DOCUMENT RESUME

ED 393 941 TM 024 993

AUTHOR Angoff, William H.
TITLE Context Bias in the Test of English as a Foreign Language.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-89-10; TOEFL-RR-29
PUB DATE Jan 89
NOTE 78p.; Table 10 contains very small print.
PUB TYPE Reports - Research/Technical (143) -- Statistical Data (110)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Context Effect; *English (Second Language); Evaluators; *Foreign Nationals; Interrater Reliability; *Item Bias; *North American Culture; Residential Patterns; Test Construction; Test Items; Test Reliability; Test Validity
IDENTIFIERS *Americanisms; Item Bias Detection; Mantel Haenszel Procedure; *Test of English as a Foreign Language; United States

ABSTRACT

This study was undertaken to test the hypothesis that items of the Test of English as a Foreign Language (TOEFL) containing reference to American people, places, customs, etc., tend to favor examinees who have spent some time living in the United States. Two samples of examinees were drawn from the March 1987 TOEFL administration, one tested in the United States consisting of 5,799 persons living in the country more than 1 year, and the other consisting of individuals with less than 1 month residence in the United States tested in their native countries (n=21,652 worldwide). Mantel Haenszel analyses were carried out for each of the 146 operational items of the test. In a separate part of the study, five raters were asked to rate test items to determine explicit references to some aspect of America. Ratings were highly reliable, and all but 16 of the 146 items were unanimously judged to have or not to have some reference to Americana. Of the TOEFL items, however, only one gave a consistent advantage, found in every region studied, to persons with U.S. residence experience. No support was provided for the hypothesis that TOEFL items give an advantage to persons who have lived in the United States for some time. Four appendixes contain the distribution of candidates by region, instructions for a pilot study and the formal study, and a display of raters' assignments of Americana to each test item. (Contains 3 figures, 22 tables, and 11 references.) (Author/SLD)

TOEFL®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 2
JANUARY

Context Bias in the Test of English as a Foreign Language

William H. Angoff

ETS

EDUCATIONAL TESTING S

The Test of English as a Foreign Language (TOEFL) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of over thirty organizations, public and private, that were 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 and 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, schoool systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and 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 TOEFL is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English-as-a-second-language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. Local research may sometimes require access to TOEFL data. In such cases, the program may provide this data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1988-89) members of the TOEFL Research Committee include the following:

Patricia L. Carrell (Chair)	Southern Illinois University
Lily Wong Fillmore	University of California at Berkeley
Fred Genesee	McGill University
Russell G. Hamilton	Vanderbilt University
Frederick L. Jenks	Florida State University
Harold S. Madsen	Brigham Young University

Context Bias in the
Test of English as a Foreign Language

by

William H. Angoff

Educational Testing Service
Princeton, New Jersey 08541

RR-89-10

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

ACKNOWLEDGMENTS

The author wishes to express his deep appreciation to Jennifer Nelson for her expert work in assembling and analyzing the data for this study. He also owes a debt of gratitude to Paul W. Holland for his statistical advice in the design and analysis stages of the work, and wishes to thank the five ETS staff members who gave of their time and thought in rating the items of TOEFL for evidence of reference to Americana: Richard Adams, Edward Curley, Miriam Levin, Kathleen Rabiteau, and Michael Zieky. Finally, he wishes to express his appreciation to Mary Varone and June Quattromani for their conscientious and expert attention to the production of this manuscript.

This study was supported by funds provided by the TOEFL testing program.

ABSTRACT

This study was undertaken to test the hypothesis advanced by Traynor (1985) that items of the TOEFL test that contain references to American people, places, regions, customs, institutions, etc., tend to favor examinees who have spent some time living in the United States. Two samples of examinees were drawn from the March 1987 administration of the test--one, tested in the United States, consisting of individuals who had lived in this country for more than a year, the other, tested in their native countries, consisting of individuals who had spent less than a month in the United States. Mantel-Haenszel (1959) analyses were carried out for each of the 146 operational items of the test, using data from the two groups, separately by region of origin, and also combined across regions. In all analyses, the subjects in the two groups were matched on the total score on the section in which the item was contained.

In a separate part of the study, five raters were engaged and asked to rate the items of the test, judging whether they contained explicit reference to some aspect of Americana. These ratings were found to be highly reliable; all but 16 of the 146 were unanimously judged by the raters as having reference or <u>not</u> having reference to Americana.

Of all 146 items in TOEFL, only one gave a consistent advantage, found in every one of the five regions of the world studied here, to examineees tested domestically. It is noteworthy, however, that this item made no reference to Americana. Several other items also showed an advantage to domestic candidates, but, with two exceptions, the advantage was found in only one region of the world. The two exceptions showed a domestic advantage in only two regions. Of the five most "significantly" aberrant items, only one was judged (by four of the five raters) to be an "Americana" item.

Bivariate distributions of "Americana" scores on the items of TOEFL versus their Mantel-Haenszel indices were prepared to determine whether there was any relation between the two. None was found.

On the strength of the information provided in the tables of this report, and on the strength of the review of the items cited here, it may be concluded that there is no support for the hypothesis that TOEFL items that make reference to American people, places, institutions, customs, etc., tend to advantage TOEFL candidates who have lived in the United States for a year or more over those who have spent little (one month or less) time in the country.

INTRODUCTION

In an article published in 1985, Traynor offered several criticisms of the Test of English as a Foreign Language (TOEFL). Among them was the concern. stated here as a hypothesis, that the items in TOEFL that draw their context from Americana--e.g., American individuals, places, events, objects, institutions, customs--favor unfairly those examinees who have spent substantial amounts of time in the United States, enough time to familiarize themselves with various aspects of the American scene. Correspondingly, by hypothesis, the test disadvantages those examinees who reside and are tested in their native countries and have spent little or no time in the United States, certainly not enough time to acquaint themselves with the various aspects of Americana. By implication, this hypothesis is assumed to have validity even though TOEFL is intended to measure generalized skills and knowledge of English proficiency and, at least by intent, does not depend on specialized subject area knowledge that the examinee may or may not bring to it. Although many of the items do speak of American persons, places, events, and so on, they are written to contain within themselves all the knowledge that the examinee will need to answer them correctly.

It should be mentioned, in connection with Traynor's assertion, that generally similar claims have been made about other tests in other contexts. For example, it has been asserted that reading comprehension paragraphs dealing with Black history or literature would be relatively easier for Blacks to read and understand than items of the same difficulty for Whites, but cast in the context of White history and literature. Similarly, it has been asserted, and with some evidential support, that paragraphs designed to test only reading skills, not content, but dealing with scientific matters, would favor science-oriented students over humanities-oriented students and, perhaps, men over women.

In general support of these hypotheses, recent developments in schema theory would suggest that the background knowledge examinees bring to a test may itself exert an effect on their test performance. Familiarity with the persons, places, institutions, and concepts mentioned in items may place them in a more easily understood context. The foregoing is a view argued convincingly by Melendez and Prichard (1985) in regard to comprehension of foreign language reading materials. At the least, one can theorize further that examinees who are confronted with a familiar item context will approach the item with the confidence that they can deal with it successfully; examinees who are unfamiliar with the context may, at least conceivably, become so disconcerted that they will fail to give full attention to the cognitive task and will answer the item incorrectly. There may also be other psychological and cognitive mechanisms that will account for the hypothesized effect.

There are other articles in addition to Traynor's that have dealt with issues generally related to this effect. Carrell (1984) speaks of the "interactive process between the reader's background knowledge and the text" as an integral part of reading comprehension. Thus, she says, "...[M]uch of the meaning understood from a text is really not actually *in* the text, per se, but in the reader, in the background or schematic knowledge of the reader." (Original emphasis.) Bernhardt (1984) draws the same conclusion: "Generic reading tests may be easily biased if they do not consider reader background knowledge and may, therefore, be incapable of accurately measuring reading ability....[R]eading tests which do not account for reader prior knowledge cannot achieve validity since they do not acknowledge a crucial factor in the comprehension construct." Carrell (1984) also points out, in more direct relevance to the hypothesis in question, "One of the most obvious reasons a particular schema [i.e., the reader's previously acquired knowledge structure] may fail to exist for an ESL reader is that the schema is specific to a given culture and is not part of a particular reader's background."

In this connection, Erickson and Molloy (1983) collected data on native and nonnative speakers of English, some of each group majoring in engineering, others not. They found, as expected, that the mean scores of the native speakers on listening comprehension and reading comprehension tests exceeded the mean scores of the nonnative speakers both on items with engineering content and on items with nonengineering content. They also found that the means for engineering students, both native and nonnative, exceeded the means for the nonengineering students, both native and nonnative, on items with engineering content. Clearly, then, both familiarity with the language and familiarity with the content had an effect on performance. Similar differences between engineering and nonengineering students, though not as striking, were found on general-language (nonengineering) items.

A study conducted by Alderson and Urquhart (1985) revealed the same effects as those found by Erickson and Molloy. Alderson and Urquhart concluded: "The hypothesis was supported that students from a particular discipline would perform better on tests based on texts taken from their own subject discipline than would students from other disciplines. That is, students appear to be advantaged by taking a test on a text in a familiar content area." This effect was further confirmed by Hale (1988), who found that students who specialized in humanities and social sciences outperformed students in the biological and physical sciences on text related to the former fields of study. As expected, the reverse was true for students of biological and physical sciences on text related to their own fields of study.

The issue addressed in the present research study, while certainly related to the work of the investigators cited above, is a little more specific. It asks whether the *context*--not the targeted content--of an item that is nontechnical and *for which all the information needed to answer the item correctly is contained within the item* tends to favor those individuals who are relatively better informed regarding the item's context.

A search of the ERIC data base was undertaken to discover whether there were other articles related to this topic, using the following descriptors in various combinations, as follows: test items, or item bias, or test bias, or cultural bias, combined with second language learning, or foreign language testing, or TOEFL, or English (second language). The results were then combined with the terms racial differences, or cultural differences, or culture-fair tests. These were then also combined with the term culture contact, denoting familiarity with another culture. This search revealed only one such study, by Schmeiser and Ferguson (1978), in which tests of English and social studies, containing items with Black and White content, were analyzed for bias. The results showed that there was no significant interaction between items of different content and the performance of Black and White examinees. That is, items with Black content were not easier for Blacks than for Whites, relative to items with neutral or White content.

Although the foregoing findings, if they may be generalized, are reassuring to those engaged in test development in the sense that the context of a test item has no significant effect on the measure of the skill (say, reading comprehension) to which the item is addressed, it would be useful to test the hypothesis further in the context of the performance of foreign students tested abroad versus foreign students tested domestically on TOEFL. It was to this end that the present study was designed. The particular intent of the study was to determine whether foreign students who have lived in the United States for some length of time, presumably long enough to become acquainted with information generally familiar to Americans, are more successful in comprehending the sense of TOEFL items than foreign students of the same level of relevant ability who have never lived in the United States. The content of the items could vary widely, including, say, an anecdote taken from the life of a famous American; the architecture of American office buildings; the foraging habits of certain American animals; the characteristics of an American lake, seashore, or mountain; voting patterns in the U.S. Senate; or the heights, weights, and longevity of American women. (The list could go on.) As indicated above, the question to be asked in the case of each item is: Do foreign students who have lived in America for some length of time have more success on these kinds of items than foreign students of the same relevant English language ability who come from the same regions of the world but have never lived in the United States.

DEFINITION OF STUDY SAMPLES

The procedure planned for this study called for the selection of a group of candidates tested domestically and a group tested overseas, and for a comparison of their responses to each of the items on a form of TOEFL. The specific procedure for developing the samples used in the study follows. By way of introduction, however, it should be mentioned that, as a matter of routine, the centers at which the candidates take TOEFL are clearly identified on their answer sheets. It was therefore a simple matter to divide the total candidate population at a particular administration into those who had taken the test domestically and those who had taken the test in foreign centers.

The candidates selected for study were those who took one of the 1987 operational forms of the test at a regular administration. For purposes of the study, the domestic candidates were restricted to those tested in the United States and Canada. However, in preparation for the study, the answer sheet administered to these students asked them to indicate how many months they had lived in the United States (allowing, for simplicity's sake, a maximum of "99 months"--i.e., 8 years, 3 months--"or more"). The restriction in the question to U.S. residency effected the exclusion, from the study sample, of those living in Canada. The distributions of the amount of time they had lived here are given in Table 1, separately for students coming from countries in Africa, the Americas (i.e., Central and South America), Asia, Europe, the Mideast, and the Pacific, and for all these areas combined. From these distributions samples were selected for study who had spent more than one year in the United States. Further, because of the small number (33) coming from the Pacific region, and the further reduction of that number when the sample was restricted to those living in the United States for more than one year (13), it was decided, because these data would be so unreliable, to drop the students from the Pacific area from the study entirely.

The answer sheet administered to students tested in foreign countries asked whether or not the student had spent more than one month in the United States. The purpose of this question, carried out once the data were tabulated, was to remove from the study sample all those who had indeed spent more than one month here, during which time they, like the domestic sample, might also have become acquainted with aspects of Americana. The distributions of these responses are given in Table 2, separately by region and combined.

The principal observations to be made in Tables 1 and 2 are that the candidates tested in foreign countries outnumber those tested domestically by a factor of almost two to one, and that of all the regions Asia contributes by far the largest number, indeed the majority of TOEFL candidates. Most of the Asian candidates come from Hong Kong, Japan, Korea, and Taiwan. For purposes of reference, Appendix A presents a breakdown of all the candidates tested from July 1984 though June 1986 by region of the world and by country within region. The distributions of the amount of time spent by the domestic candidates living in the United States (Table 1) is highly skewed--note that the means in each region are considerably higher than the medians--and that in general half of these candidates had spent less than 9 months living in the United States. Note also that, because the candidates who had spent more than 99 months in the United States are grouped in the same interval with those who had spent only 99 months in the United States, the means shown at the foot of the distributions are lower than they would have been had the intervals been extended to represent exactly how long the candidates had actually lived in the United States. Table 2 shows that the overwhelming majority (over 80%) of the candidates tested in foreign countries had spent no more than a month in the United States. Understandably, there is variation from region to region in this regard, as there is for the domestic candidates shown in Table 1.

