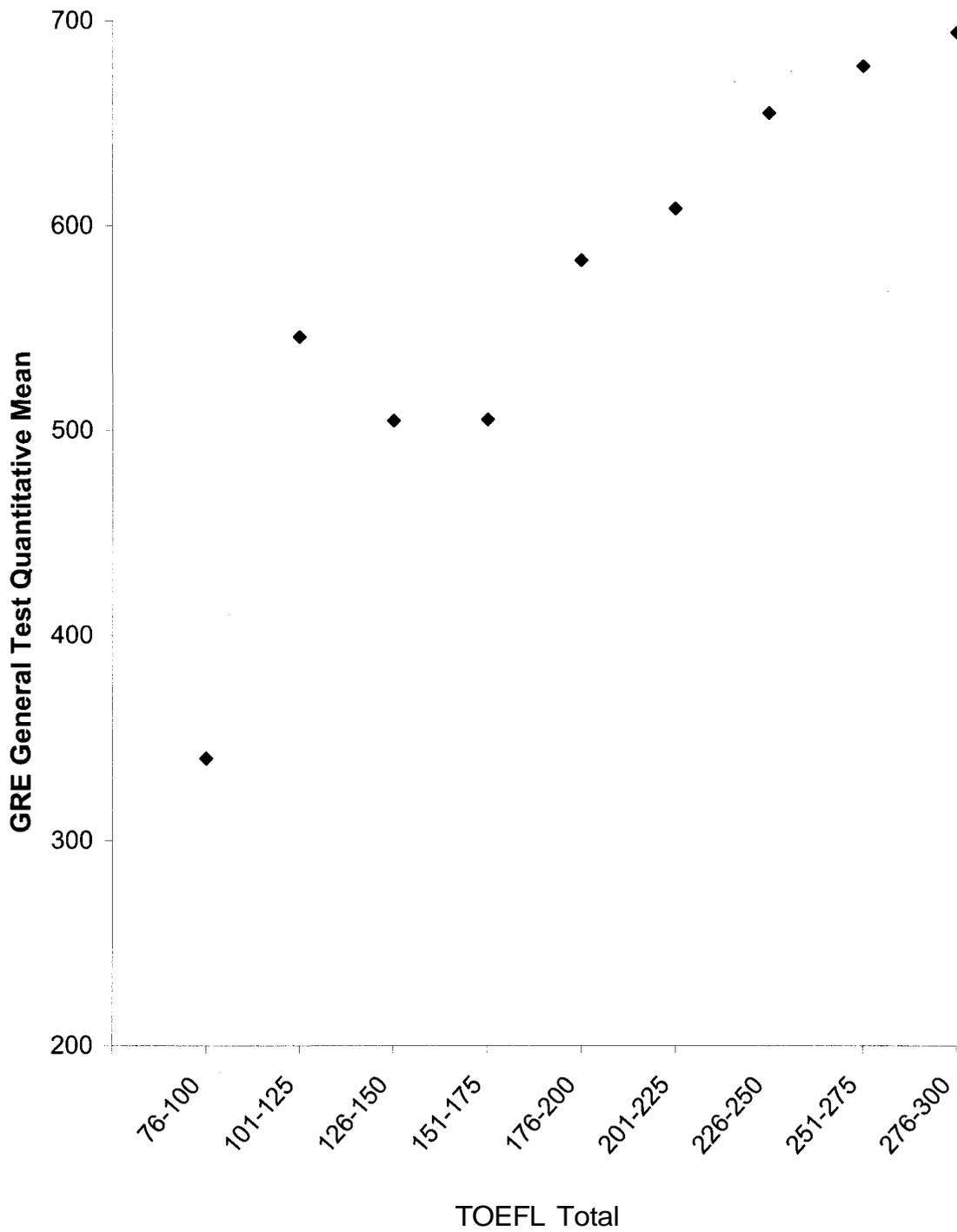
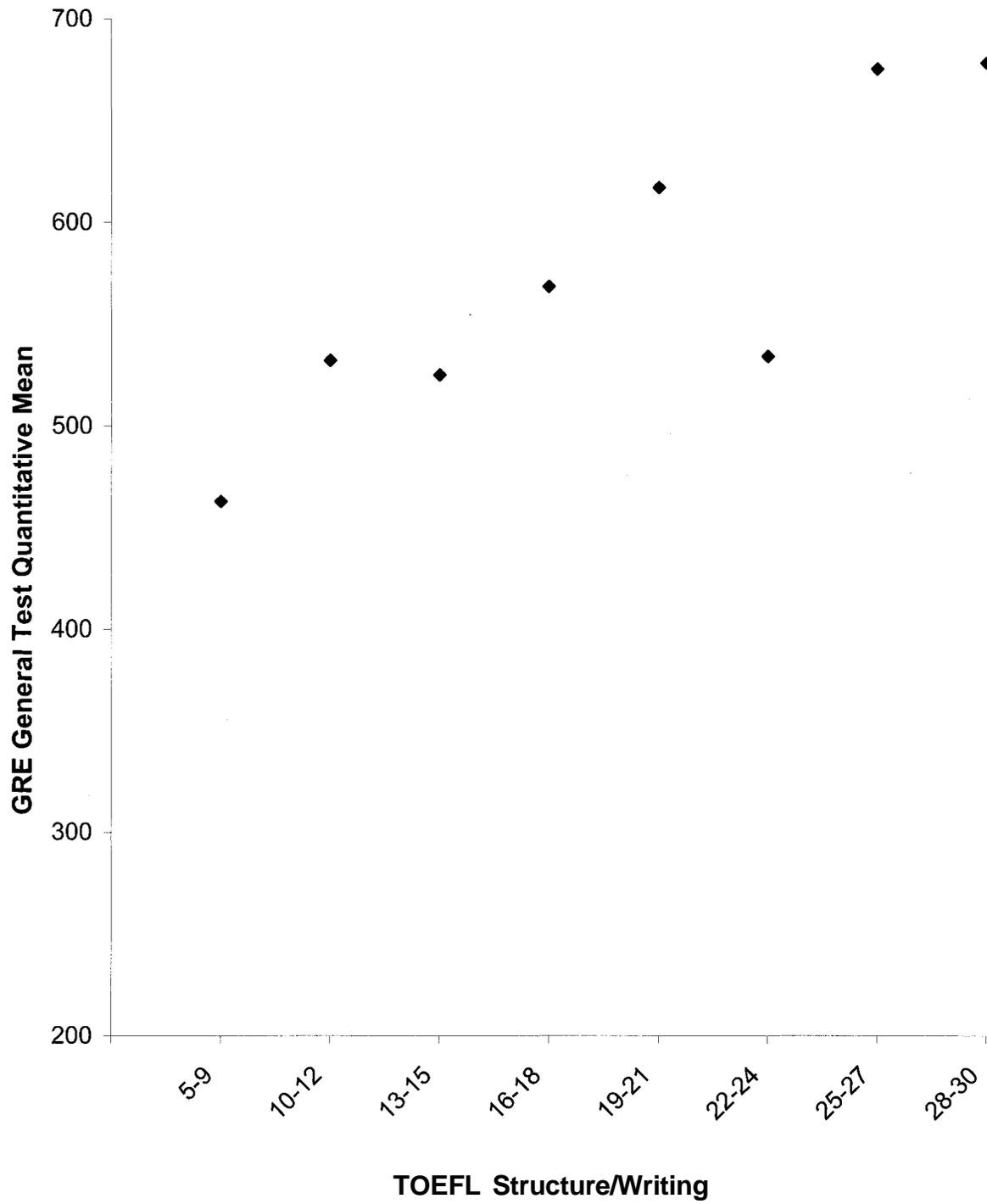*Figure 1.* Regression of GRE General Test quantitative scores on TOEFL total scores for ESL speakers.
