DOCUMENT RESUME

ED 365 108 FL 021 655

AUTHOR Barnwell, David Patrick
TITLE Problems of Articulation and Testing: Lessons from the 1920s.
PUB DATE Nov 93
NOTE 44p.; Paper presented at a Conference on Research Problems in Adult Language Learning (Ohio State University, Columbus, OH, November 1993).
PUB TYPE Information Analyses (070) -- Historical Materials (060) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Articulation (Education); Educational History; *Language Tests; Second Language Instruction; *Second Language Learning; *Standardized Tests; Statewide Planning; *Test Construction; *Testing Problems; Testing Programs; Test Reliability; Test Validity

ABSTRACT

Language testing historians have tended to ignore a significant period in the evolution of language tests, the years 1883-1929. In the earliest years, testing focused on knowledge about, not of, the language and reflected the teaching of Latin and Greek more than that of living languages. Grammatical formalism and translation were emphasized, and standardized testing was in its infancy. While the first World War slowed the pace of educational research, military needs gave impetus to standardization. In the 1920s, the quality of educational research increased. The war and public s' ›ol enrollments altered the pattern of foreign language study in the United States, motivating more screening and aptitude testing and increased standardization in vocabulary, translation, reading, and listening assessment. A major project in New York involved large-scale testing of secondary school students of French and redesign of a statewide examination. Results indicated significant problems in articulation and consistency of student progress. This and related testing research mark an important turning point in foreign language testing. The project produced 16 standardized foreign language tests of French, German, Spanish, and Italian. During the same period, standardized testing of speech skills was attempted. This era, largely ignored, should be acknowledged for its contributions. (MSE)

PROBLEMS OF ARTICULATION AND TESTING: LESSONS FROM THE 1920S

Paper read at Conference on Research Problems in Adult Language Learning, November 1993
Ohio State University

David Patrick Barnwell

PROBLEMS OF ARTICULATION AND TESTING:

LESSONS FROM THE 1920S

David Barnwell

PROBLEMS OF ARTICULATION AND TESTING: LESSONS FROM THE 1920S

> Nothing short of chaos prevails in the classification of our modern language students. The fact that a course is of a given length may have and usually does have little relation to the knowledge of the subject attained by a given class.
> (Algernon Coleman The Teaching of Modern Foreign Languages, 1929)

Bernard Spolsky's (1978) division of the history of language testing into three stages is well known. Spolsky posited a "pre-scientific or traditional" stage, a "psychometric-structuralist" stage, and (what was then) today's "psycholinguistic-sociolinguistic" stage. While later Spolsky came to realize that the question was more complicated than could be summed up in three neat eras, this tripartite division came to provide a handy label for those who wished to make quick generalizations about the history of foreign language testing. Unfortunately, the formulation was quite wrong, for Spolsky had missed an entire generation of language tests.

At the inaugural meeting of the MLA in 1883 the primary goals of language learning were established as literary and philological (Lodeman 1887). Numerous examples survive of the kind of foreign language testing which gave expression to these goals in the decades before the First World War. The College Boards, established in 1900, for many years essentially followed the format of combining some literary translation with the gauging of metalinguistic knowledge through asking students for explicit discussion of grammatical rules. They stressed overt grammatical

knowledge--the ability to verbalize rules and exceptions. The Boards' examinations often called for the translation of sentences so heavily "seeded" with particular grammar points as to be almost ludicrous or nonsensical. No opportunity was given to the student to create with the language, or even to integrate its disparate grammatical elements. Students were expected to know about the language rather than know the language. The format showed the influence of the teaching of Latin and Greek, and made no allowance for the fact that French and German were living languages. All instructions were given in English (MLA Report 1899).

While grammatical formalism was probably at its strongest in the teaching and testing of German, the position for French was little different. In 1917 Meras published an analysis of 178 French examinations set by American universities. The main testing technique used was translation from English, though in addition dictation, questions on readings, and free composition were employed. As in the case of German, considerable discrete grammatical points were tested--about 70% of the papers asked for explicit conjugation of verb tense paradigms, while "time and again" such things as noun-adjective concordance, possessive adjectives and pronouns were tested. Students were presented with such tasks as

DISCUSS FULLY THE POSITION OF ADJECTIVES

or

GIVE RULES FOR THE AGREEMENT OF PAST PARTICIPLES

or

MAKE A DIAGRAM SHOWING THE RELATIVE POSITION OF ALL PERSONAL PRONOUN OBJECTS WHEN STANDING BEFORE THE VERB.