Table 3 gives the distributions, by region, of the domestic and foreign students who were selected for this study. The domestic sample, it is recalled, were those who had spent more than one year in the United States; the foreign students were those who had spent no more than one month in the United States. These definitions were imposed on the data in order to magnify the difference in exposure of the study samples to U.S. culture without reducing their sizes beyond what was deemed necessary to conduct reliable studies.

It will be noted in Table 3 that the foreign candidates selected for study outnumber the domestic candidates by a ratio of 3.7 to 1, with the Asians representing (again) the large majority in both samples. This factor, the ratio of foreign to domestic candidates, varies considerably from one region to another, from 10.2 for Europe to 1.2 for the Americas. As just indicated, the largest number of candidates in the combined samples comes from Asia (N=19,396), and represents about 70% of the entire study sample of 27,451. The smallest, 1,109 (4% of the total) comes from Africa. It should be emphasized that these are the numbers of cases used in the present study of context bias; because the examinees were selected on the basis of the length of time they had spent in the United States. it is not to be inferred that they represent the proportions of TOEFL candidates in any year who come from the different regions of the world. Those proportions may be determined by reviewing the numbers shown in Appendix A. At the same time, it should be noted that the proportions of candidates in the study coming from each region are not strikingly different from the corresponding proportions in the candidate group before selection for the study.

It should also be noted in passing that the total number of cases actually used in the study samples (shown in Table 3) is 62 short of the total number satisfying the selection criteria. Of these, 51 came from the foreign samples, 11 from the domestic samples. Each of these cases was dropped because of some internal inconsistency that appeared in its identification.

Tables 4-9 present the distributions of test scores for the selected samples. Tables 4 and 5 present distributions of raw scores (rights only; all sections of TOEFL are scored rights only) on Section 1 (Listening

Comprehension) of TOEFL for candidates from each of the five major regions of the world considered in this study: Africa, Central and South America, Asia, Europe, and the Mideast. The first of these tables, Table 4, gives these distributions for the candidates tested in the United States (domestic samples); the second, Table 5, gives corresponding distributions for the candidates tested in their native countries (foreign samples).

Several characteristics of these groups are immediately evident in both Table 4 and Table 5. The highest-performing group in Listening Comprehension, shown in both tables, are the European candidates, who score more than half a standard deviation above the next-highest-performing group, those from the Americas. Quite possibly, one of the reasons for the superiority in the performance of these groups is the similarity to English, however slight, of the European languages spoken by most of these candidates, particularly as seen in the common Latin- or Germanic-derived cognates in English and several of the European languages. Possibly also, the greater amount and quality of English instruction in the European countries may contribute to this superiority. The lowest-performing groups are the foreign test candidates from the Mideast and Asia (Table 5), in these instances, perhaps, because of the absence of these advantages. It should be noted, however, that, leaving aside the high European means, there is remarkably little variation from region to region in their mean scores on Listening Comprehension.

The Europeans are also the highest-scoring on Section II of TOEFL (Structure and Written Expression), seen in both Table 6 (domestic) and Table 7 (foreign), although the patterns of mean score levels among the other regional groups is slightly different in this section from those in the Listening Comprehension section. Here, the Africans are the next highest-performing, possibly because of the formal emphasis given in the curriculums in Africa to the structure of the English language. As before, the candidates from Asia and the Middle East are the lowest-scoring, perhaps due in part to the dramatic differences in the orthography of these languages and the orthography of English. But again, the variation in means across the regions is quite small.

The distributions for Section III (Vocabulary and Reading Comprehension) in Tables 8 (domestic) and 9 (foreign) show the same superiority in performance for the Europeans. The students from Central and South America are next highest in performance, quite possibly, again, because of greater cognate similarities of the Spanish and Portuguese languages with English than is the case with the African, Asian, and Mideast languages. In general, these tables show that the candidates from regions with non-Roman orthographics rank last in all but one of these six tables. Finally, it is noted that the variation in means across the regions is substantially greater in Section III than we have observed in Sections I and II.

It is interesting to note in Tables 4 and 5 that the mean of each of the domestic groups exceeds that of its corresponding foreign group, a superiority that (speculatively) may be related to their greater exposure to, and possibly greater motivation to learn, spoken English as a consequence of living in this country. It is not, incidentally, to be regarded as supportive of the Traynor hypothesis; the latter is concerned not with the general superiority of one group over the other, but with the possibility of special advantage enjoyed by the domestic group on items that have a specific reference to Americana. In any case, the superiority of the domestic group shown on Section I is not at all evident in Sections II (see Tables 6 and 7) and III (see Tables 8 and 9), which are based on written, rather than spoken, language. There, the foreign groups are more often higher-scoring than the domestic groups, but, it is noted, not by very much. The combined foreign-group mean on Section II exceeds that of the combined domestic-group mean by less than one-tenth of a standard deviation. On Section III the superiority in their performance is less than one-sixth of a standard deviation.

Although some speculations are offered here for the rank orders of these regional groups--even though, as noted above, the variations among them are slight, particularly in Sections I and II of the test--it should also be noted that the speculations are, in fact, nothing more than that. Reasons for these differences that have been offered include variation in cognate similarities, orthographic similarities, and possible variation in curriculum similarities. Other reasons may include factors related to the self-selection of students to take the tests, similarities between English and several foreign languages with respect to written and spoken language structures, educational policies and practices in different countries (especially in regard to requirements for the inclusion of English in the curriculum), and frequency of contact with English-language materials and native speakers.

Finally, it will be of some interest to compare the performance of the candidates in the study samples with a combined reference group of TOEFL candidates taken from several administrations of the test. The table below provides this comparison. It gives scaled score means on each of the three sections of TOEFL for the domestic and foreign study samples, separately and combined, and for the combined group of 714,731 TOEFL candidates, tested in the period July 1984 to June 1986, some of whom took the test in domestic, and others, in foreign centers.

Section	Scaled Score Means for the Examinees in the Study			Scaled Score Means for the Reference Group[1] Tested 7/84 to 6/86
	Domestic	Foreign	Total	
I - List. Comp.	54.0	50.4	51.2	51.2
II - Struct. and Written Exp.	50.4	51.1	50.9	51.3
III - Vocab. and Read. Comp.	49.8	50.9	50.7	51.1
Nos. of Cases	5,799	21,652	27,451	714,731

[1] Taken from: Educational Testing Service, 1987; p. 21.

The mean values in the table reveal that the combined study group performed at very nearly the same level on all three sections of TOEFL as did the reference group, differing from it on every section by less than one-half of a mean scaled score point. The differences between the domestic study sample and the reference group are greater, however, showing superior average performance for the domestic group on Section I (by 2.8 scaled score points), but inferior performance on Sections II and III (by 0.9 points and 1.3 points, respectively). In evaluating these comparisons it should be recalled that, in an effort to magnify the difference in exposure of the domestic and foreign study groups to the particulars of American culture, the domestic candidates chosen for the study were intentionally defined as those who had lived in the United States for more than 12 months. (The foreign sample consisted of those who had spent no more than a month in the United States.) As a consequence, as many as 58.6% of the domestic candidates were removed from the original group of those candidates, leaving in the study sample those who had a substantial familiarity with spoken English and with Americana generally. (By way of contrast, only 16.6% of the foreign candidates were removed from their original group.) It is therefore not surprising that the domestic group was so superior to the reference group on the items of Section I. Interestingly, that familiarity is not visible on Sections II and III, which call for responses to written material and are therefore not as dependent as the items in Section I on residency in the United States.

Tables 10 and 11 give the intercorrelations among the three sections of TOEFL for each of the six regional groups and for the total group as well as scaled-score means and standard deviations for these groups. As in the preceding six tables, the first of these two tables gives the data for the domestic samples, the second, for the foreign samples. In reviewing these tables, we find that the correlations reported in them are not much different from those found in other studies of TOEFL; they generally range from the mid-.60s to the high .70s, with higher correlations between Section II (Structure and Written Expression) and Section III (Vocabulary and Reading Comprehension) than between either of those and Section I (Listening Comprehension), suggesting that the skills called upon in Sections II and III are more similar to each other than between either of them and Section I. We see also that the correlations among the three sections are not noticeably higher for the combined groups than for the five constituent groups. This is at least partly so because the pattern of pairs of centroids (the bivariate means) for one group versus another do not fall on a straight line; in any bivariate comparison, some groups score higher on one dimension than another, relative to the other groups in the set of five. If the centroids did fall on a straight line, the variance of the combined group and, consequently, the correlations in the combined group, would be consistently higher than in the individual groups. We see, finally, that all the domestic groups show higher scaled score means on Section I than on Sections II and III, indicating a relative superiority in Section I over Sections II and III for the domestic study sample in comparison with the original scaling group, undoubtedly due to their greater exposure to spoken English than was true of the scaling group. This consistent superiority in Section I over Sections II and III is not at all evident in the foreign study sample, who had not had the benefit of that degree of exposure to spoken English.

PROCEDURE

The study was organized in three phases. In the first phase, ratings were made of each of the items under study as to the presence or absence of reference to some aspect of Americana. In the second phase, studies of differential item functioning (DIF) were undertaken to determine whether or not the two principal groups--those tested domestically and those tested in their native foreign countries--showed noticeably different success rates on the items, even after matching on English language ability, as measured by the total score on the particular section of the test containing the item. The third, evaluative, phase of the study was an attempt to rationalize the high values of DIF observed in the study and to compare the results of the first two phases, to determine whether or not the items with large DIF values were in fact the items that were rated as "Americana" items.

Rating of the Items

As indicated above, the purpose of the item rating exercise was to classify, or rate, each item of the test form used in the study as to the presence or absence in the item--either in the stem, in the options, or in the referent (e.g., the paragraph on which it was based)--of some aspect of Americana.

In a pilot study of the rating process, carried out in connection with the preparation of the proposal for the study, the raters were asked to go through the 150 items of an earlier form of the test and classify each item as (1) referring to some aspect of Americana--e.g., containing the name or names of an American person or geographical region, place, or natural phenomenon (river, mountain range, volcano, etc.), a characteristically American institution, concept, custom, etc; (2) not referring to someone or something just described; or (3) ambiguous in this regard. Further. the raters were asked to rate the items on a scale of 1 to 5 (5 = highly American; 1 = not American).

The outcome of the pilot study made it clear that the ratings could not be made on a continuum as, for example, from "weaker reference to Americana" to "stronger reference to Americana," but could only be placed in one or the other nominal category--reference to Americana or not--as would be the case, for example, in the nominal categories man/woman, Swedish/Hungarian, baseball/golf, teacher/lawyer. Further, it was evident that the instructions given in the pilot study were not sufficiently explicit. There were areas of ambiguity in the classification, Americana/non-Americana, that required more detailed instruction. As a result of the experience gained in the pilot study, it was decided in the present formal study to give the raters of the test form studied here more specific guidance for making judgments, especially in those instances that had previously been thought to be ambiguous. For reference and comparison, Appendix B shows the memorandum that was sent to the raters describing their task for the pilot study; Appendix C is the memorandum that was sent to the raters giving them their instructions for the current, formal, study. It should be noted that in the second memorandum the raters were asked to classify the items as (1) making reference to Americana or (2) not. No opportunity was afforded them for a rating of degree of Americana or of ambiguity.

In the pilot study six people who were regularly engaged at ETS as test development specialists working on verbal tests (including foreign language tests) were chosen as raters. Staff members whose daily work involved the development of TOEFL items and tests, however, were excluded from the rater group on the presumption that their earlier work in constructing this form (and other forms) of TOEFL was carried out with a particular definition of Americana as they saw it. It was judged that a separate group, not normally engaged in constructing TOEFL items, would engage themselves in the rating task with no predispositions for the definition of Americana and would therefore have less difficulty accommodating themselves to the definition as given to them in the rating instructions.

In the conduct of the formal study, the same raters who participated in the pilot study were asked to participate again. As it turned out, one of the six original raters found that she could not participate in the formal study, and so the number of raters was reduced from six to five.

Analysis of Differential Item Functioning

Quite independently of the rating procedure, a measure was obtained of the differential functioning of each item, i.e., the degree to which the item appeared to favor the domestic group over the foreign group. The question addressed in this analysis was whether the domestic and foreign groups enjoy the same degrees of success on the various items after controlling for differences in ability for these two groups, as measured by their scores on the section of TOEFL in which the item appears. The index of this measure, referred to here as D-DIF (differential item functioning, expressed approximately on the ETS delta scale of difficulty), was the Mantel-Haenszel (MH) index (Mantel & Haenszel, 1959; Holland & Thayer, 1988). This index may be described by considering the following 2x2 table that describes the frequencies of correct (1) and incorrect (0) responses to an item by the

		Item Performance		
		1	0	
Tested Group	D	a_i	b_i	$N_{Di} = a_i + b_i$
	F	c_i	d_i	$N_{Fi} = c_i + d_i$
		$N_{1i} = a_i+c_i$	$N_{0i} = b_i+d_i$	$N_i = a_i+b_i+c_i+d_i$

domestically tested (D) and foreign tested (F) candidates, all of whom have fallen in the same interval (i) of scores on the relevant section of TOEFL (and are in that sense matched on ability), and come from the same region of the world, as defined in the TOEFL Test and Score Manual, 1987-88 edition (Educational Testing Service, 1987). The Mantel-Haenszel index (α) may be described as calculated at one particular score interval (i) of the section of TOEFL of which the item is a part:

$$\alpha_i = \frac{p_{Di}}{q_{Di}} \Big/ \frac{p_{Fi}}{q_{Fi}} = \frac{\dfrac{a_i}{a_i + b_i}}{\dfrac{b_i}{a_i + b_i}} \quad \frac{\dfrac{c_i}{c_i + d_i}}{\dfrac{d_i}{c_i + d_i}} = \frac{a_i}{b_i} \Big/ \frac{c_i}{d_i} = \frac{a_i d_i}{b_i c_i}, \qquad (1)$$

where p_{Di} is the proportion of the domestic candidates in score interval[2] i who answered the item correctly, and $q_{Di} = 1 - p_{Di}$. Similarly, p_{Fi} is the proportion of the foreign candidates who answered the item correctly, and $q_{Fi} = 1 - p_{Fi}$. Thus, α_i is the ratio of the odds (p/q) that the domestic candidates in a particular region have answered the item correctly, divided by the odds that the foreign candidates in the same region have answered the item correctly. If there is no difference in the performance of the two groups on this item within this score interval, then α_i will be equal to 1. If, however, the two groups function differently--if, for example, in this score interval the domestic group from a particular region performs better on the item than the foreign group from that region, $\alpha_i > 1$. If, on the other

[2] The score intervals were defined as follows: 0-2, 3-4, 5-6, ..., 49-50, on Section I; 0-2, 3-4, 5-6, ..., 37-38 on Section II; and 0-4, 5-7, 8-10, ..., 56-58 on Section III.

hand, the foreign group performs better than the domestic group, $\alpha_i < 1$.