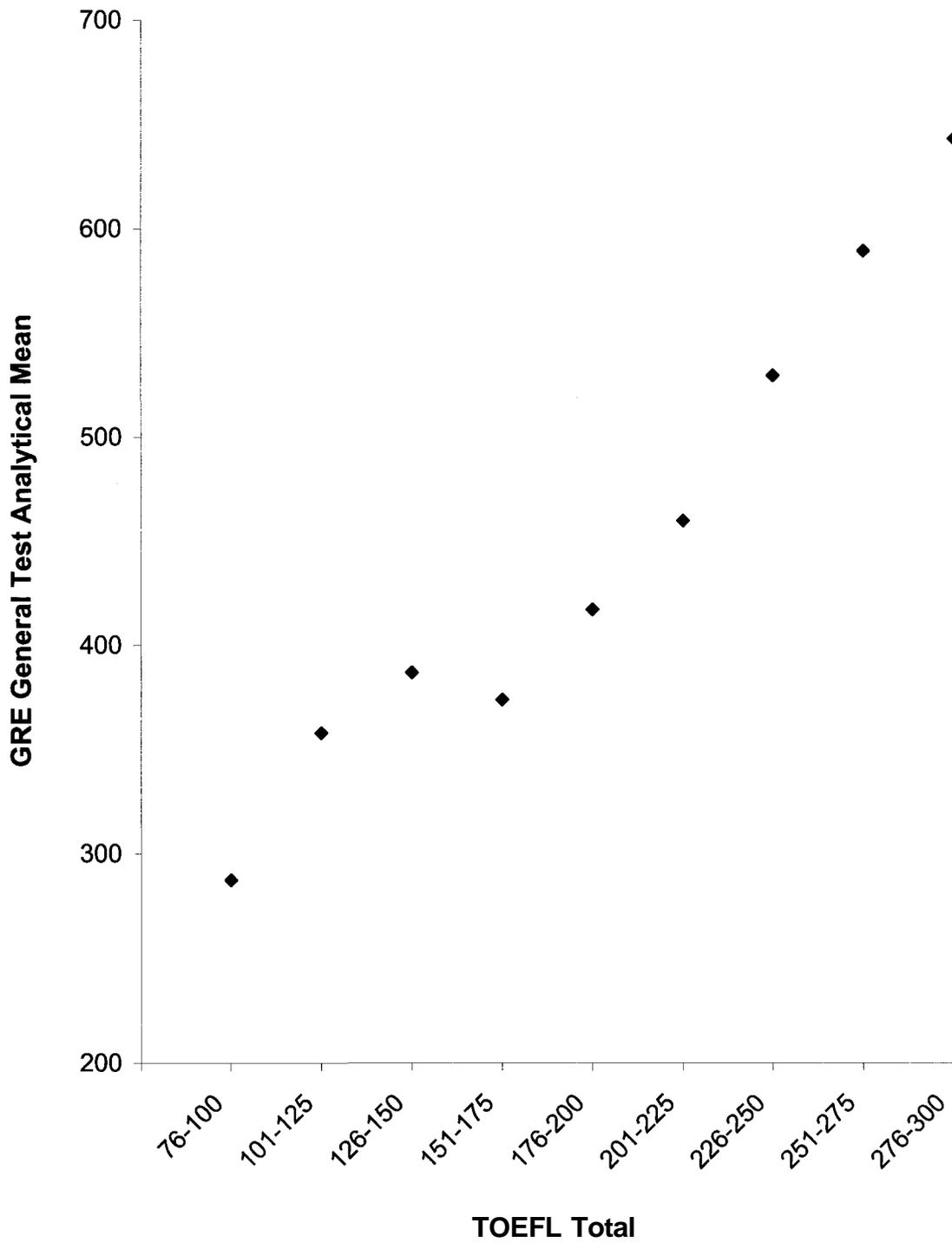
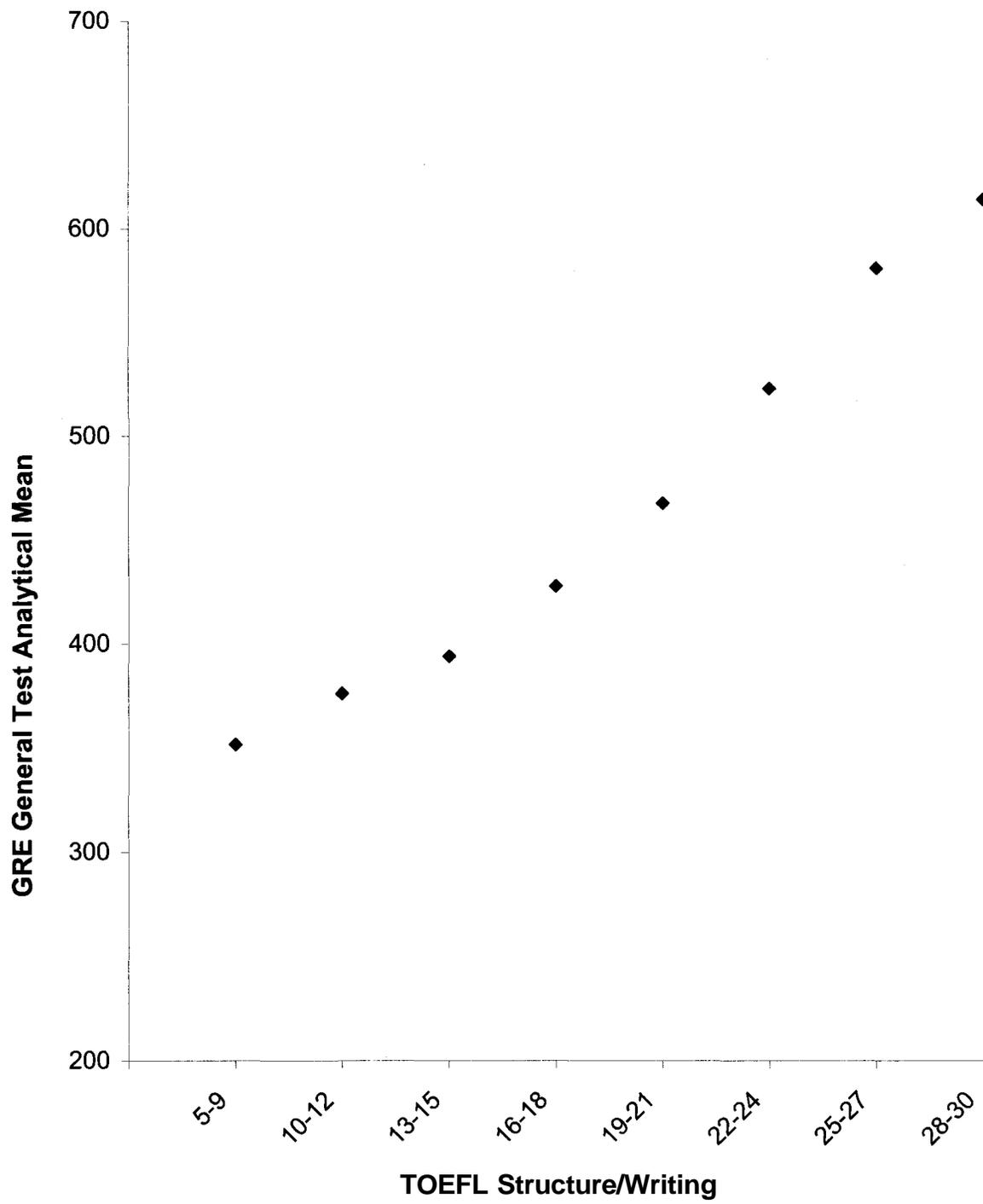*Figure 2.* Regression of GRE General Test analytical scores on TOEFL total scores for ESL speakers.