This kind of foreign language testing was being carried out simultaneous to the beginnings of the great movement towards the measurement of human differences. In 1904 E.L Thorndike published the monumental "Introduction to the Theory of Mental and Social Measurements", perhaps the first book solely devoted to the topic of educational or psychological measurement. Binet's pioneer work with intelligence testing in Europe became known in the United States in 1910 when Goddard published an adaptation of the Binet test to American conditions. The more famous Stanford revision of the Binet Scale came some years later (Terman 1916). The first attempt to draw up a standardized test for a particular school subject comes in 1908 with Stone's test of arithmetic. This was followed by an increasing stream of tests and scales for the measurement of such diverse things as handwriting, drawing ability, spelling.

Though the First World War in some ways slowed the accelerating pace of educational research, in another way it gave impetus to the great movement in testing. The United States Army, faced with the need to classify its hundreds of thousands of recruits and place them in services where their talents could be used to best advantage, needed tests that could be administered to groups rather than individuals. The military turned to the work of psychological testers such as Otis (1918), and created the Army Alpha and Beta tests. In all, tests, not just of intelligence but of a wide range of vocational abilities, were administered to perhaps two million men.

POST WAR

Once peace was declared, the energy of educators and psychologists received free expression. The years after the World War witnessed an unprecedented expansion in the quantity of educational research. The number of doctoral degrees in education granted rose from 53 in 1918, through 94 in 1923 to 189 in 1927. The quantity of other research reports published in education rose in parallel, from 165 in 1918 to 333 in 1927 (Monroe 1928, 46). Research institutions began to spring up on all sides. The Iowa Child Research Station, set up as in 1917 at the University of Iowa, the Bureau of Educational Research at the University of Illinois, established in 1918, Teachers College Institute of Educational Research, established in 1922, all were to play a major role in the burgeoning research movement. More public bodies concerned with research were also set up. The American Council on Education was founded in 1918, the Research Department of the National Education Association being instituted four years later. Through bodies such as the Carnegie Foundation and the Russell Sage Foundation it was relatively easy to secure financial assistance for educational and research projects.

The War had effected a fundamental change in the pattern of foreign language study in the United States. German, for so long the most commonly taught language, in the public systems at any rate, fell into a chasm from which it was never to ascend. Before the War about 25% of all public school pupils had been taking German. In 1921-2 the figure was less than 1%. Which language gained through German's loss varied according to geography. Where French had always been strong, in the Northeast and South, and espe-ially

in the private schools, it picked up much of the slack. Elsewhere, especially in the West, and throughout the public schools, Spanish enjoyed dramatically increased enrollments. All of this came against a background of rapidly growing high school enrollments generally, as the hitherto semi-exclusive doors of the high schools were democratically thrown open to all. Millions of pupils were beginning to come into the expanded high school system. Schools were swamped by large increases in student numbers, and the student body was becoming much more heterogeneous than before. The educational system was thus faced with teaching large numbers of children who were dissimilar in background to those the schools had been used to. High schools were no longer seen as mere feeder institutions for elitist universities. Rather were they viewed as offering an education that was intrinsically worthwhile, especially when given the vastly increased numbers and variety of background of those in attendance. The new types of students now flocking through the high schools required a new kind of test.

Some argued that the influx of large numbers of students was damaging the prospects of the best students. Barlow (1926, 32) the president of the AATS, told his organization that

> we have in the early terms in high school a horde of low I.Qs. These low I.Qs get into our modern language classes together with their more gifted brethren. The pace has to be set for the latter. The weak ones cannot keep step, and they swell the percentage of failure at the end of the term, after having received attention that could more profitably have been given to their brighter comrades.

Barlow noted the efforts made in New York in the early 1920s to permit only the brightest pupils entering from elementary schools to take foreign language in the high school.