The Mantel-Haenszel procedure estimates a common odds ratio across all matched categories. The form of its index is given as follows:

$$\hat{\alpha} = \frac{\sum_i p_{Di} q_{Fi} N_{Di} N_{Fi} / N_i}{\sum_i q_{Di} p_{Fi} N_{Di} N_{Fi} / N_i} = \frac{\sum_i a_i d_i / N_i}{\sum_i b_i c_i / N_i} , \qquad (2)$$

which is the average factor by which the odds that a member of the domestic group responds correctly to the item exceeds the odds that a member of the foreign group responds correctly to the item.

For the sake of convenience, $\hat{\alpha}$ is transformed to another scale, yielding an index, which we will refer to as MH D-DIF, by means of the equation, MH D-DIF = 2.35 $\ln(\hat{\alpha})$. This transformation centers the index about the value 0 (which corresponds to the absence of differential item functioning), and puts it on a scale roughly comparable to the ETS delta scale of item difficulty. Positive values of MH D-DIF indicate that the items favor the domestic group; negative values indicate that the items favor the foreign group. In the interpretation of all such indices it should be emphasized again that for each item studied the abilities of the groups were first matched on the basis of the total score on the section containing the item. Further, it should be noted that in the range of moderate item difficulty--between 30% and 70%--the absolute value of MH D-DIF is approximately equal to 10 times the absolute difference in proportion correct for the domestic group and the matched members of the foreign group. Thus, |MH D-DIF| = 1 corresponds to a difference between the matched groups of 10 percentage points; |MH D-DIF| = 2 corresponds to a difference between the matched groups of 20 percentage points.

Conceivably, a question of design might be raised in connection with the present study, inasmuch as the domestic sample, unlike the foreign sample, undoubtedly benefited from their experience living in the United States, and the two groups were matched on a variable that was differentially affected by their different experiences. It could be argued that the domestic group should have taken the test before leaving for the United States and matched with the foreign group on the basis of scores unaffected by their U.S. experience; and that they and the foreign group should have taken the test again for purposes of the DIF analysis after the period of differential experiences with Americana. The position taken in this study, however, is that the intent of the study was to identify items that showed differential effects after taking into account, by the process of matching, any differences that may have existed between the two groups, and stemming from whatever conditions that may have caused those differences to exist. It is this latter position that was taken in the design and conduct of the study.

Relationship Between Ratings of Americana and Mantel-Haenszel Results

The third phase of the study called for a comparison between the ratings of Americana given to the items and the independent measures of differential item functioning. For this comparison, distributions of the Mantel-Haenszel indices were prepared separately for each category of rating of Americana. The intent here was to search for a trend in the MH D-DIF values as a function of the ratings of Americana, to determine whether the items that were more clearly thought to have an American reference did indeed tend to have higher MH values.

RESULTS

Analysis of Ratings of Reference to Americana

The ratings given to the items--1 (having reference to Americana) or 0 (not having reference to Americana)--are shown in Appendix D separately by item and by rater. An examination of these ratings indicates quite clearly that, for the most part, the raters classified the items in the same way. In the last column (Sum) it is seen that most of the items, by far--a total of 130 out of 146 items--were rated unanimously; 105 were rated 0 (non-Americana) by all raters, and 25 were rated 1 (Americana) by all raters. Only 2 items out of the 50 in Section I (Listening Comprehension), 4 of the 38 items in Section II (Structure and Written Expression), and 10 of the 58 in Section III (Vocabulary and Reading Comprehension) failed to be rated unanimously; of those 16, 7 achieved near-unanimity (4 to 1, either way) and 9 showed a 3-to-2 split.

Table 12 shows the intercorrelations (phi-coefficients) among the five raters, separately for Sections I, II, and III of TOEFL, and for all sections combined. The sizes of these intercorrelations, ranging from .73 to 1.00, attest again to the high degree of agreement among the raters in the judgment of an item's reference to Americana. The reliabilities of the sum of the ratings given by the five raters were .972 for the items of Section I (Listening Comprehension), .970 for the items of Section II (Structure and Written Expression), .954 for the items of Section III (Vocabulary and Reading Comprehension), and .963 for all the items of TOEFL combined. Apparently, the instructions that the raters were asked to follow were sufficiently clear to ensure high levels of rater reliability.

Analysis of Differential Item Functioning by Region of Origin

The second phase of the analysis consisted of subjecting each of the 146 items of TOEFL to a statistical evaluation of the degree to which the item favored TOEFL candidates tested in the United States over TOEFL candidates tested in their native lands, after matching the individual regional groups for ability as measured by the total score in that section of the test in which the item appeared.

Tables 13-15 present distributions of the original ("unpurified"; see discussion of Tables 16-18 for explanation) Mantel-Haenszel indices of D—DIF, one for each of the sections of TOEFL, separately by region of the world and for all regions combined. Two distributions are given in each table for the total of all regions--one, in which the candidates were matched only on total score, and a second, in which the candidates were matched by total score and also by their region of origin.

Several general observations can be made in these tables. The first is, as expected, that the means of the distributions all hover about zero, a natural outcome of the matching itself, in which the aggregate advantage on some items of each section of the test enjoyed by the domestic group is balanced by the aggregate advantage on other items enjoyed by the foreign group, once they are matched on overall ability on the section.

Second, there are noticeable differences from one regional group to another with respect to the dispersions within the distributions. Of all the regions, Asia, for example, shows the smallest dispersion. To some extent these differences in dispersion are a function of the sizes of the groups; the groups with the largest numbers (e.g., Asia) would be expected to have the smallest dispersions because of the smaller component of random error in the DIF indices. But one may speculate that other factors also enter here: the degree to which the total score is an effective matching variable--i.e., the degree to which it correlates with item performance for candidates coming from each region; and, at least by hypothesis, the degree to which living in the United States does in fact exert a differential effect, positively or negatively, on item performance, after matching on total score.

A third observation made in these tables is that the dispersions in the total group distributions are generally smaller than in any of the individual regions, at least in part because of the greater size of the total group. Finally, it is seen that the distributions in which the matching was done by score _and_ region generally have smaller dispersions than those in which the matching was done by score alone. This is quite likely because region is a relevant matching variable, serving to reduce random variation beyond that afforded by the total score on the relevant section of the test.

If one searches within Tables 13—15 for evidence of discrepant, or aberrant, items, one finds in Table 13 (Section I of TOEFL) that there is indeed at least one item in all but two of the regional distributions (the Latin Americas and the European regions), and also in the total—group distributions, whose MH D-DIF value lies at some distance from the rest of the MH values in the distributions. This is item 13 in the Listening Comprehension section. It will be discussed, along with other discrepant items, in connection with Table 19.

Tables 16-18 parallel Tables 13–15, differing from them in only one respect. This is that items with relatively large positive MH D–DIF values, values of 2.00 or greater, have been removed from the matching variable in the analysis for each regional group, thus "purifying" the matching variable, as it were, of items with large D–DIF values. This step is a response to the possible criticism, from those who hold to the hypothesis that the test is largely biased in favor of the domestically tested candidates, that the presence of biased items in the matching variable causes that variable itself to be biased and as a result tends to obscure the data that would otherwise reveal bias in other items.

One general observation is noteworthy in Tables 16–18. This is that the effort of "purifying" the matching variable of items with relatively large MH values did little to change the MH values of the items. The differences in these values under the two conditions were quite small. As expected, however, the new, purified, values were in nearly all instances slightly larger (i.e., more positive) than the original, unpurified, values. This is so because only those items with relatively large positive values--those presumably favoring the domestic groups--were removed in the purification process, not those items with large positive or large negative values, as is sometimes done in such analyses. The intent behind the "one–sided" purification was to give the hypothesis, that domestic candidates are unfairly advantaged by the test, a stronger chance of being supported.[3] As a consequence, the differences between the mean MH values in Tables 13–15 and those in the corresponding Tables 16–18 are in the positive direction. In some instances, as in the comparison of means in Tables 14 and 17, based on Section II data, there are no differences at all. This is so because there were no items in Section II with MH D–DIF values of 2.00 or greater, and therefore no need to purify the total score.

As will later be observed in examining Table 19, only three items in the test (item 13 in Section I and items 4 and 7 in Section III) meet the operational criteria for item discrepancy in more than one regional group; except for those three items, the relatively large MH D–DIF values are idiosyncratic with respect to the group in which they were observed. Further, it should be noted that only one of the three exceptional items referred to above--indeed, the only such item in the test, item 13 in Section

[3] It is recalled that there was an additional, earlier attempt, in the selection of the samples for this study, to strengthen the opportunity for the hypothesis to be supported. There, instead of choosing all, or samples of all, the domestic and foreign candidates for whom data were available, we chose only those domestic candidates who had lived in the United States a year or more, during which time they would have had more opportunity to familiarize themselves with Americana. We also chose for the study those foreign candidates who had spent less than one month in the United States, thus removing anyone from the study sample who might have visited here long enough to become acquainted with Americana.

I--yields a large MH D—DIF value in every region studied, and is the only one that appears consistently in all the analyses of the total of all regional groups combined. It is therefore the only item in the test that can reasonably be considered an item that distinguishes domestic students from foreign students in the sense considerd here.

Conceivably, there might be items of American content that favor domestic over foreign examinees only in some regions of the world, not all. Such items would suggest a lower degree of familiarity with Americana in those regions than exists in others. But this is not how one reads Traynor's (1985) concern about TOEFL, which is that foreign-tested examinees are generally at a disadvantage relative to domestically tested examinees. Furthermore, it is interesting that Table 19 shows a greater number of significant DIF values in the Americas and Europe regions, regions that would appear to have more in common with the United States in language and culture than the Africa, Asia, and Mideast regions rather than less.

In support of the foregoing conclusion, it will be interesting to see the behavior of the items in the test in graphic form. Figures 1, 2, and 3 show "delta plots" of these items, section by section. These delta plots were formed by (1) calculating the percent-pass for each item in the domestic group, and, separately, again in the foreign group; (2) converting each percent-pass figure to the usual ETS delta scale by reading the normal deviate value (z) corresponding to that percentage in a table of the normal curve; and (3) expressing the value on a 13-4 scale by the formula, $\Delta = -4z + 13$. The delta for each item as observed in the domestic group was then plotted against its delta as observed in the foreign group. These are the delta plots seen in Figures 1, 2, and 3.

As a procedural note, it is recalled from Table 3 that the foreign samples were each larger than the corresponding domestic sample, but by different proportions. In forming the samples for which these deltas were calculated, care was taken to "control" the two groups, as it were, with respect to the numbers of cases in each region. Therefore, further random sampling was undertaken so that each foreign group would contain exactly 1.2 times the number in its corresponding domestic group. (The figure of 1.2 was arrived at by observing that it was the smallest of all five regional ratios--observed in the Americas region--representing the ratio of the number of foreign to domestic candidates.) All surplus foreign cases in each region were randomly removed from that region's foreign sample. As is undoubtedly apparent, the purpose of this sampling was to ensure that the total foreign group had the same proportionate representation from each region as was contained in the total domestic group while ensuring that the representativeness of the candidates in each foreign group relative to all the study group candidates in their region was not disturbed.

Figure 1 is an excellent illustration of the delta plot outcome. Here we see that the items are somewhat dispersed, in the sense that they do not cluster as closely about the 45° line or any other line parallel to it as is true of the item points in Figures 2 and 3. With one exception, the item points in Figure 1 appear above the 45° line, indicating that they were more difficult for the foreign group than for the domestic group. This observation is consistent with the observation, made in Tables 4 and 5, that each of the foreign groups found the Listening Comprehension section of the test more difficult than did its corresponding domestic group. It is an understandable finding, inasmuch as the domestic group had had more exposure to spoken English than had their foreign compatriots.

But perhaps most interesting in Figure 1 is the appearance of one item in the Listening Comprehension section that is clearly separate from the rest. This item is especially difficult for the foreign group, more so than the other items in the section. As expected from the earlier discussion, this item is item 13.

The delta plots for Sections II and III (Figures 2 and 3) are unremarkable. In both plots the items cluster closely about the line, and slightly below it on average, indicating, as seen earlier in Tables 6 and 7 and again in Tables 8 and 9, that the foreign groups are generally higher-scoring than the corresponding domestic groups in Sections II and III, the nonlistening parts of the test--but not by much. As seen in these plots most of the items appear as slightly more difficult for the domestic groups than for the foreign groups. In neither the Section II nor the Section III plot are there items that are noticeably aberrant. It remains now to examine in detail the items in TOEFL that appeared to be especially aberrant and to consider whether some rationale can be deduced for their behavior. Their MH D-DIF values can be seen in Table 19.