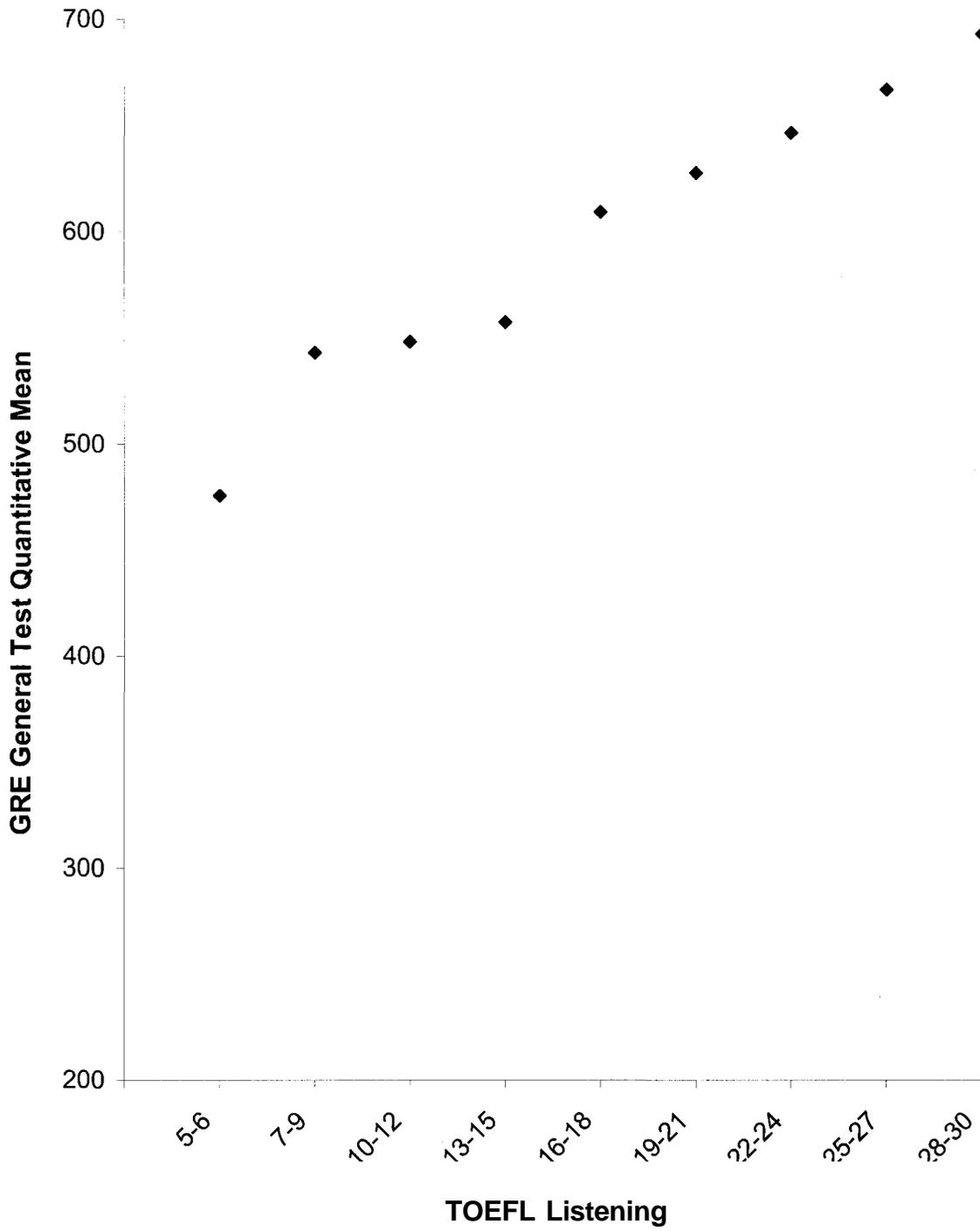
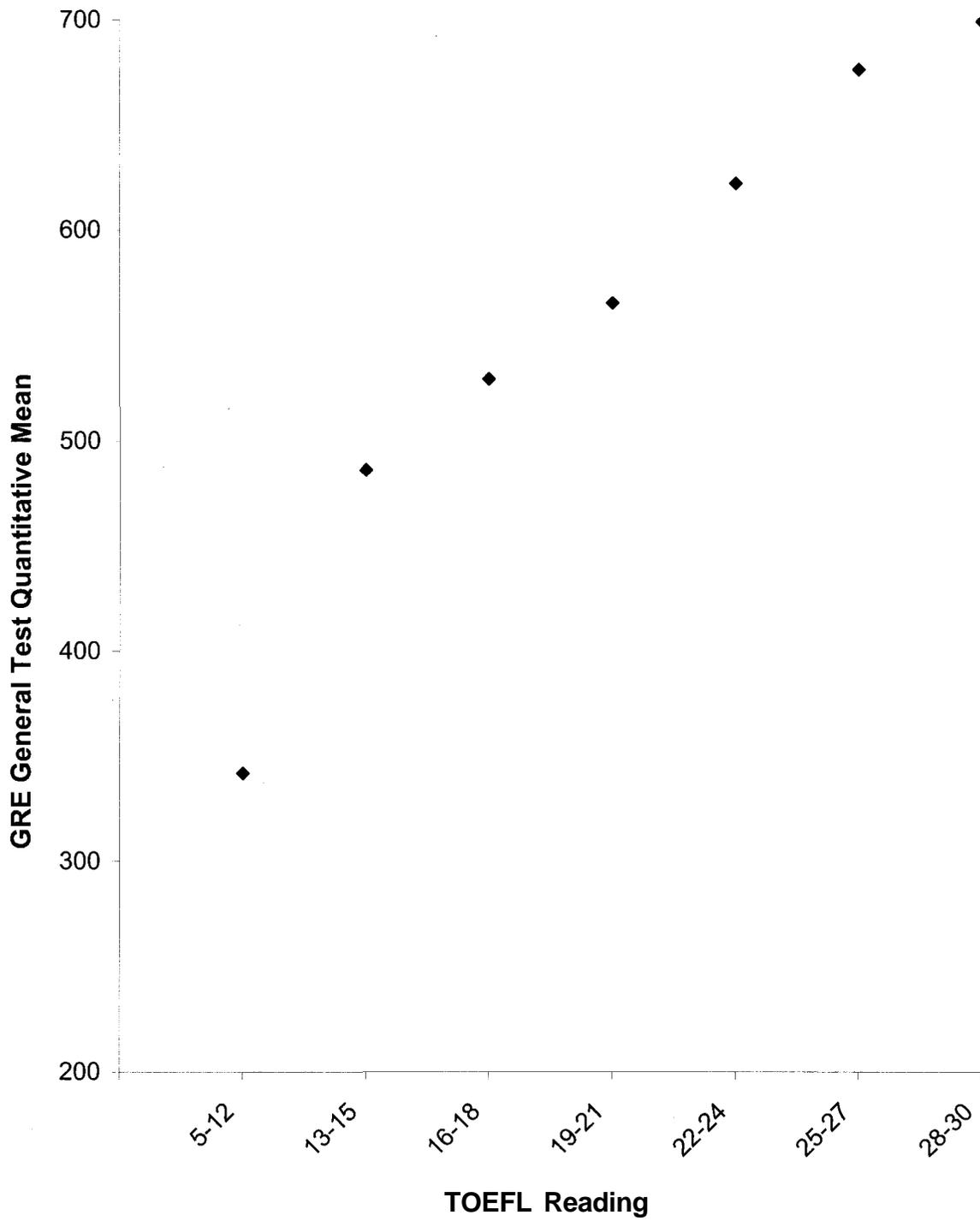*Figure 3.*   Regression of GRE General Test quantitative scores on TOEFL Listening scores for ESL speakers.