An interesting study on the place of languages in the San Francisco area high school system was carried out in the mid 1920s. George Rice of the University of California visited 22 modern language classes in 1925-6. He saw no evidence of any attempt by teachers to cater for individual ability differences among their students--what he called "elasticity". He further surveyed students to see how long they spent in foreign language home study. Generally the better students spent less time on study than the weaker ones. Rice argued that schools were failing in their duty to the brighter students in not demanding more of them. "With the ever-increasing inclusion in our high school population of students of inferior linguistic ability, the average capacity and average achievement is being continually lowered and the superior pupil is getting less education". Rice called for more grouping by ability level and greater use of prognosis tests. Where these were not possible he urged greater "elasticity" in teaching, that would require more from the bright students and not give the weak a sense of failure. Coleman (1929) echoed this: "It would also be highly advantageous both to pupils and to effective teaching in modern languages if school authorities would cooperate in grouping students on the basis of their previous scholastic record and of scores on intelligence tests, and if they would make it possible to reclassify or drop from the subject those who do not keep up with their classes, whether from incapacity or from other causes".

Faced with this reality many educators and administrators looked for ways of putting order on the incipient chaos. Some teachers instituted trial periods from two weeks to a year to select those who would and cull those who would not benefit from further

language study. Others recommended weeding out deficient students by examining their I.Q. scores and their scores in English. More formal aptitude tests--known then as prognosis tests--were published. One, the Barry Prognosis Test, used a little Spanish, on the very sensible theory that to find out how well a student will do on Spanish, you should see how well he does on some Spanish. Barry claimed to have calculated correlations of around 0.60 between scores on his test and teachers' marks in a subsequent Spanish course. Other commonly used early tests were the Iowa Foreign Language Placement Test, the Symonds Prognosis Test, the Luria-Orleans test, the Wilkins Test, the George Washington University test. All of these were oriented towards academic success in the language, so they tended to focus on testing ability to put grammatical rules into practice in particular exercises, and also on translating to/from the specimen language. None of the 1920 tests sought to measure a person's ability to discriminate foreign language sounds. They really were very like verbal intelligence tests, since they focused on the ability to decipher written texts. Word-based rather than discourse-based, they were oriented towards the most academically inclined students. They were used not so much to see which students had an aptitude for foreign language but rather which students did not. A later reviewer wrote "Foreign language prognosis tests of the Symonds and Luria-Orleans type are usually excellent means for reducing foreign language enrollments in nonfunctional courses taught by teachers incapable of adjusting either method or content to the needs, interests and abilities of children" (Kaulfers 1940 p. 1341).

Barlow, the AATS president, was especially shrewd in noticing a weakness of aptitude testing that has often gone undetected. He pointed out that a prognosis test provided no basis for predicting if a particular student would profit from foreign language study. In fact one could not speak about someone benefiting from foreign language study unless one could define the benefits of foreign language study. "The vast majority leave school with little book knowledge, and soon lose what they have. These people have, however, grown in various ways, and have developed certain attitudes of mind ... a prognosis test which does not take them into account, but merely indicates that one pupil is apt to be slower than another in his progress in language study is not satisfactory" (1926, 33). Fife (1930, 35) noticed that teachers surveyed were far more optimistic about their students reaching the <u>indirect</u> objectives of foreign language study rather than the direct--things such as increased command of English, a clearer understanding of the nature of language, and knowledge of the contributions to civilization of foreign peoples. Fife noted ironically "It is not possible to escape the inference that the teacher is more optimistic about the the success of his pupils in those fields which he cannot test, the fields of transfer values and of the formation of desirable intellectual and social habits, than in those fields in which he tests at regular intervals." Today, seven decades later, it seems that our profession has still made no effort to quantify the non-linguistic outcomes of foreign language education.

In 1925 there were approximately 20,000 public schools in the United States. About half of these offered one or more modern languages. These were generally the larger and more urban schools. About one in six taught no language at all. The remainder taught Latin or much more rarely Greek. While many schools offered Latin but no modern language, it was very rare to find a school that offered a modern language but no Latin. Latin was thus well ahead of all other languages, being almost as strong in terms of enrollment as all modern languages combined. This was especially so in the rural and smaller schools. There were about three quarters of a million enrollments in modern languages, in a total secondary enrollment of almost 3 millions. The actual number of students studying foreign language would have been less, since there were some duplicate enrollments. While it is hard to put a number on this, it is clear that only a minority of high school students--probably fewer than 20%--were enrolled in modern language classes. One of the elements that had brought this about was that in the Midwest, the traditional center of German study, no language had supplanted German in its post-War decline.