Table 19 is a display of all the items in TOEFL whose MH D-DIF values (a) exceeded ± 1.5 and (b) differed significantly ($t \geq 1.96$) from 1.0, favoring the domestic group, or from -1.0, favoring the foreign group. The item that is most noteworthy in this table has already been identified, item 13 in Section I. It has original (i.e., "unpurified") MH D-DIF values of 4.93 for the Africa region, 2.19 for the Americas region, 3.75 for the Asia region, 3.47 for Europe, and 3.51 for the Mideast, all positive and sizable values, indicating that the item favors the domestic group in every region. It also appears in the total group, when matching is done only on score (MH D-DIF = 3.68), and again when the total group is matched on score and region (MH D-DIF = 3.66). Every one of the MH values is statistically discrepant when evaluated against the criteria that have been adopted in the ETS testing programs in which differential item functioning studies have been conducted with respect to ethnic group and in studies conducted with respect to sex. No other item in the test shows this level and degree of consistent aberrant behavior, appearing in all five regions as well as in the combined group across all regions.

It will be of some interest to examine item 13 and speculate on the reasons for its aberrant behavior. The item depends on a recorded voice on tape making the statement, "The number you have reached is not in service. Please check your listing and dial again." The examinee is then asked to choose from among the following explanations for the statement:

(A): The call can't be completed as dialed.
(B): The service department received a call.
(C): The check was sent to the phone company.
(D): The line is busy and should be tried again.

The correct response, of course, is (A), which was chosen by 75% of the domestic examinees and by 35% of the foreign examinees in the combined group, across all regions.[4] One can only speculate on the reason, or reasons, for the difference in the success rates, but item analysis data tell us that many more (43%) of the examinees in the foreign group were attracted by response (D) than were examinees in the domestic group (16%), possibly because of the similarity of the wording in that response and the wording in the tape. On the other hand, the language of the item is generally characteristic of the language conventions of American telephone operators. It is entirely possible, therefore, that it was not any explicit American content in this item, but the nature of the language, more familiar to domestic than to foreign examinees, that caused the item to favor domestic candidates. If this is indeed the case, it may be useful for the developers of TOEFL, and their advisory committee members, to consider, as a matter of policy, whether the test should or should not contain items that reflect the characteristic types of English language normally used in different contexts in the United States.

In any case, for whatever reason, item 13 in Section I consistently disfavors the foreign candidates. At the same time, it is important to note, in the context of this study, that item 13 was not judged to be an "Americana" item in the sense of the Traynor hypothesis by any of the five raters. Indeed, none of the items in Section I that were flagged as positively discrepant (i.e., favoring the domestic group) was judged by any of the raters to be an Americana item. This was also true of the one flagged item, item 8, in Section II. However, one of the positively discrepant items in Section III, item 4, was judged as an Americana item by four of the five raters.

Another item that appeared as discrepant was item 24, also in the Listening Comprehension section. What is of some interest here is that it appeared only in the European region, but in that region it had an even higher MH D-DIF value (3.62) than did item 13 in the European region. The script for this item reads as follows:

(W): What happened to your leg? Did you fall down skiing?
(MB): I tripped over the carpet at work.
(MA): What happened to the man?

[4] It should be noted that the percent-pass figures cited here, and in discussions of other aberrant items, were observed in the original item-analysis data, tabulated before matching on total score was undertaken.

The response options follow:

(A): He hurt himself on a skiing trip.
(B): He was hit by a car.
(C): He fell on his pet.
(D): He had an accident at work.

Clearly, response (D) is correct, and 96% of the European domestic group but only 76% of the European foreign group answered the item correctly. There were some small numbers of the foreign group that were attracted to response (B)--11%--and to response (C)--8%--possibly because of component parts in the word "carpet" of the word "car" in response (B) and the word "pet" in response (C). (Fewer people in the domestic group--less than 2% and less than 1%, respectively--made these errors.) But again, this confusion, if this is indeed what caused the failures for the foreign group, occurred only in the European region.

Only one item in Section II--item 8--appeared to be discrepant, and to a relatively small degree (MH D-DIF = 1.65), smaller than in items 13 and 24 in Section I, and smaller than in the items in Section III that will be discussed shortly. Furthermore, it appeared only in the European region.

The stem of item 8, Section II, reads as follows:

> Tar forms during the distillation of organic matter
> such as coal, wood, oils, fats, ---- of various sorts.

The options given for the omission are these:

(A): wastes or
(B): wastes as
(C): are wastes
(D): and wastes

Sixty-three percent of the domestic European group answered this item correctly, but only 43% of the foreign European group answered it correctly. In no other region was a difference of this magnitude observed, in either direction. Again, the reason for the item's discrepant behavior is elusive, at best.

Item 4 in Section III appeared as discrepant in both the Americas (MH D-DIF = 2.35) and the Mideast (MH D-DIF = 1.71) regions. The stem of the item reads as follows:

> In The Race Track, also titled Death of a Pale Horse, Albert Pinkham Ryder portrays the involuntary remembrance of a nightmare.

The response options for this item, calling for a synonym of the double-underscored word, are:

(A): sickness
(B): fear
(C): horse race
(D): bad dream

The difference in performance between the two Americas groups was only .07 (93% correct for the domestic group and 86% correct for the foreign group). The difference in performance between the two Mideast groups was only .10 (88% correct for the domestic group and 78% correct for the foreign group). One can only speculate about the reasons for the confusion, and, as with other items, it is difficult to account for the difference in performance between the domestic and foreign groups or for the fact that the item appeared as discrepant in two such dissimilar regions.

Finally, item 7 in Section III also appeared discrepant in two regions, in this instance, the Americas (MH D—DIF = 2.26) and Europe (MH D--DIF = 2.77), which do share, to some extent, similarities in language. (The inhabitants of all the countries in Central and South America speak Latin-derived languages; and at least some of the European languages derive from Latin.) The stem in item 7 is as follows:

Most rivers <u>overflow</u> their banks about once every two years.

The response options read:

(A): stay over
(B): flip over
(C): flood
(D): carry over

The difference in the p-values for the two Americas groups was about .17 (the p-values for the domestic and foreign groups were .59 and .42, respectively), and the difference in the p-values for the two European groups was about .26 (the p-values for the domestic and foreign groups were .81 and .55, respectively). The confusion in this item may have come from the appearance of the word part "over-" in the stem as well as in two of the four options of the item. The item appeared to be of moderate-to-easy difficulty for both domestic and foreign groups in the Americas region as well as in the European region.

Thus, while it may be possible to speculate, on the reasons for the confusions and failure of some examinees to answer some of these items correctly, it is not at all clear why these confusions were more damaging to the foreign candidates than to the domestic candidates.

<u>Relationship Between Americana Ratings and Item Discrepancies Indices</u>

In another part of this analysis, an attempt was made to examine the relationship between the ratings of Americana described earlier in this report with the Mantel-Haenszel results. The results of this attempt are shown in Tables 20-22. These tables give distributions of MH D-DIF values for items that had "Americana scores" of 0, 1, 2, 3, 4, and 5. These scores were formed separately for each item simply by counting the number of raters who classified the item as having reference to American people, places, institutions, customs, etc. For example, if two raters classified an item as an Americana item and the other three classified it as a non-Americana item, the item received an Americana score of 2.

As may be seen in Tables 20-22, there seems to be no relationship at all between the Americana scores and the MH values. This conclusion can be drawn in a simple visual review of the bivariate distributions and can be verified by scanning the means at the foot of each column of frequencies. In none of these tables do we see a progression in the means to show that as the Americana score increases (or decreases), the mean MH values also increase (or decrease). In point of fact, there is no progression at all in the mean MH values.

In this connection it may be pointed out that four of the five discrepant items discussed above had an Americana scores of 0; none of the raters could discern any American content in them. The fifth, item 4 in Section III, having to do with the work of Albert Pinkham Ryder, received an Americana score of 4, probably because the work of an identified American writer was named. But there does not appear to be any relation between the American content of the stem of the item in relation to its options and in relation to the statistical observation that the item is discrepant. On the strength of the information provided in these and earlier tables, and on the strength of the review of the items cited here, it may be concluded that there is no support for the hypothesis that TOEFL items that make reference to American people, places, institutions, customs, and so forth, tend to advantage TOEFL candidates who have lived in the United States for a year or more over those who have spent little (one month or less) or no time here.

SUMMARY AND CONCLUSIONS

This study was undertaken in an effort to examine the claim advanced by Traynor (1985) that much of TOEFL is "American," which means, he maintains, "that a student with knowledge of American history, American geography, American sport, etc., has a decided advantage, both practically and psychologically" over a student who is not familiar with Americana. In an effort to convert Traynor's claim to a testable hypothesis, groups of students were formed, those who had been tested in the United States and had lived here more than a year, and those who had been tested (on the same form of the test) in foreign countries and who had never lived in the United States for more than a month. These restrictive definitions of the two groups were effected in order to magnify the difference in their acquaintance with Americana (while still ensuring enough data to carry out the intended study reliably), and thus to give the hypothesis a greater opportunity to be supported by the data.

Once these groups were defined, the 146 operational items of TOEFL were subjected to a Mantel-Haenszel (1959) analysis, which determined the degree to which the domestically tested group fared better on an item than the corresponding foreign-tested group, after the groups were matched on ability, as measured by the total score on the section of the test that contained the item. This analysis was carried out separately for each of the five regions of the world from which TOEFL candidates come in sufficient numbers. It was also carried out for the combined group, aggregating across regions, matching first on total score alone, as just carried out, and, again, matching on total score and region. Thus, for each item several Mantel-Haenszel (MH D-DIF) indices were found, separately for region and for the overall total group. Items were identified for which any MH D-DIF index exceeded both of two criteria: (a) that the absolute value of the index exceeded 1.5 (positive value indicating an advantage for the domestic group, negative value indicating an advantage for the foreign group) and (b) that the absolute value of the index exceeded 1.00 by an amount that yielded a t-value of 1.96 or greater.

Several items were found that satisfied the criteria stated above, but only one that did so in all regions of the world. This item, and three other items that either yielded larger MH D-DIF values or were found to satisfy these criteria in more than one region of the world (none were found in more than two regions) were found and discussed. A fifth item was also considered, principally because it was the only one in that section of the test (Section II) that satisfied the operational criteria given above. In no instance could a rationale be found for the large MH D-DIF index in a way that was related to the Traynor hypothesis.

In another part of the study, five raters were asked to review the items in the test and judge whether or not they contained American content. The first outcome of this judgment process worth noting was that there was a high degree of agreement among the raters; the large majority of the items were judged unanimously by the raters as either having American content or not. On very few of the items was there any disagreement among the raters

In any case, for each item we developed an Americana score, simply the number of raters who judged the item to contain specific American content. Thus, the Americana score had a possible range from 0 (no one judged it an Americana item) to 5 (all five raters judged it an Americana item). For each level of score, 0-5, distributions were run of MH D-DIF indices to determine whether there was any relationship between the two scores, or indices. Tables showing these distributions were prepared, one for each of the three sections of TOEFL. Our findings were that there was no relation at all between the Americana scores and the MH D-DIF values, indicating again that the American content of the item was irrelevant in determining the size of the MH D-DIF index. The conclusion drawn from these analyses is that there is no evidential support, at least in these data, for the hypothesis that the

American content in some of the items of TOEFL favors students who have lived in the United States for any length of time.

REFERENCES

Alderson, J. C., & Urquhart, A. H. (1985). This test is unfair: I'm not an economist. In Hauptman, P. C., Leblanc, R., and Wesche, M. B. (Eds.) Second-language performance testing. Ottowa: University of Ottowa Press, pp. 25-43.

Bernhardt, E. B. (1984). Toward an information processing perspective in foreign language reading. The Modern Language Journal, 68, 322-331.

Carrell, P. L. (1984). Schema theory and ESL reading: Classroom implications and applications. The Modern Language Journal, 68, 333-343.

Educational Testing Service (1987). TOEFL test and score manual, 1987-88 Edition. Princeton, NJ: Educational Testing Service.

Erickson, M., & Molloy, J. (1983). ESP test development for engineering students. In J. W. Oller (Ed.), Issues in language testing research (pp. 280-288). Rowley, MA.: Newbury House.

Hale, G. A. (1988). The interaction of student major-field group and text content in TOEFL reading comprehension. TOEFL Research Report No. 25. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Melendez, E. J., & Pritchard, R. H. (1985). Applying schema theory to foreign language reading. Foreign Language Annals, 18, 399-403.

Schmeiser, C. B., & Ferguson, R. L. (1978). Performance of black and white students on test materials containing content based on black and white cultures. Journal of Educational Measurement, 15, 193-200.

Traynor, R. (1985). The TOEFL: An appraisal. ELT Journal, 39, 43-47.

Table 1

Distributions of Reported Amount of Time
Spent Living in the U.S., by Region, for the Candidates
Tested Domestically, by Region of Origin

	Region of Native Country						
No. of Months in U.S.	Africa	Americas	Asia	Europe	Mideast	Pacific	Total
95 +	37	124	256	28	61	3	509
90 – 94	1	3	18	4	4	0	30
85 – 89	1	2	17	4	4	0	28
80 – 84	5	9	62	9	7	0	92
75 – 79	5	5	37	2	3	0	52
70 – 74	4	16	83	10	8	0	121
65 – 69	2	10	53	8	4	0	77
60 – 64	18	24	116	15	22	0	195
55 – 59	5	2	86	5	7	0	105
50 – 54	7	10	94	11	9	0	131
45 – 49	15	24	188	19	29	2	277
40 – 44	6	15	217	13	30	0	281
35 – 39	16	39	334	18	82	0	489
30 – 34	17	29	370	28	78	0	522
25 – 29	18	16	285	17	61	0	397
20 – 24	25	66	586	44	152	1	874
15 – 19	35	73	806	61	186	5	1,166
10 – 14	53	114	886	51	203	2	1,309
5 - 9	145	269	1,990	341	453	6	3,204
0 – 4	178	241	3,078	260	398	14	4,169
No. of Cases	593	1,091	9,562	948	1,801	33	14,028
Mean	21.03	26.20	17.32	16.97	18.43	17.45	18.28
Median	8.27	11.10	8.41	6.96	11.42	6.25	8.60
Std. Dev.	27.44	31.75	22.26	23.29	21.43	28.34	23.48

Table 2

Distributions of Responses to the Question:
"Have You Spent More Than One Month in the U.S.?"
for the Candidates Tested in Foreign Countries,
by Region of Native Country

More Than One Month in the U.S.?	Africa	Americas	Asia	Europe	Mideast	Pacific	Total
Yes	166(12.2)	534(46.9)	2,074(11.8)	1,143(26.8)	351(22.0)	0(0)	4,268(16.4)
No	1,194(87.8)	604(53.1)	15,540(88.2)	3,117(73.2)	1,248(78.0)	2(100)	21,705(83.6)
Total	1,360	1,138	17,614	4,260	1,599	2	25,973

Note: Numbers in parentheses are the percentages of the total in each region.