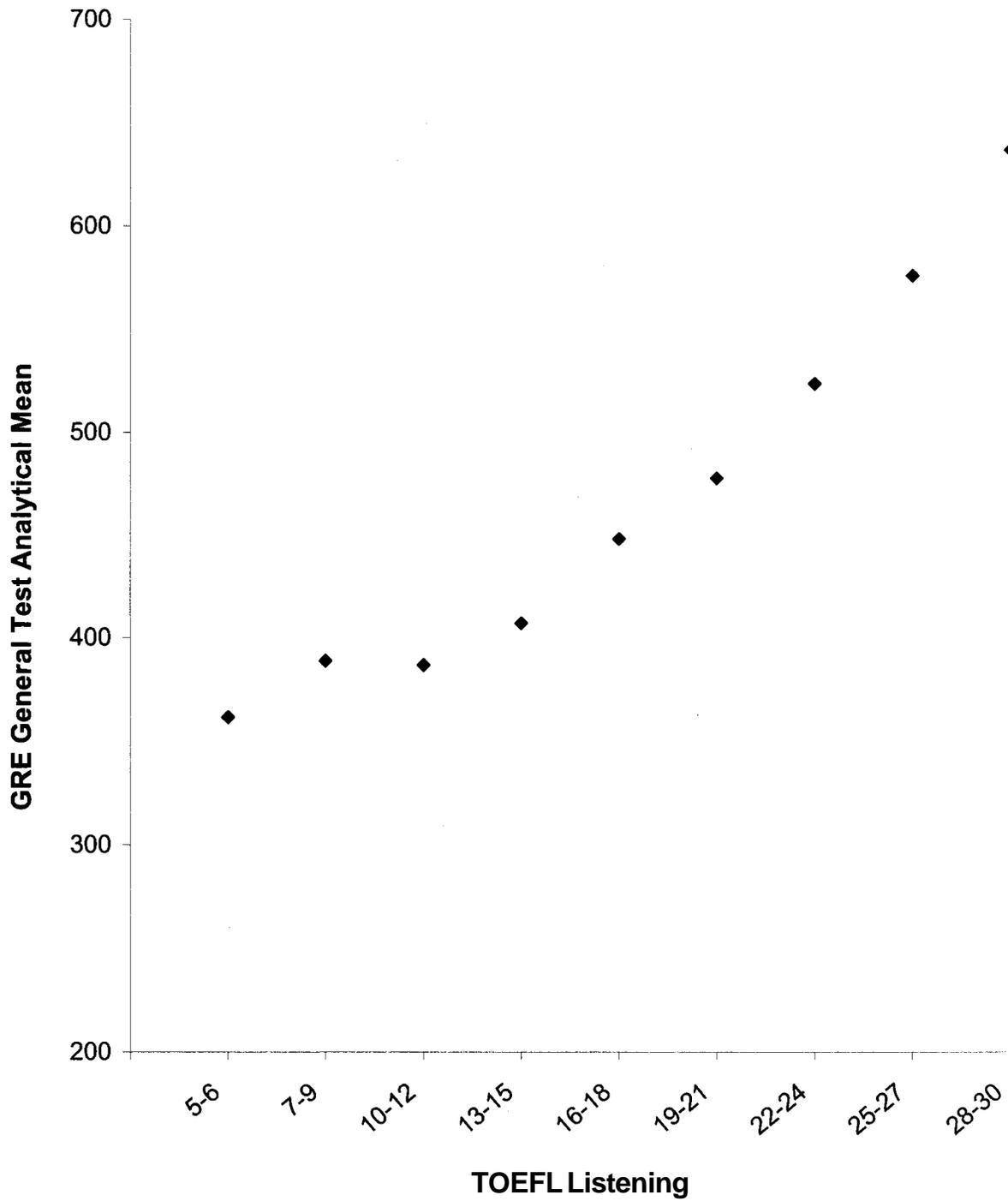
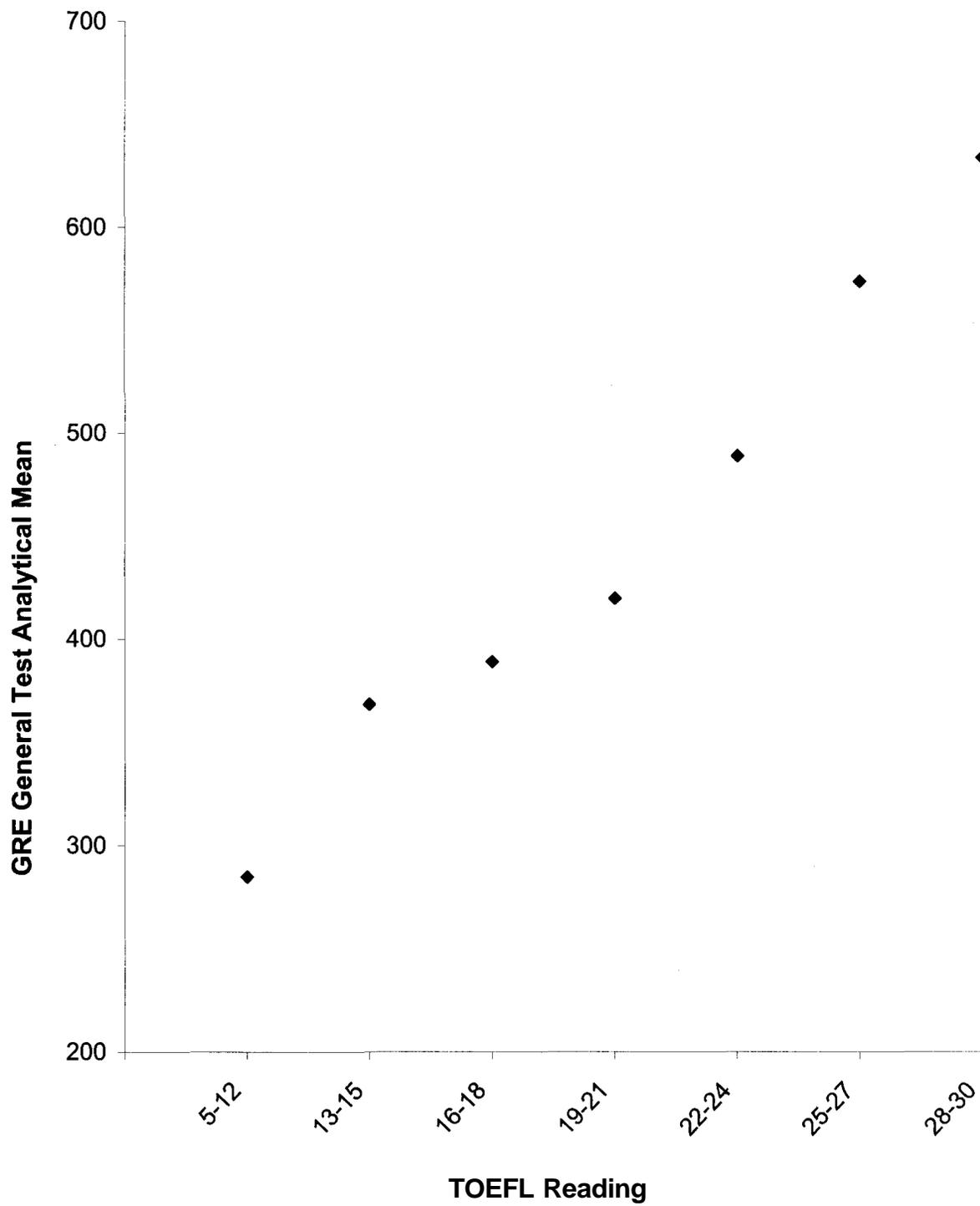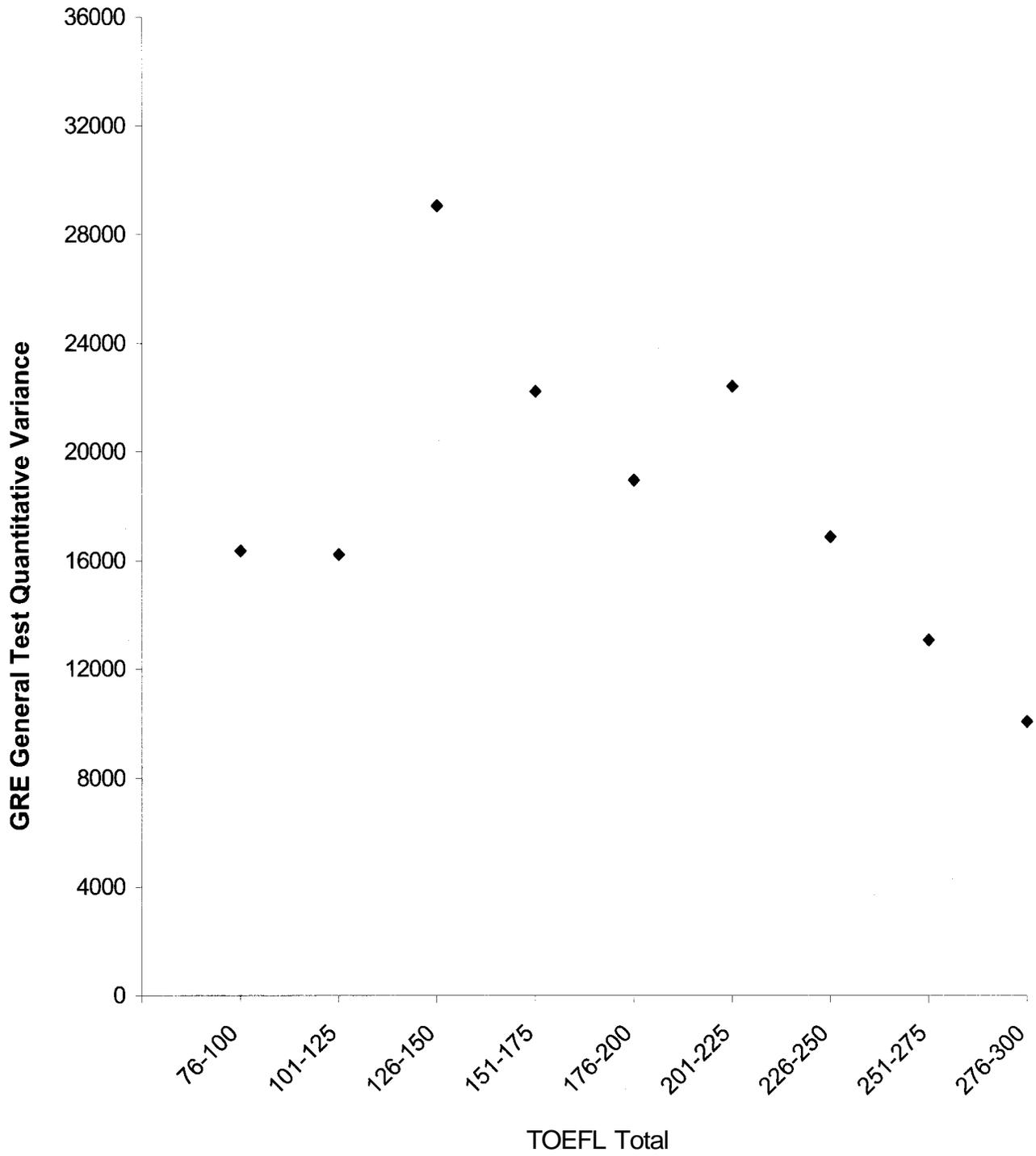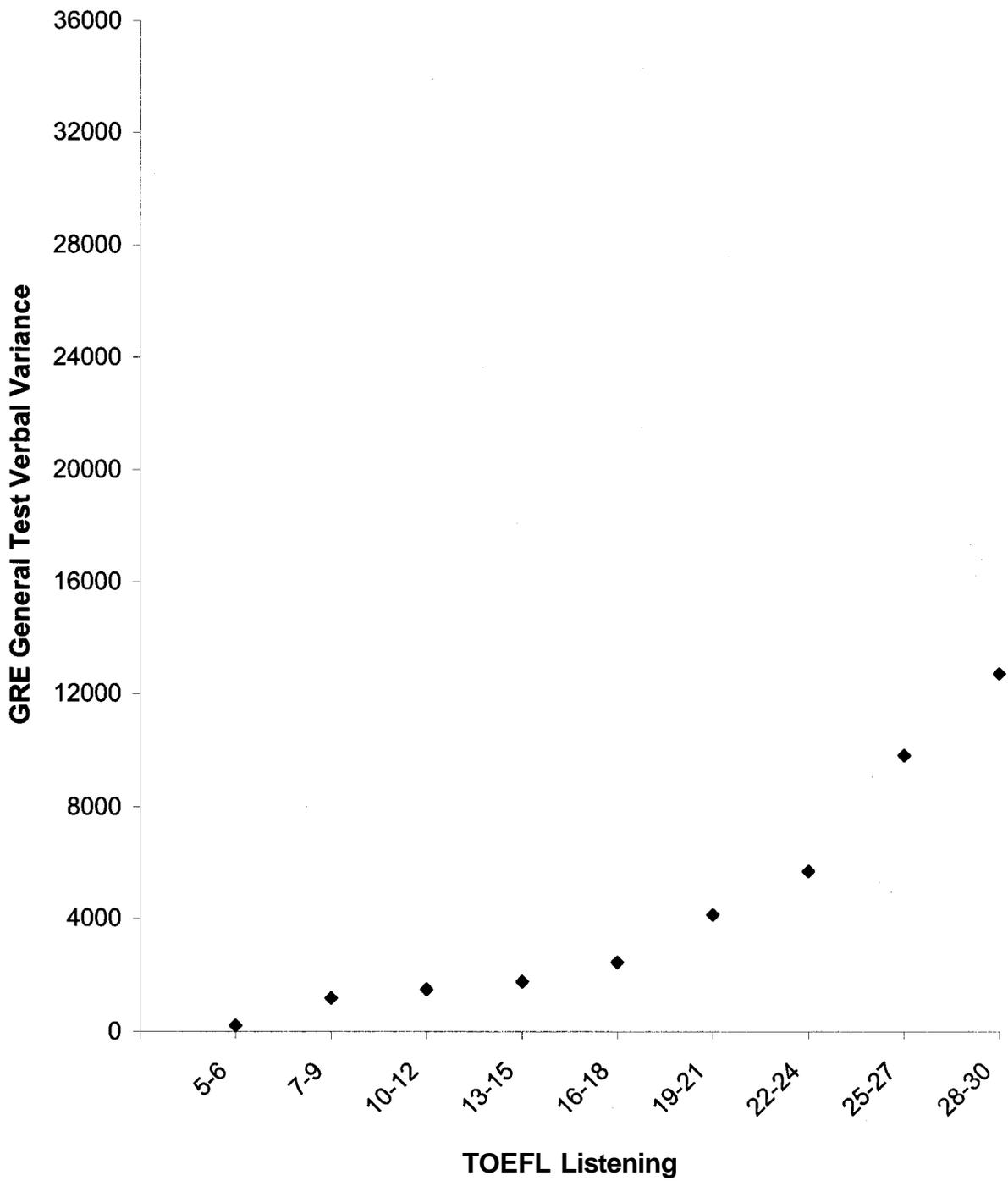*Figure 4.* Regression of GRE General Test analytical scores on TOEFL Listening scores for ESL speakers.

Figure 5. Regression of GRE General Test quantitative scores on TOEFL Structure/Writing scores for ESL speakers.

Figure 6. Regression of GRE General Test analytical scores on TOEFL Structure/Writing scores for ESL speakers.

Figure 7. Regression of GRE General Test quantitative scores on TOEFL Reading scores for ESL speakers.