A survey taken in 1925 found that about 83% of students in the high schools (public + private) went no further than two years of study. 57% went no further than finishing their first year. Third year registration in the modern languages was about 16% of first year enrollment, while 4th year was a mere 2% of first year (Wheeler 1928). As Coleman put it (1929, 27) "In fact there appears to prevail in public school circles the judgment, implied if nowhere definitely formulated, that the normal course in a a

foreign language, ancient or modern, should last only two years if the student's general secondary education is to be provided for properly. This probably ... results from more recent tendencies in curriculum revision in behalf of the large groups of young people who formerly did not enter the secondary school". But another important factor was that of late starters--about one-fifths of beginners in language courses were already in the 11th grade, and had therefore at most two years in which "to develop the language power and to cultivate the attitudes by which the contribution of the subject to their education will be chiefly determined". Short exposure to foreign languages was institutionalized by the fact that there were a large number of schools which only offered two years of French or Spanish, and quite a few which only offered one year. Throughout the system, attrition rates, or "discontinuance" as it was called at the time, were for modern languages twice that for the schools as a whole. In fact, though the numbers of and in junior high schools were growing rapidly, teaching modern languages at junior school appeared to be a complete waste of time due to the problem of discontinuance.

This had a knock-on effect at the college level, where there was much "lost motion" in the transfer from high school. Here the strengths and weaknesses of high school enrollment patterns were to some extent reversed. At the college level by the 1920s Latin had lost its pre-eminent place. Proportionately more college students took foreign language than did at school high school. Beginning German was especially popular, since anti-German sentiment had caused it to dropped from most high schools since the War.

The acute problems facing the expanding educational system combined with the general impetus in educational research to heighten the interest in measurement, since a fundamental requirement for research is the ability to quantify its outcomes. Vivion Hermon of the University of Wisconsin in 1925 calculated that in the fifteen years since 1910 over 300 intelligence or educational tests had been produced. Monroe in 1927-8 estimated that not fewer than 30 million copies of standardized or semi-standardized tests were being used in the United States each year. It was not uncommon for sales of tests in the more popular areas such as mathematics or reading to approach a million copies. There were tests and scales for rating everything, from the quality of maintenance of school buildings to the performance of clergymen. The Army itself contributed to the testing boom, by dumping its vast supplies of psychological tests on the post-war civilian market.

Inevitably, foreign language testing participated in the great testing wave. The testing formats employed by such as the College Boards had never been universally popular with teachers. One writer declared himself opposed to the traditional type of test, especially for its stress on information about the language rather than assessment of use of the language itself. He argued that "the candidate should be tested in ability to apply French grammar in the construction of real French sentences, and not in ability to organize grammatical facts in valueless lists and synopses, or in meaningless rules and diagrams" (Meras 1917, 293). Heuser (1921) in describing College Board examinations in

German given during the pre-War years, expressed a commonly held attitude to formal grammar: "No amount of declining is worth anything, unless the respective cases can be used in sentences".

As early as 1916, a series of tests, including some in foreign language, was published by Charles Starch as part of his book "Educational Measurements". These tests are of a type hitherto not seen, quite dissimilar to those put out by the College Boards. The examination consisted of two parts, a vocabulary test and a translation test. Starch based his vocabulary test on a list of 100 words taken at random from a foreign language dictionary. In the case of German these were the first words of every 23rd page, the 23 being merely the frequency of selection needed in order to yield 100 words. Students were provided with two lists of 100 words, one in English and the other in the target language. They had to match each foreign word with its English "equivalent". Starch defended his method by pointing out that it provided a comprehensive random sample of vocabulary. He suggested that a score on his test would indicate the percentage of words in the entire foreign language vocabulary that a person knew. "If a pupil knows 25 words of each list it means that he knows 25% of the entire vocabulary" (p.175). But the very randomness of the word selection method confounded common sense; it meant that the entire dictionary range of target language frequency and register was encountered, with no attempt to graduate for likelihood of use in the spoken or even written language. Thus, in the French test, words such as CONDYLIEN COPHROPHAGE REGREFFER CHRYSOCALE EMMITOUFLER came up, along with such everyday words as AVOIR JETER BAS SCIENTIFIQUE. As one contemporary critic put it

"it tested a promiscuous vocabulary, which we do not seek to teach" (Handschin 1920, 220).