Table 3

Distribution of Cases in the Study Samples

Region of Native Country	Domestic Samples (More Than One Year in the U.S.)	Foreign Samples (One Month or Less in the U.S.)	Total
Africa	236 (4.1)	1,193 (5.5)	1,429
Americas	505 (8.7)	604 (2.8)	1,109
Asia	3,900 (67.3)	15,496 (71.6)	19,396
Europe	308 (5.3)	3,115 (14.4)	3,423
Mideast	850 (14.7)	1,244 (5.7)	2,094
Total	5,799 (100.1)	21,652 (100.0)	27,451

Note: Numbers in parentheses are the percentages of the totals (5,799 or 21,652) in all regions.

Table 4

Distributions of Raw Scores
On Section I, Listening Comprehension Section of TOEFL,
for the Domestic Samples

Raw Score (Rights Only)	Region of Native Country					
	Africa	Americas	Asia	Europe	Mideast	Total
50	3	3	12	12	2	32
48 – 49	12	32	71	50	20	185
46 – 47	9	51	173	49	30	312
44 – 45	23	62	236	39	41	401
42 – 43	27	43	275	28	55	428
40 – 41	18	38	333	32	71	492
38 – 39	22	39	347	22	66	496
36 – 37	22	31	368	14	96	531
34 – 35	19	39	388	18	86	550
32 – 33	22	36	340	11	75	484
30 – 3i	14	31	342	10	56	453
28 – 29	12	21	263	7	82	385
26 – 27	12	19	223	11	51	316
24 – 25	10	13	170	2	37	232
22 – 23	3	16	126	0	33	178
20 – 21	5	11	105	2	19	142
18 – 19	0	10	53	1	14	78
16 – 17	1	7	40		9	57
14 – 15	1	2	15		3	21
12 – 13	1	1	13		3	18
10 – 11			6		1	7
8 – 9			0			0
6 – 7			1			1
4 – 5						
2 – 3						
0 – 1						
No. of Cases	236	505	3,900	308	850	5,799
Mean	36.52	36.93	34 38	41.49	34.02	35.01
Std. Dev.	7.61	8.54	7.68	6.77	7.63	7.90

Table 5

Distributions of Raw Scores
On Section I, Listening Comprehension Section of TOEFL,
for the Foreign Samples

	Region of Native Country					
Raw Score (Rights Only)	Africa	Americas	Asia	Europe	Mideast	Total
50		3	22	23	2	50
48 – 49	25	9	137	127	19	317
46 – 47	47	16	313	239	32	647
44 – 45	59	28	406	277	40	810
42 – 43	43	35	460	319	47	904
40 – 41	61	24	578	308	49	1,020
38 – 39	58	53	721	260	57	1,149
36 – 37	89	52	914	241	65	1,361
34 – 35	79	36	1,059	262	72	1,508
32 – 33	88	46	1,213	229	103	1,679
30 – 31	86	51	1,292	205	78	1,712
28 – 29	80	58	1,353	168	104	1,763
26 – 27	80	31	1,327	161	113	1,712
24 – 25	77	43	1,324	111	109	1,664
22 – 23	82	36	1,242	79	91	1,530
20 – 21	80	30	1,090	48	98	1,346
18 – 19	61	16	871	27	61	1,036
16 – 17	42	17	547	14	43	663
14 – 15	32	9	386	14	34	475
12 – 13	15	8	153	2	15	193
10 – 11	7	3	69	1	6	86
8 – 9	2		16		5	23
6 – 7			3		1	4
4 – 5						
2 – 3						
0 – 1						
No. of Cases	1,193	604	15,496	3,115	1,244	21,652
Mean	30.45	31.54	29.02	36.54	29.21	30.26
Std. Dev.	9.34	8.65	8.35	7.73	8.87	8.76

Table 6

Distributions of Raw Scores
On Section II, Structure and Written Expression Section of TOEFL,
for the Domestic Samples

Raw Score (Rights Only)	Region of Native Country					
	Africa	Americas	Asia	Europe	Mideast	Total
38	8	10	41	16	5	80
36 - 37	25	39	199	55	28	346
34 - 35	26	48	333	55	33	495
32 - 33	25	57	434	47	45	608
30 - 31	15	54	461	20	59	609
28 - 29	25	44	445	33	82	629
26 - 27	19	58	445	29	94	645
24 - 25	21	41	391	16	112	581
22 - 23	19	41	368	18	113	559
20 - 21	16	29	274	8	75	402
18 - 19	6	26	225	4	61	322
16 - 17	13	23	134	5	55	230
14 - 15	9	17	78	2	36	142
12 - 13	6	9	37		26	78
10 - 11	2	4	23		9	38
8 - 9	1	5	7		12	25
6 - 7			3		1	4
4 - 5			1		2	3
2 - 3			0		0	0
0 - 1			1		2	3
No. of Cases	236	505	3,900	308	850	5,799
Mean	27.23	26.89	26.87	30.93	23.94	26.67
Std. Dev.	7.25	6.89	6.10	5.42	6.55	6.42

Table 7

Distributions of Raw Scores

On Section II, Structure and Written Expression Section of TOEFL,

for the Foreign Samples

Raw Score (Rights Only)	Region of Native Country Africa	Americas	Asia	Europe	Mideast	Total
38	66	6	222	62	14	370
36 – 37	190	40	978	338	48	1,594
34 -- 35	142	58	1,355	488	69	2,112
32 – 33	122	74	1,626	476	71	2,369
30 – 31	101	68	1,825	413	96	2,503
28 – 29	110	80	1,695	389	106	2,380
26 – 27	94	57	1,663	281	126	2,221
24 – 25	85	64	1,543	221	130	2,043
22 – 23	82	43	1,328	161	120	1,734
20 – 21	65	33	1,022	117	131	1,368
18 – 19	49	31	831	65	118	1,094
16 – 17	26	15	569	46	86	742
14 – 15	28	18	402	28	51	527
12 – 13	22	9	245	20	40	336
10 – 11	6	6	124	9	24	169
8 – 9	4	2	49	0	7	62
6 – 7	1		14	1	4	20
4 – 5			3		3	6
2 – 3			0			0
0 – 1			2			2
No. of Cases	1,193	604	15,496	3,115	1,244	21,652
Mean	28.84	27.22	26.84	29.63	24.11	27.21
Std. Dev.	6.97	6.36	6.43	5.51	6.85	6.49

Table 8

Distributions of Raw Scores
On Section III, Vocabulary and Reading Comprehensior Section of TOEFL,
for the Domestic Samples

Raw Score (Rights Only)	Africa	Americas	Asia	Europe	Mideast	Total
	Region of Native Country					
57 – 58	6	10	26	20	3	65
54 – 56	7	33	138	62	14	254
51 – 53	15	47	196	38	19	315
48 – 50	15	37	271	41	18	382
45 – 47	13	53	339	32	20	457
42 – 44	19	52	425	18	39	553
39 – 41	17	64	415	38	55	589
36 – 38	24	35	411	13	65	548
33 – 35	29	46	409	20	92	596
30 – 32	24	40	353	14	92	523
27 – 29	15	32	312	4	108	471
24 – 26	20	20	252	5	93	390
21 – 23	19	11	186	2	104	322
18 – 20	9	14	106	1	79	209
15 – 17	2	6	43		34	85
12 – 14	1	5	14		8	28
9 – 11	1		2		4	7
6 – 8			1		1	2
3 – 5			0		0	0
0 – 2			1		2	3
No. of Cases	236	505	3,900	308	850	5,799
Mean	36.11	39.73	37.21	46.15	30.37	36.85
Std. Dev.	10.56	10.31	9.70	8.69	9.74	10.32

Table 9

Distributions of Raw Scores

On Section III, Vocabulary and Reading Comprehension Section of TOEFL,

for the Foreign Samples

Raw Score (Rights Only)	Region of Native Country Africa	Americas	Asia	Europe	Mideast	Total
57 – 58	21	12	193	120	4	350
54 – 56	60	37	599	392	20	1,108
51 – 53	77	63	961	429	32	1,562
48 – 50	88	75	1,214	366	31	1,774
45 – 47	96	65	1,401	325	54	1,941
42 – 44	121	63	1,611	310	54	2,159
39 – 41	131	65	1,632	253	66	2,147
36 – 38	120	69	1,569	229	82	2,069
33 – 35	129	35	1,540	194	102	2,000
30 – 32	96	45	1,418	156	130	1,845
27 – 29	95	24	1,191	119	173	1,602
24 – 26	70	22	925	100	159	1,276
21 – 23	35	12	625	65	143	880
18 – 20	28	13	362	37	94	534
15 – 17	14	2	176	15	61	268
12 – 14	11	2	61	2	28	104
9 – 11	0		10	3	7	20
6 – 8	1		2		4	7
3 – 5			0			0
0 – 2			6			6
No. of Cases	1,193	604	15,496	3,115	1,244	21,652
Mean	38.20	41.27	37.93	43.48	30.30	38.40
Std. Dev.	10.08	9.53	9.89	9.90	10.09	10.29

Table 10

Intercorrelations Among the Sections of TOEFL
for the Domestic Samples

Africa; N = 236

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.756	.720	55.22	6.10
Section II	.756	1.000	.799	51.31	9.13
Section III	.720	.799	1.000	49.39	7.27

Americas; N = 505

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.755	.705	55.64	6.78
Section II	.755	1.000	.807	50.75	8.58
Section III	.705	.807	1.000	51.74	7.13

Asia; N = 3,900

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.608	.665	53.51	5.99
Section II	.608	1.000	.788	50.55	7.44
Section III	.665	.788	1.000	50.06	6.51

Europe; N = 308

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.707	.676	59.43	5.66
Section II	.707	1.000	.772	55.95	7.45
Section III	.676	.772	1.000	56.31	6.26

Mideast; N = 850

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.724	.706	53.18	5.95
Section II	.724	1.000	.774	47.10	7.63
Section III	.706	.774	1.000	45.45	6.76

Total Group; N = 5,799

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.656	.678	54.03	6.22
Section II	.656	1.000	.796	50.38	7.86
Section III	.678	.796	1.000	49.83	7.02

Table 11

Intercorrelations Among the Sections of TOEFL
for the Foreign Samples

Africa; N = 1,193

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.676	.707	50.60	7.09
Section II	.676	1.000	.808	53.44	9.03
Section III	.707	.808	1.000	50.78	6.90

Americas; N = 604

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.678	.603	51.34	6.61
Section II	.678	1.000	.777	51.01	7.79
Section III	.603	.777	1.000	52.79	6.49

Asia; N = 15,496

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.649	.663	49.46	6.27
Section II	.649	1.000	.804	50.57	7.89
Section III	.663	.804	1.000	50.56	6.67

Europe; N = 3,115

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.656	.634	55.24	6.18
Section II	.656	1.000	.784	53.99	7.12
Section III	.634	.784	1.000	54.37	6.90

Mideast; N = 1,244

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.724	.764	49.63	6.71
Section II	.724	1.000	.802	47.36	8.26
Section III	.764	.802	1.000	45.33	7.12

Total Group; N = 21,652

	Section			Scaled Score	
	I	II	III	M	S.D.
Section I	1.000	.663	.674	50.42	6.65
Section II	.663	1.000	.804	51.06	8.03
Section III	.674	.804	1.000	50.88	7.01

Table 12

Intercorrelations Among Raters for the Items in TOEFL

Section I; No. of Items = 50

			Rater			Mean	S.D. of
	Z	L	R	C	A	Rating	Ratings
Z	1.00	.86	.86	.86	.73	0.08	0.27
L	.86	1.00	1.00	1.00	.86	0.06	0.24
R	.86	1.00	1.00	1.00	.86	0.06	0.24
C	.86	1.00	1.00	1.00	.86	0.06	0.24
A	.73	.86	.86	.86	1.00	0.08	0.27

Reliability of the sum of 5 ratings: .972

Section II; No. of Items = 38

			Rater			Mean	S.D. of
	Z	L	R	C	A	Rating	Ratings
Z	1.00	.76	1.00	1.00	.86	0.21	0.41
L	.76	1.00	.76	.76	.88	0.32	0.46
R	1.00	.76	1.00	1.00	.86	0.21	0.41
C	1.00	.76	1.00	1.00	.86	0.21	0.41
A	.86	.88	.86	.86	1.00	0.26	0.44

Reliability of the sum of 5 ratings: .970

Section III; No. of Items = 58

			Rater			Mean	S.D. of
	Z	L	R	C	A	Rating	Ratings
Z	1.00	.76	.91	.76	.83	0.28	0.45
L	.76	1.00	.76	.93	.76	0.40	0.49
R	.91	.76	1.00	.76	.91	0.28	0.45
C	.76	.93	.76	1.00	.68	0.40	0.49
A	.83	.76	.91	.68	1.00	0.28	0.45

Reliability of the sum of 5 ratings: .954

Total Test; No. of Items = 146

			Rater			Mean	S.D. of
	Z	L	R	C	A	Rating	Ratings
Z	1.00	.78	.93	.84	.83	0.19	0.39
L	.78	1.00	.80	.89	.82	0.26	0.44
R	.93	.80	1.00	.86	.89	0.18	0.39
C	.84	.89	.86	1.00	.76	0.23	0.42
A	.83	.82	.89	.76	1.00	0.21	0.40

Reliability of the sum of 5 ratings: .963

Table 13

Distributions of Original ("Unpurified") Mantel-Haenszel Indices, by Region, for Section I, Listening Comprehension