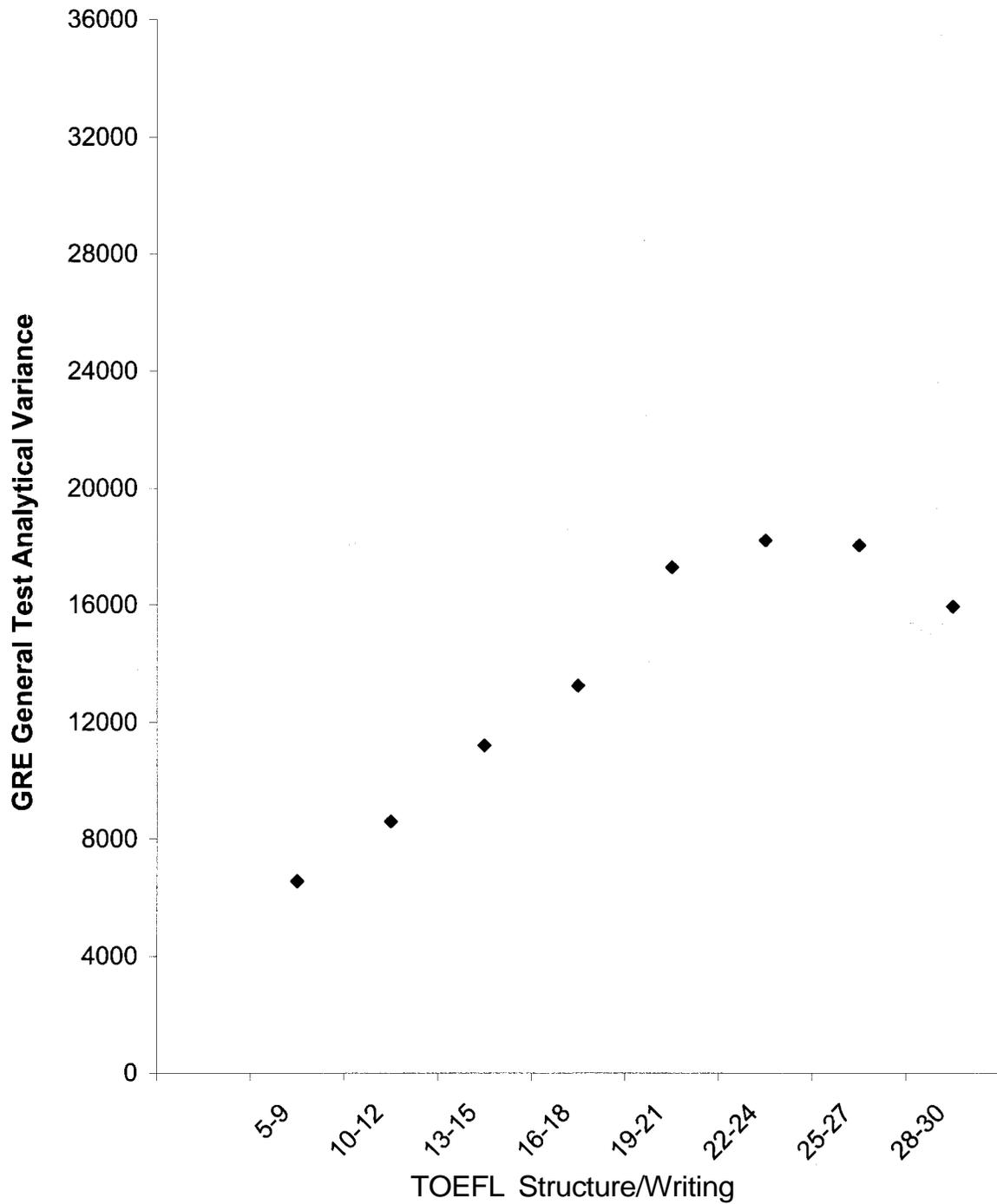*Figure 8.* Regression of GRE General Test analytical scores on TOEFL Reading scores for ESL speakers.

Figure 9. Trends in GRE General Test quantitative score variances associated with TOEFL total scores for ESL speakers.

Figure 10. Trends in GRE General Test verbal score variances associated with TOEFL Listening scores for ESL speakers.

Figure 11. Trends in GRE General Test quantitative score variances associated with TOEFL Listening scores for ESL speakers.
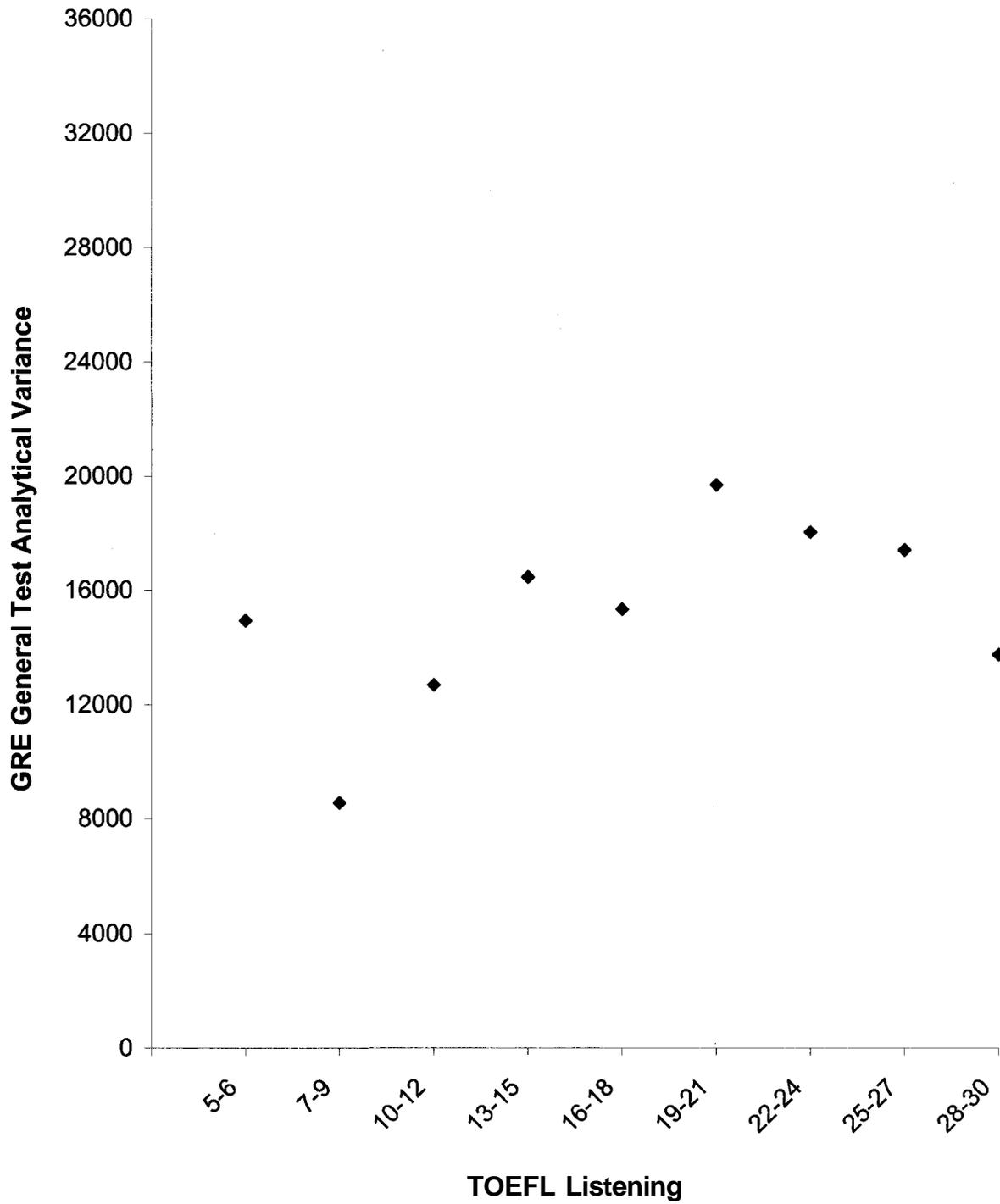
Figure 12. Trends in GRE General Test analytical score variances associated with TOEFL Listening scores for ESL speakers.
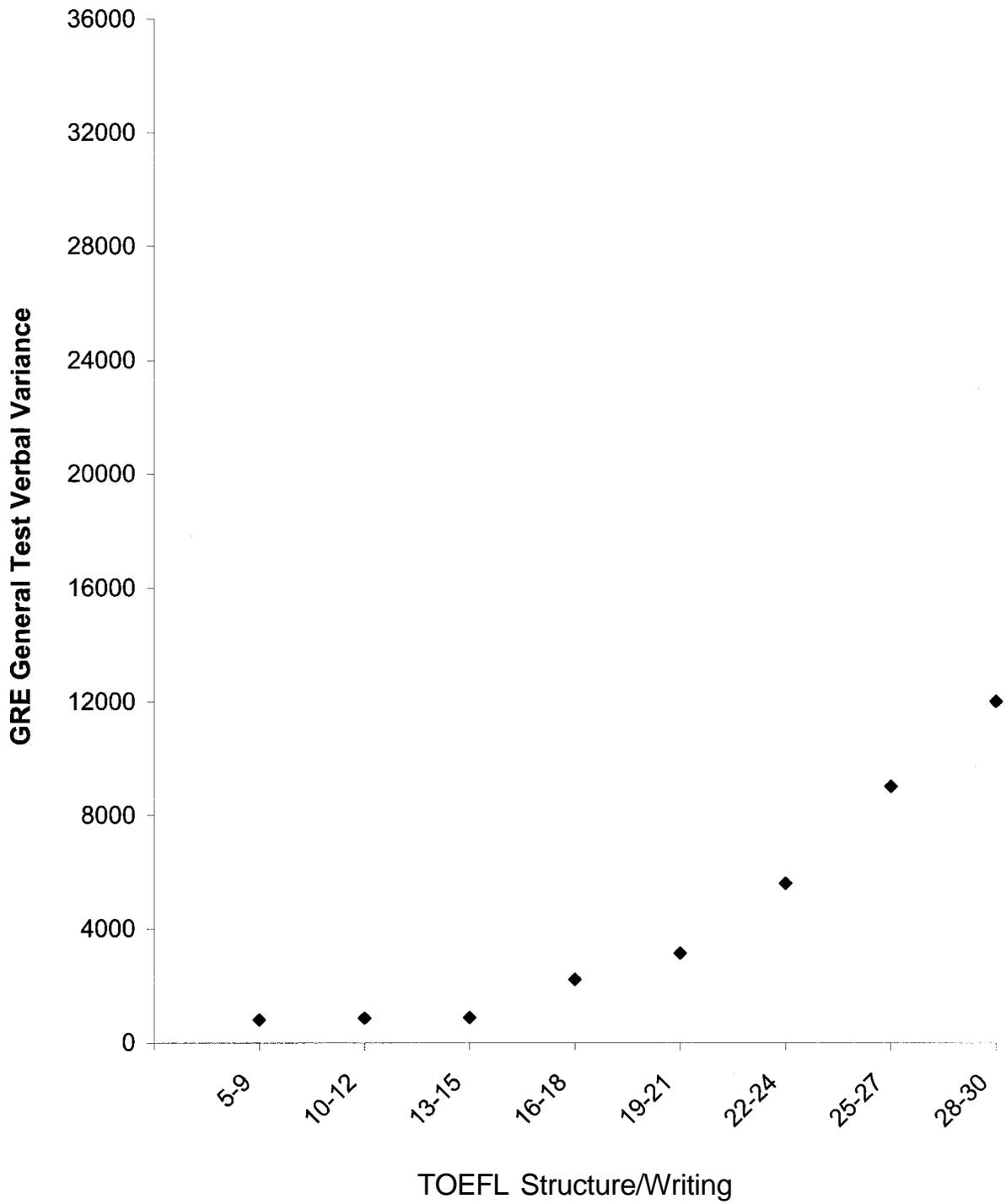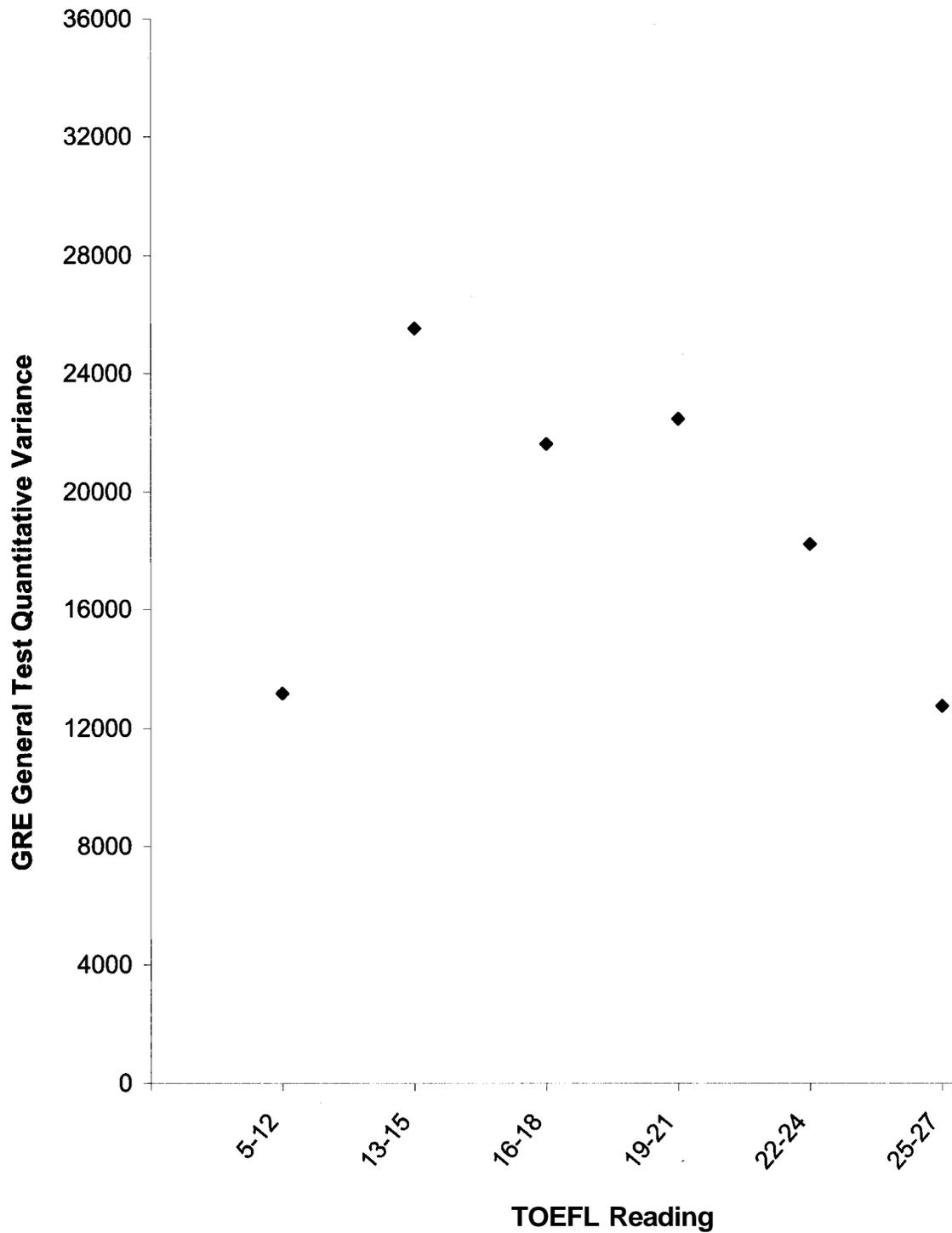
Figure 13. Trends in GRE General Test verbal score variances associated with TOEFL Structure/Writing scores for ESL speakers.