No. of Items = 50

	Region of Native Country					Total Group	Total Group
MH D-Dif Index	Africa	Americas	Asia	Europe	Mideast	(Matched on Score)	(Matched on Score and Region)
4.50- 4.999	1						
4.00- 4.499							
3.50- 3.999			1	1	1	1	1
3.00- 3.499				1			
2.50- 2.999				1			
2.00- 2.499		1		1			
1.50- 1.999	2	1		1	1		
1.00- 1.499	3	2	2	3	3	4	3
0.50- 0.999	7	7	5	2	8	3	3
0.00- 0.499	13	11	13	8	3	14	13
-0.50--0.001	6	16	18	14	18	13	14
-1.00--0.501	10	7	10	10	12	14	15
-1.50--1.001	6	2	1	6	4	1	1
-2.00--1.501	1	1					
-2.50--2.001		2		2			
-3.00--2.501	1						
Mean	-0.03	-0.06	-0.02	-0.05	-0.05	-0.03	-0.03
Std. Dev.	1.14	0.83	0.77	1.22	0.84	0.78	0.76

Table 14

Distributions of Original ("Unpurified") Mantel-Haenszel Indices, by Region, for Section II, Structure and Written Expression

No. of Items = 38

	Region of Native Country					Total Group	Total Group
						(Matched	(Matched on
MH D-Dif Index	Africa	Americas	Asia	Europe	Mideast	on Score)	Score and Region)
3.00- 3.499							
2.50- 2.999							
2.00- 2.499							
1.50- 1.999	1			5	1		
1.00- 1.499	1	2	2	3	1		1
0.50- 0.999	6	5	4	5	3	4	3
0.00- 0.499	13	10	15	9	13	15	15
-0.50--0.001	10	11	13	5	13	16	16
-1.00--0.501	4	9	3	6	7	2	3
-1.50--1.001	3	1	1	2		1	
-2.00--1.501				3			
-2.50--2.001							
-3.00--2.501							
Mean	0.04	.00	0.05	0.17	0.04	0.05	0.05
Std. Dev.	0.66	.58	0.47	0.98	0.52	0.45	0.43

Table 15

Distributions of Original ("Unpurified") Mantel-Haenszel Indices, by Region, for Section III, Vocabulary and Reading Comprehension

No. of Items = 58

	Region of Native Country					Total Group (Matched on Score)	Total Group (Matched on Score and Region)
MH D-Dif Index	Africa	Americas	Asia	Europe	Mideast		
3.00- 3.499							
2.50- 2.999				1			
2.00- 2.499		2					
1.50- 1.999		1		2	1		
1.00- 1.499	5	2	1	5	1		
0.50- 0.999	8	11	4	7	7	8	5
0.00- 0.499	21	10	26	14	20	23	27
-0.50--0.001	10	14	20	14	20	20	20
-1.00--0.501	9	12	5	9	5	5	5
-1.50--1.001	4	3	2	4	3	2	1
-2.00--1.501	1	2		2			
-2.50--2.001		1					
-3.00--2.501					1		
Mean	-0.01	-0.04	0.00	0.02	-0.01	0.01	0.01
Std. Dev.	0.68	0.88	0.46	0.85	0.65	0.42	0.42

Table 16

Distributions of "Purified" Mantel-Haenszel Indices, by Region, for Section I, Listening Comprehension

No. of Items = 50

	Region of Native Country					Total Group (Matched on Score)	Total Group (Matched on Score and Region)
MH D-Dif Index	Africa	Americas	Asia	Europe	Mideast		
5.00- 5.499	1						
4.30- 4.999							
4.00- 4.499							
3.50- 3.999			1	2	1	1	1
3.00- 3.499				1			
2.50- 2.999				1			
2.00- 2.499		1		1			
1.50- 1.999	4	1		1	1	1	
1.00- 1.499	1	4	2	2	5	3	4
0.50- 0.999	9	6	6	5	6	5	5
0.00- 0.499	13	11	14	10	8	13	13
-0.50--0.001	5	18	18	16	21	15	18
-1.00--0.501	10	6	8	7	7	11	8
-1.50--1.001	5	1	1	2	1	1	1
2.00--1.501	1						
2.50--2.001	1	2		2			
3.00--2.501							
3.50--3.001							
Mean	0.12	0.02	0.09	0.21	0.07	0.06	.08
Std. Dev.	1.15	0.84	0.78	1.22	0.83	0.78	0.76

Table 17

Distributions of "Purified" Mantel-Haenszel Indices, by Region, for Section II, Structure and Written Expression

No. of Items = 38

	Region of Native Country					Total Group (Matched on Score)	Total Group (Matched on Score and Region)
MH D-Dif Index	Africa	Americas	Asia	Europe	Mideast		
3.00- 3.499							
2.50- 2.999							
2.00- 2.499							
1.50- 1.999	1			5	1		
1.00- 1.499	1	2	2	3	1		1
0.50- 0.999	6	5	4	5	3	4	3
0.00- 0.499	13	10	15	9	13	15	15
-0.50--0.001	10	11	13	5	13	16	16
-1.00--0.501	4	9	3	6	7	2	3
-1.50--1.001	3	1	1	2		1	
-2.00--1.501				3			
-2.50--2.001							
-3.00--2.501							
Mean	0.04	0.00	0.05	0.17	0.04	0.05	0.05
Std. Dev.	0.66	0.58	0.47	0.98	0.52	0.45	0.43

Table 18

Distributions of "Purified" Mantel-Haenszel Indices, by Region, for Section III, Vocabulary and Reading Comprehension

No. of Items = 58

	Region of Native Country					Total Group	Total Group
						(Matched	(Matched on
MH D-Dif Index	Africa	Americas	Asia	Europe	Mideast	on Score)	Score and Region)
3.00- 3.499							
2.50- 2.999				1			
2.00- 2.499		3					
1.50- 1.999				2	1		
1.00- 1.499	5	2	1	6	1		
0.50- 0.999	8	13	4	7	7	8	5
0.00- 0.499	21	11	26	13	20	23	27
-0.50--0.001	10	15	20	15	20	20	20
-1.00--0.501	9	8	5	9	5	5	5
-1.50--1.001	4	4	2	3	3	2	1
2.00--1.501	1	1		2			
-2.50--2.001		1					
-3.00--2.501					1		
Mean	-0.01	0.02	0.00	0.09	-0.01	0.01	0.01
Std. Dev.	0.68	0.87	0.46	0.86	0.65	0.42	0.42

Table 19

MH D-DIF Values for Statistically Discrepant Items*

		Region of Native Country										Total Group (Matched on Score)		Total Group (Matched on Score and Region)	
		Africa		Americas		Asia		Europe		Mideast					
Section	Item No.	Orig.	Purif.	Orig.	Purif.	Orig.	Purif.	Orig.	Purif.	Orig.	Purif.	Orig.	Purif.	Orig.	Purif.
I	2	----	----	----	----	----	----	-2.38	----	----	----	----	----	----	----
	5	----	----	-2.33	-2.31	----	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	-2.22	----	----	----	----	----	----	----
	13	4.93	5.10	2.19	2.40	3.75	3.86	3.47	3.73	3.51	3.58	3.68	3.75	3.66	3.78
	24	----	----	----	----	----	----	3.62	3.97	----	----	----	----	----	----
	28	----	----	----	----	----	----	2.15	2.57	----	----	----	----	----	----
	30	----	----	----	----	----	----	----	----	----	1.61	----	----	----	----
	32	-1.82	----	----	----	----	----	----	----	----	----	----	----	----	----
	38	-2.52	-2.36	----	----	----	----	----	----	----	----	----	----	----	----
	44	----	----	----	----	----	----	----	----	----	----	----	1.51	----	----
II	8	----	----	----	----	----	----	1.65	1.65	----	----	----	----	----	----
III	4	----	----	2.35	2.24	----	----	----	----	1.71	1.71	----	----	----	----
	6	----	----	----	----	----	----	----	----	-2.84	-2.84	----	----	----	----
	7	----	----	2.26	2.14	----	----	2.77	2.83	----	----	----	----	----	----
	9	----	----	1.97	2.03	----	----	----	----	----	----	----	----	----	----
	27	----	----	-2.31	-2.29	----	----	----	----	----	----	----	----	----	----
	39	----	----	----	----	----	----	-1.82	----	----	----	----	----	----	----

*Discrepant items are items whose MH D-DIF values equal or exceed ± 1.5 and depart significantly ($t \geq 1.96$) from 1.0 (i.e., favoring the domestic group) or from -1.0 (i.e., favoring the foreign group). Numbers appear only in those cells for which the MH D-DIF values meet the foregoing criteria.

Table 20

Distributions of MH D-DIF Values for TOEFL Items
by Americana Score (Sum of Ratings
of American Reference)
Section I - Listening Comprehension

	Americana Score					
MH D-DIF Index	0	1	2	3	4	5
3.5- 3.999	1					
3.0- 3.499						
2.5- 2.999						
2.0- 2.499						
1.5- 1.999						
1.0- 1.499	3					
0.5- 0.999	3					
0.0- 0.499	11	1				1
-0.5--0.001	13					1
-1.0- 0.501	13	1				1
-1.5- 1.001	1					
-2.0- 1.501						
-2.5- 2.001						
-3.0- 2.501						
No. of Cases	45	2	0	0	0	3
Mean	-.02	-.13	-	-	-	-.10

Note: MH values are based on the total groups of domestic and foreign candidates, matched on total score and region of origin

Table 21

Distributions of MH D–DIF Values for TOEFL Items
by Americana Score (Sum of Ratings
of American Reference)
Section II – Structure and Written Expression

	Americana Score					
MH D-DIF Index	0	1	2	3	4	5
1.0- 1.499	1					
0.5- 0.999	1		1			1
0.0- 0.499	10	1				4
-0.5--0.001	11	1	1			3
-1.0- 0.501	3					
No. of Cases	26	2	2	0	0	8
Mean	-.03	-.06	.45	-	-	.21

Note: MH values are based on the total groups of domestic and foreign candidates, matched on total score and region of origin

Table 22

Distributions of MH D-DIF Values for TOEFL Items
by Americana Score (Sum of Ratings
of American Reference)
Section III – Vocabulary and Reading Comprehension

	Americana Score					
MH D-DIF Index	0	1	2	3	4	5
0.5- 0.999	2				1	2
0.0- 0.499	14	1	2	1		9
-0.5--0.001	15		3			2
-1.0- 0.501	2		1		1	1
-1.5- 1.001	1					
No. of Cases	34	1	6	1	2	14
Mean	-.04	.34	-.12	.35	-.15	.15

Note: MH values are based on the total groups of domestic and foreign candidates, matched on total score and region of origin

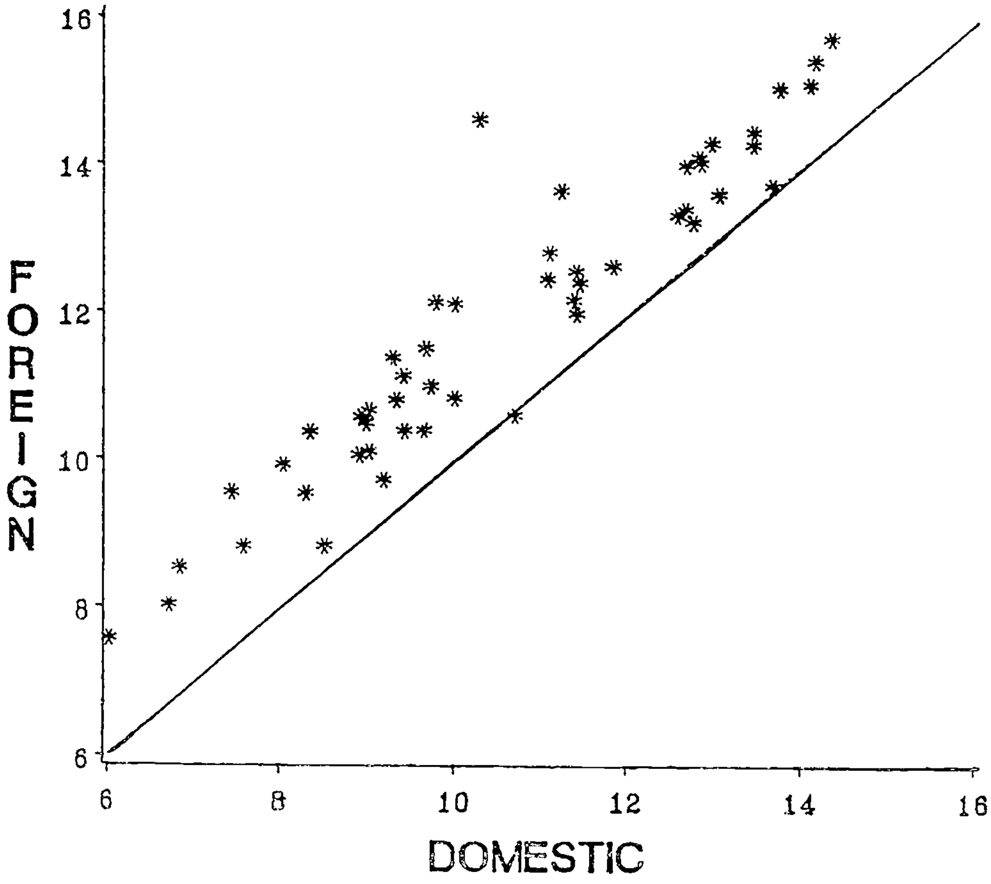

Figure 1. Plot of Deltas for the Combined Domestic Group (N = 5,799) vs the Combined Foreign Group (N = 6,957) on Section I (Listening Comprehension)