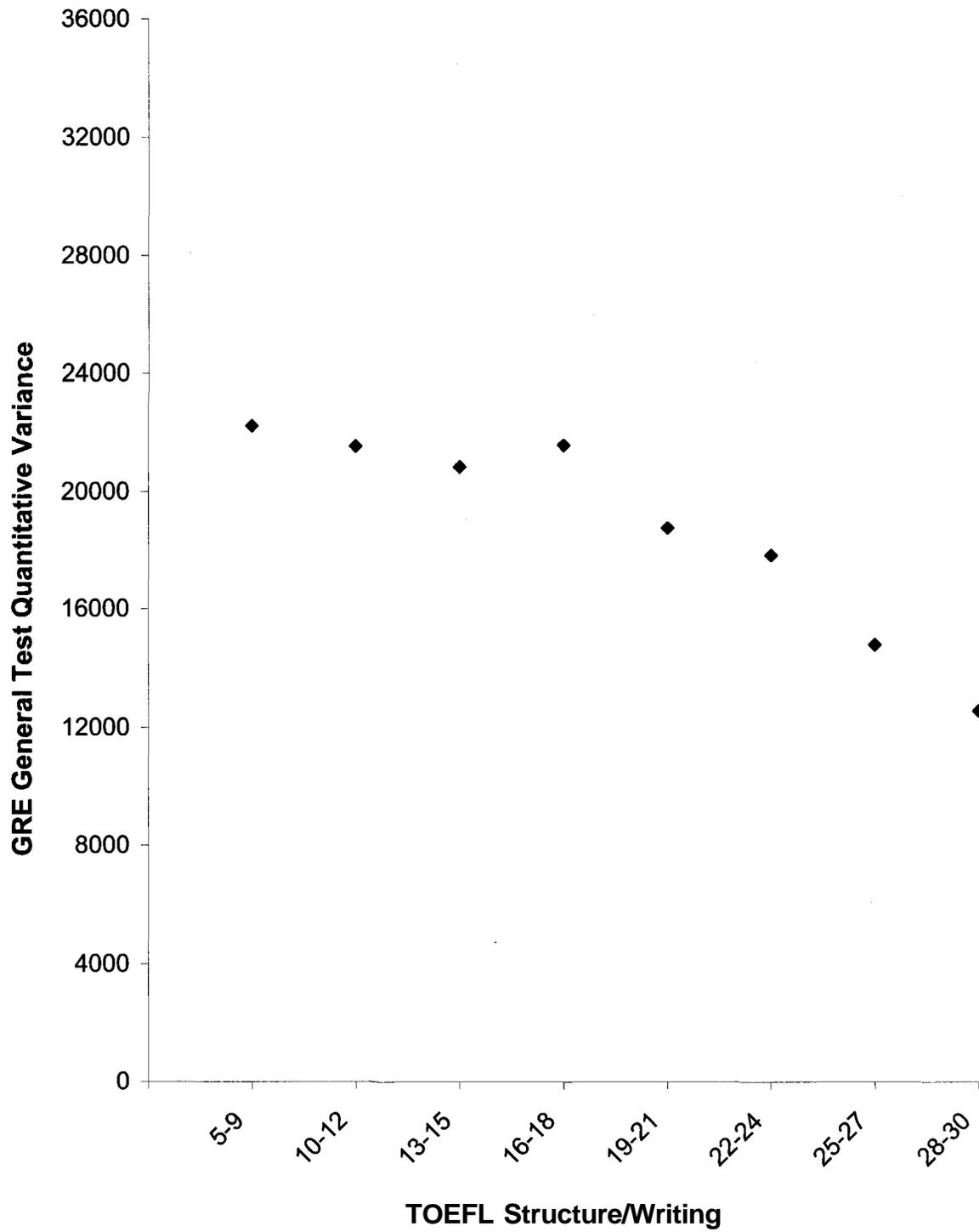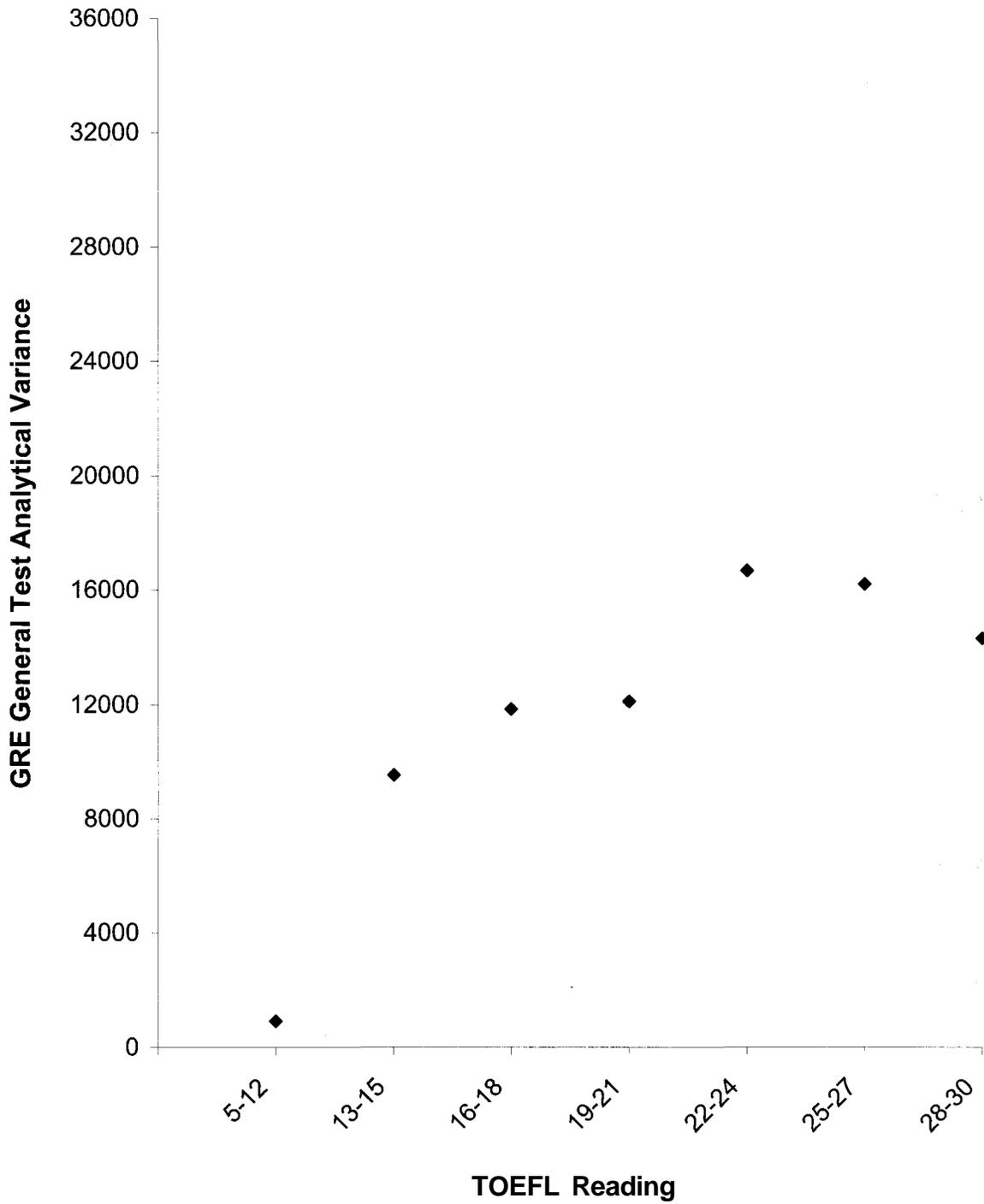
Figure 14. Trends in GRE General Test quantitative score variances associated with TOEFL Structure/Writing scores for ESL speakers.

Figure 15. **Trends in GRE General Test analytical score variances associated with TOEFL Structure/Writing scores for ESL speakers.**

Figure 16. Trends in GRE General Test quantitative score variances associated TOEFL Reading scores for ESL speakers.

Figure 17. Trends in GRE General Test analytical score variances associated with TOEFL Reading scores for ESL speakers.