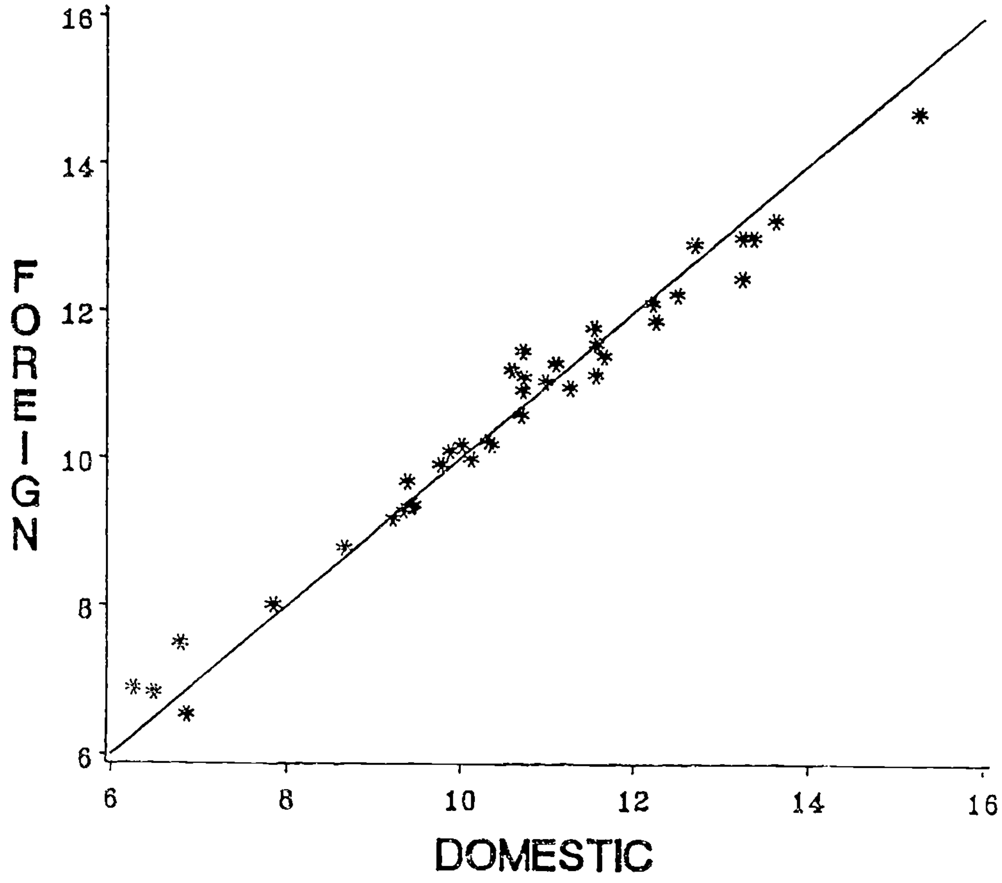

Figure 2. Plot of Deltas for the Combined Domestic Group (N - 5,799) vs the Combined Foreign Group (N - 6,957) on Section II (Structure and Written Expression)

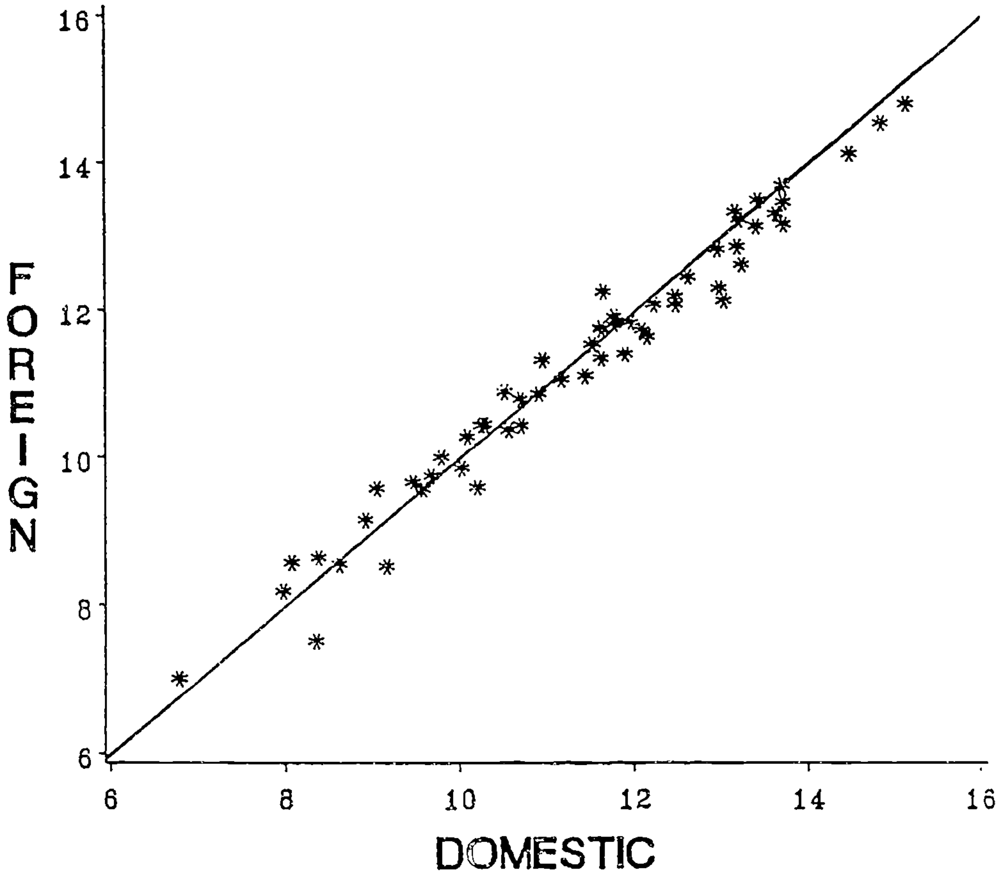

Figure 3. Plot of Deltas for the Combined Domestic Group (N - 5,799) vs the Combined Foreign Group (N ~ 6,957) on Section III (Vocabulary and Reading Comprehension)

Appendix A

Distribution of TOEFL Candidates by Geographic Region and Native Country, from July 1984 to June 1986

Table 10. TOEFL Total and Section Score Means — Nonnative English-Speaking Examinees Classified by Geographic Region and Native Country*

(Based on 714,731 students seeking admission to institutions in the United States who took TOEFL from July 1984 through June 1986)†

Geographic Region and Native Country	Number of Cases	Listening Comprehension	Structure and Written Expression	Vocabulary and Reading Comprehension	Total Score Mean	Total Number of Cases January 1978 — June 1986
AFRICA						
Algeria	819	51	52	51	513	3,998
Angola	16	50	52	52	513	78
Benin (Dahomey)	40	48	50	49	494	137
Botswana	102	55	58	55	562	583
Burundi	68	49	53	52	513	183
Cameroon	713	50	54	51	519	2,182
Cape Verde Islands	51	52	50	46	501	189
Central African Republic	23	46	47	48	453	79
Chad	9	·	·	·	·	37
Comoros	3	·	·	·	·	4
Congo	32	49	52	50	501	104
Djibouti	59	46	46	47	469	141
Egypt	6,544	49	51	49	491	25,530
Equatorial Guinea	6	·	·	·	·	17
Ethiopia	2,473	50	51	50	501	7,289
Gabon	36	49	50	48	489	126
Gambia	190	51	55	52	523	530
Ghana	1,470	54	60	57	573	7,894
Guinea	165	44	45	45	450	358
Guinea-Bissau	33	47	45	44	454	43
Ivory Coast	273	49	50	50	495	1,545
Kenya	1,338	56	59	58	571	6,011
Lesotho	66	55	58	55	558	219
Liberia	154	51	54	50	516	1,073
Libya	449	52	48	46	489	7,079
Madagascar	79	52	54	54	534	218
Malawi	84	53	60	58	569	302
Mali	208	48	52	51	501	553
Mauritania	80	45	45	46	453	108
Morocco	1,544	49	50	48	491	4,681
Mozambique	17	49	50	49	495	84
Niger	67	48	52	49	494	195
Nigeria	6,140	50	54	52	519	58,773
Rwanda	56	48	53	52	510	132
Sao Tome and Principe	9	·	·	·	·	20
Senegal	283	49	52	51	508	747
Seychelles	21	59	58	56	576	43
Sierra Leone	229	51	57	54	540	1,073
Somalia	864	49	48	48	487	2,187
South Africa, Republic of	503	57	59	57	578	1,636
Sudan	812	48	49	47	482	3,890
Swaziland	216	55	59	55	565	411
Tanzania	612	53	56	54	545	2,324
Togo	108	48	51	50	495	281
Tunisia	1,296	49	52	50	503	3,169
Uganda	244	55	59	56	568	1,153
Upper Volta (Burkina Faso)	44	49	52	51	507	143
Zaire	348	48	50	49	494	1,172
Zambia	249	58	60	57	583	1,079
Zimbabwe	254	56	60	57	579	902
AMERICAS						
Antigua and Barbuda	1	·	·	·	·	1
Argentina	2,173	55	56	57	557	7,301
Bahamas	7	·	·	·	·	29
Barbados	0	·	·	·	·	12
Belize	4	·	·	·	·	29
Bermuda	2	·	·	·	·	8
Bolivia	782	54	52	52	525	4,245
Brazil	4,424	52	52	54	528	17,746
Canada	1,438	57	54	55	554	4,446
Cayman Islands	0	·	·	·	·	1
Chile	1,473	53	52	55	533	6,107
Colombia	4,712	52	51	53	521	19,679
Costa Rica	825	56	54	55	549	3,077
Cuba	754	52	52	55	529	2,723
Dominica	1	·	·	·	·	1
Dominican Republic	944	53	51	53	523	3,113
Ecuador	1,164	53	50	52	520	4,812
El Salvador	1,009	54	51	53	525	4,999
Grenada	0	·	·	·	·	4
Guatemala	993	55	53	54	538	3,543
Guyana	8	·	·	·	·	134
Haiti	1,415	51	51	50	508	5,784
Honduras	820	55	53	54	535	3,311
Jamaica	6	·	·	·	·	140
Mexico	5,814	54	52	54	535	27,381
Netherlands Antilles	515	58	55	54	550	2,047
Nicaragua	802	53	52	54	530	2,716
Panama	1,785	53	50	51	515	8,059
Paraguay	131	54	53	54	536	507
Peru	2,630	53	51	53	522	10,083
Puerto Rico	3,195	53	52	53	526	9,727
Saint Vincent and the Grenadines	2	·	·	·	·	2
Suriname	188	57	55	52	548	808
Trinidad and Tobago	15	60	60	56	586	129
United States	1,209	58	56	55	563	4,377
Uruguay	283	56	57	58	570	983
Venezuela	3,172	52	50	52	512	47,380
Virgin Islands	2	·	·	·	·	10
West Indies Associated States	5	·	·	·	·	30
ASIA						
Afghanistan	490	52	48	48	496	1,243
Bangladesh	3,468	48	48	50	482	15,749
Bhutan	55	53	54	54	537	150
Brunei	203	56	54	54	543	1,048
Burma	593	52	51	51	512	2,019
China, People's Republic of	19,219	50	53	51	515	52,515
Hong Kong	64,417	52	50	51	510	263,386
India	32,021	55	58	58	573	107,237
Indonesia	22,499	50	47	48	484	59,845
Japan	62,659	49	50	49	496	195,501
Kampuchea (Cambodia)	427	50	46	46	477	1,062
Kiribati	6	·	·	·	·	23
Korea	64,030	48	51	52	505	176,989
Laos	258	51	45	46	473	876
Macao	1,466	50	50	50	500	5,220
Malaysia	41,451	53	52	52	525	133,488
Mauritius	213	56	59	57	575	640
Mongolia	26	49	48	49	486	83
Nepal	746	52	55	55	542	2,107
Pakistan	14,415	51	51	51	507	39,771
Philippines	8,020	58	58	57	578	36,860
Singapore	5,491	58	57	57	571	22,641
Sri Lanka	3,698	52	53	53	528	9,644
Taiwan	88,401	49	50	50	498	291,328
Thailand	22,471	48	48	48	480	80,868
Vietnam	6,751	51	50	50	503	26,399
EUROPE						
Albania	5	·	·	·	·	26
Andorra	7	·	·	·	·	27
Austria	508	59	58	57	584	1,655
Belgium	1,232	58	59	58	583	4,664
Bulgaria	69	55	55	55	551	229
Cyprus	4,060	52	51	48	501	19,129
Czechoslovakia	346	56	54	55	552	1,064
Denmark	812	61	58	56	584	2,733
Finland	730	60	57	56	576	2,505
France	9,578	54	56	56	554	27,085
German Democratic Republic	298	59	56	55	565	605
Germany, Federal Republic of	6,750	59	59	58	[illegible]	21,389
Great Britain	137	58	56	56	565	789
Greece	8,807	54	53	51	524	40,893
Greenland	6	·	·	·	·	24
Hungary	314	56	54	55	548	942
Iceland	701	60	55	54	567	2,256
Ireland	9	·	·	·	·	70
Italy	3,000	55	57	58	564	10,088
Liechtenstein	15	59	57	57	579	[illegible]
Luxembourg	67	55	60	59	593	256
Madeira	3	·	·	·	·	9
Maldives	13	·	·	·	·	30
Malta	49	63	64	62	630	150
Monaco	29	53	53	54	531	77
Netherlands	1,618	62	60	59	603	6,576
Norway	2,606	59	56	54	564	9,359
Poland	2,121	55	53	53	535	5,956
Portugal	504	57	55	55	559	1,934
Romania	734	55	54	55	549	2,277
San Marino	1	·	·	·	·	2
Spain	3,031	55	55	56	553	8,544
Sweden	1,553	61	58	57	587	5,966
Switzerland	1,920	56	57	55	570	6,216
Turkey	5,168	52	51	50	513	17,958
Union of Soviet Socialist Republics	1,095	55	53	53	539	8,452
Vatican	5	·	·	·	·	24
Yugoslavia	857	56	54	54	545	2,675
MIDDLE EAST						
Bahrain	829	51	46	44	469	3,[illegible]
Iran	11,586	52	49	47	495	65,926
Iraq	1,639	50	47	45	479	8,578
Israel	5,028	56	52	51	529	15,[illegible]79
Jordan	10,406	50	47	45	472	49,466
Kuwait	4,943	50	44	43	455	16,510
Lebanon	10,919	52	50	48	499	39,162
Oman	1,346	50	45	43	460	2,849
Qatar	709	49	43	42	444	2,482
Saudi Arabia	10,300	48	45	44	457	42,310
Syria	3,591	51	48	43	483	13,856
United Arab Emirates	1,920	49	43	42	445	5,830
Yemen	1,458	48	46	44	460	4,020
PACIFIC REGION						
American Samoa	778	50	46	44	464	5,313
Australia	56	54	50	48	506	427
Caroline Islands	815	48	45	44	457	2,973
Fiji Islands	198	56	55	53	545	721
Guam	2	·	·	·	·	13
Mariana Islands	53	54	50	48	507	220
Marshall Islands	187	48	44	43	447	721
Nauru	8	·	·	·	·	73
New Zealand	5	·	·	·	·	25
Papua New Guinea	24	55	56	54	547	82
Solomon Islands	3	·	·	·	·	19
Tahiti	14	·	·	·	·	14
Tonga	54	49	47	45	468	466
Tuvalu	21	53	53	52	524	21
Western Samoa	6	·	·	·	·	85

*Because of the unreliability of statistics based on small samples, means are not reported for subgroups of less than 15 for a total of 171 examinees.

†Includes 32,9[illegible]0 students who did not report their country of birth or who reported English as their native language.

Appendix B

Memorandum of Instructions Given to Item Raters in Pilot Study, Dated December 3, 1985

Memorandum for:

Subject: Classification of TOEFL items for a research project

Date: December 3, 1985

From: William H. Angoff

I am at present writing a proposal for consideration by the TOEFL Research Committee to determine whether items that have American (U.S.) content, or refer to U.S. people, places, institutions, etc. are in any way disadvantageous to those foreign candidates who have never lived in the U.S. Accordingly, I am planning to do an item discrepancy (item bias) study to determine whether such items are in fact more difficult for candidates tested in foreign countries as compared with candidates, matched for ability and nationality, but tested domestically.

The Research Committee is also interested to know the reliability of the process of classifying items into those that have reference to American content vs those that do not have such reference. I have selected a small group of test development specialists, of whom you are one, in the hope that you will all be willing to serve as raters for this reliability study. I am attaching a TOEFL testbook, Form 3HATF5 (disclosed) No. . I would appreciate it if you would take the time to go through the 150 items of this test form, one by one, and classify them as (1) referring to some aspect of Americana—e.g., the name of an American person or geographical region, place, or natural phenomenon (e.g., river, mountain range), a characteristically American institution, concept, custom, etc.; (2) not referring to someone or something of the sort described in (1); or (3) an item that is ambiguous in this regard. Additionally, if you feel that this whole matter is one of degree, would you also rate the items on a scale of 1 to 5 (5-highly American, 1-not American)? If you feel that the items cannot be so rated, but that they simply contain reference to Americana or that they do not, would you give the "Americana" items a checkmark ( ), the non-Americana items an x-mark (x), and the ambiguous items a mark of (0)? Make your notations on a separate sheet of paper and write your name on that sheet. May I collect your rating sheets on the morning of December 10?

Thank you very much for your help. The project-job for this work is 579-44.

Appendix C

Memorandum of Instructions Given to Item Raters in Formal Study, Dated September 28, 1987

Memorandum for:

Subject: Classification of TOEFL Items for a Research Project: Context Bias on TOEFL

Date: September 28, 1987

From: William H. Angoff

You may recall that about two years ago, in connection with a pilot study I was doing at the time (December 1985), I asked you to classify the items in a form of TOEFL with respect to their American content. You were being asked to judge whether or not a given item made reference to American individuals or groups of individuals, American places, things, institutions, customs, etc.. The pilot study was carried out in preparing a proposal I was then writing to determine whether or not items that contained such references were disadvantageous to TOEFL candidates who were tested in foreign countries, as compared with those tested domestically. The proposal was accepted. I am now engaged in the study itself, and I need to go through the same exercise of having the items rated and studied for the reliability of the rating process. As I did two years ago, I have selected a small group of test development specialists, of whom you are one, in the hope that you will all be willing to serve as raters for this reliability study. Please call and let me know (Ext. 1551), and if you are, I will ask Test Files to send you a numbered testbook, which you will be asked to return to Test Files after you have made your ratings.

If you agree to participate in this rating process, I would ask you to go through the 150 items of this test form, one by one, and classify them as (1) referring to some aspect of Americana, or (2) not. Having profited from the results of the earlier rating exercise, I believe I can now give you some detailed guidelines for making the judgments I am asking you to make. They are as follows:

An item is to be classified as "Referring to Americana" only if the context has a clear American reference; that is,

(1) a reference to a person who is known by you to be an American (e.g., Jonas Salk, Amelia Earhart, Benjamin Franklin);

(2) a reference to an American group of persons (e.g., the Amish people, the Huron nation, Nisei);

(3) a reference to an American place or region (e.g., the Alamo, Boston, the Appalachian Mountains, the Midwest);

(4) a reference to an American institution (e.g., Congress, the State of the State address, the party whip);

(5) a reference to an American event (e.g., the Civil War, the landing of the Pilgrims, the Reconstruction Period);

(6) a reference to an American, or typically American, custom (e.g., distribution requirements in college, rodeos, Valentine's Day);

American names for institutions, concepts, people, etc. that are not solely American (e.g., drug store, movies, volleyball, two-party system, the Mormon Church, string quartets, common law) should not be classified as American; similarly for fictitious names that are known and used not only in the U.S. but in other English-speaking countries (Dick, Bill, Margaret, Louise, William Long, Jane Ewell).

Only when the rater knows, or strongly believes, that a referent is American should the rater classify it so.

If a reading comprehension passage contains American referents, the rater should classify all the items based on the passage as American, even those items that do not themselves contain an American referent.

The rater should confine his (her) ratings to the referent and the context of the item. The rater should not try to judge whether or not the item will be biased in favor of examinees who have knowledge of Americana.

The rater should not attempt to judge whether a word used in the item has cognates in other languages and for that reason, or other reasons, is thought to be easier, or harder, for speakers of those languages than for others. Judgments of difficulty are not what is being sought here. The issue is simply and solely whether the context contained in the item is American or not.

You will be asked to submit your ratings within two weeks after you receive the testbook. May I hear from you within the next day or two, telling me whether you will be able to participate in the project?

Thank you very much.

Appendix D

Display of Raters' Assignments

of Americana to Each Item of TOEFL

Ratings of Americana for Section I of TOEFL

	Raters					
Item	Z	L	R	C	A	Sum
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
25	0	0	0	0	0	0
26	1	0	0	0	0	1
27	0	0	0	0	0	0
28	0	0	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	0	0
31	0	0	0	0	0	0
32	0	0	0	0	1	1
33	0	0	0	0	0	0
34	0	0	0	0	0	0
35	0	0	0	0	0	0
36	0	0	0	0	0	0
37	0	0	0	0	0	0
38	0	0	0	0	0	0
39	0	0	0	0	0	0
40	0	0	0	0	0	0
41	0	0	0	0	0	0
42	0	0	0	0	0	0
43	0	0	0	0	0	0
44	0	0	0	0	0	0
45	0	0	0	0	0	0
46	0	0	0	0	0	0
47	0	0	0	0	0	0
48	1	1	1	1	1	5
49	1	1	1	1	1	5
50	1	1	1	1	1	5

Code: 1 = Item contains a reference to Americana.
0 = Item does not contain a reference to Americana.

Ratings of Americana for Section II of TOEFL

			Raters			
Item	Z	L	R	C	A	Sum
1	0	1	0	0	1	2
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	1	0	0	1	2
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0
12	1	1	1	1	1	5
13	0	0	0	0	0	0
14	0	0	0	0	0	0
*15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	1	1	1	1	1	5
18	1	1	1	1	1	5
19	1	1	1	1	1	5
20	0	0	0	0	0	0
21	1	1	1	1	1	5
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	1	1	1	1	1	5
25	0	0	0	0	0	0
26	1	1	1	1	1	5
27	0	1	0	0	0	1
28	0	0	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	0	0
31	0	0	0	0	0	0
32	1	1	1	1	1	5
33	0	0	0	0	0	0
34	0	0	0	0	0	0
35	0	0	0	0	0	0
36	0	0	0	0	0	0
37	0	0	0	0	0	0
38	0	1	0	0	0	1
39	0	0	0	0	0	0
*40	0	0	0	0	0	0

*Not counted in the score.

Code: 1 = Item contains a reference to Americana.
0 = Item does not contain a reference to Americana.

Ratings of Americana for Section III of TOEFL

	Raters							Raters					
Item	Z	L	R	C	A	Sum	Item	Z	L	R	C	A	Sum
1	0	0	0	0	0	0	31	1	1	1	1	1	5
2	0	0	0	0	0	0	32	1	1	1	1	1	5
3	0	0	0	0	0	0	33	1	1	1	1	1	5
4	0	1	1	1	1	4	34	1	1	1	1	1	5
5	1	1	1	1	1	5	35	1	1	1	1	1	5
6	0	0	0	0	0	0	36	0	1	0	1	0	2
7	0	0	0	0	0	0	37	0	1	0	1	0	2
8	1	1	1	1	0	4	38	0	1	0	1	0	2
9	0	0	0	0	0	0	39	0	1	0	1	0	2
10	0	0	0	0	0	0	40	0	0	0	1	0	1
11	1	1	1	1	1	5	41	0	1	0	1	0	2
12	0	0	0	0	0	0	42	0	0	0	0	0	0
13	1	1	1	1	1	5	43	0	0	0	0	0	0
14	0	0	0	0	0	0	44	0	0	0	0	0	0
15	0	0	0	0	0	0	45	0	0	0	0	0	0
16	1	1	1	1	1	5	46	1	1	0	1	0	3
17	0	0	0	0	0	0	47	0	0	0	0	0	0
18	0	0	0	0	0	0	48	0	0	0	0	0	0
19	1	1	1	1	1	5	49	0	0	0	0	0	0
20	0	0	0	0	0	0	50	0	0	0	0	0	0
21	0	0	0	0	0	0	51	0	0	0	0	0	0
22	0	0	0	0	0	0	52	0	0	0	0	0	0
23	0	0	0	0	0	0	53	0	0	0	0	0	0
24	0	1	0	0	1	2	54	0	0	0	0	0	0
25	0	0	0	0	0	0	55	0	0	0	0	0	0
26	0	0	0	0	0	0	56	1	1	1	1	1	5
27	0	0	0	0	0	0	57	1	1	1	1	1	5
28	0	0	0	0	0	0	58	1	1	1	1	1	5
29	0	0	0	0	0	0	59	1	1	1	1	1	5
*30	0	0	0	0	0	0	*60	1	1	1	1	1	5

*Not counted in the score

Code: 1 = Item contains a reference to Americana.
0 = Item does not contain a reference to Americana.

TOEFL Research Reports currently available...

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language.* John L. D. Clark. November 1977.

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.* Lewis W. Pike. June 1979.

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests.* Paul J. Angelis, Spencer S. Swinton, and William R. Cowell. October 1979.

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context.* John L. D. Clark and Spencer S. Swinton. October 1979.

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.* Donald E. Powers. December 1980.

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Spencer S. Swinton and Donald E. Powers. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings.* John L. D. Clark and Spencer S. Swinton. December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance* Gordon A. Hale, Paul J. Angelis, and Lawrence A. Thibodeau. December 1980.

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language.* Donald L. Alderman and Paul W. Holland. August 1981.

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude.* Donald L. Alderman. November 1981.

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.* Kenneth M. Wilson. July 1982.

Report 12. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL.* Kenneth M. Wilson. July 1982.

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions: Validation and Standard Setting.* Donald E. Powers and Charles W. Stansfield. January 1983.

Report 14. *A Manual for Assessing Language Growth in Instructional Settings.* Spencer S. Swinton. February 1983.

Report 15. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students.* Brent Bridgeman and Sybil Carlson. September 1983.

Report 16. *Summaries of Sudies Involving the Test of English as a Foreign Language, 1963-1982.* Gordon A. Hale, Charles W. Stansfield, and Richard P. Duran. February 1984.

Report 17. *TOEFL from a Communicative Viewpoint on Language Proficiency: A Working Paper.* Richard P. Duran, Michael Canale, Joyce Penfield, Charles W. Stansfield, and Judith E. Liskin-Gasparro. February 1985.

Report 18. *A Preliminary Study of Raters for the Test of Spoken English.* Isaac I. Bejar. February 1985.

Report 19. *Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English.* Sybil B. Carlson, Brent Bridgeman, Roberta Camp, and Janet Waanders. August 1985.

Report 20. *A Survey of Academic Demands Related to Listening Skills.* Donald E. Powers. December 1985.

Report 21. *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference.* Charles W. Stansfield. May 1986.

Report 22. *Patterns of Test Taking and Score Change for Examinees Who Repeat the Test of English as a Foreign Language.* Kenneth M. Wilson. January 1987.

Report 23. *Development of Cloze-Elide Tests of English as a Second Language.* Winton Manning. April 1987.

Report 24. *A Study of the Effects of Item Option Rearrangement on the Listening Comprehension Section of the Test of English as a Foreign Language.* Marna Golub-Smith. August 1987.

Report 25. *The Interaction of Student Major-Field Group and Text Content in TOEFL Reading Comprehension.* Gordon A. Hale. January 1988.

Report 26. *Multiple-Choice Cloze Items and the Test of English as a Foreign Language.* Gordon A. Hale, Charles W. Stansfield, Donald A. Rock, Marilyn M. Hicks, Frances A. Butler, and John W. Oller, Jr. March 1988.

Report 27. *Native Language, English Proficiency, and the Structure of the Test of English as a Foreign Language.* Philip K. Oltman, Lawrence J. Stricker, and Thomas Barrows. July 1988.

Report 28. *Latent Structure Analysis of the Test of English as a Foreign Language.* Robert F. Boldt. November 1988.

00104 • S128P2 • 275630 • Printed in U S A