Starch's translation test consisted of 30 target language sentences, ranged in what he presumed to be increasing order of difficulty. He expressed no awareness of the grading problems that commonly arise when translation is evaluated. The translation was either entirely right or entirely wrong. Though Starch's formats were new he stuck to tradition in not differentiating between modern and ancient languages; the examination format for French and German followed that which he offered for Latin.

The first standardized modern language test was Charles Handschin's Silent Reading Test in French and Spanish (1919). In describing his test Handschin (1920) set out eight principles for the construction of a foreign language test, the most comprehensive exposition of testing theory drawn up to that date. His French test included a paragraph of 192 words with 10 questions, to be answered in either English or French, as well as a Comprehension and Grammar test based on completions and inflections. This test was followed by Vivion Henmon's French Test of Vocabulary and Sentence Translation in 1921. Henmon had earlier published a test for Latin, and his French test followed the same format. It consisted of sixty French words and twelve sentences, set in supposed order of difficulty. Students had to translate context-less French words and phrases to English. The entire test, though not explicitly speeded, was expected to take about twenty minutes. An attempt was made to control level of difficulty of the vocabulary, with all words used being taken from first-year

texts. The problem of rater reliability, however, was not addressed; no pointers as to scoring were given, except to score right or wrong without giving partial credit. Scoring could be carried out by a simple count or by using weighted scores on the basis of supposed difficulty levels. Though he obviously knew French, Henmon was not a language teacher--he was a testing specialist, and his tests often showed a lack of appreciation for what is important in foreign language. Henmon relied entirely on the translation of discrete context-less words and phrases. The parallel between Henmon's Latin test and his French test is significant, since it shows that no specific methodology for testing the vernacular aspects of the modern languages had yet been devised. Even though his testing was innovative and represented the most modern approach then current, Henmon did not create any structure for modern language testing that would have differentiated it from the testing of the ancient languages.

Despite such attitudes to testing, it would be fallacious to conclude that the profession, or at least its most self-conscious and innovative wings, was at this time uncaring about oral goals. Outside the academy, an interpretation test, in which the candidate interpreted from one to the other of two examiners had already been used for some years in Civil Service examinations (Lundeberg 1929, 196). Generally, though the earliest decades of this century were marked by skepticism about the feasibility of oral testing, many writers saw the value of oral ability, but only as one of several objectives of foreign language study. It should be stressed that there was for a long time a somewhat sloppy use of the term oral. Most contemporary

writers used the term "oral tests" to refer to what today would be called "auditory" or "listening" tests. There were a few of these in use at the university entrance level. Columbia University's "oral" (sic) test involved a dictation, together with written answers to auditory questions (Hayden 1920). Other universities such as Cornell and Princeton made use of formats such as this, though Princeton's also required a kind of translation, in which a student had to reproduce in English a passage he had heard in the foreign language (Decker 1925).

The mid-1920s saw major advances in the effort to put testing on a rational, even scientific basis. Commentators had for years pointed to inconsistencies in grading of the "old-type" examinations, whether arising through subjectivity of standards or mere carelessness. Ill-defined and capricious weighting systems as to relative difficulty levels or importance of particular topics caused standards to change. The small number of questions on the old-type tests made a misunderstanding of the question or ignorance of one topic all the more disastrous. It was charged that the essay-type question caused a great waste of time, for both examinee and examiner, in getting to the relevant points. The College Board's examinations in foreign language always had had a significant testing unreliability problem. An examination of passing rates on the nine College Boards subject examinations between 1910 and 1919 shows that French is in second place, with a rate of 62%. German is in second last place, with a passing rate of 50%. Standards oscillated from year to year; in 1916 73% passed intermediate French, in 1917 only 42%. In 1926 53% passed the

French examination, while this rose the following year to 82% (Kandel 1936, 52). It is hard to account for these disparities except to conjecture that great importance must have been attached to the choice of translation topics and vocabulary. As Robert (1926) showed, good performance on the translation was in itself almost sufficient to ensure a pass.

Against these were now set the "new-type" examinations. These basically exemplified the application to the classroom of the formats that had been created for standardized tests such as those of intelligence. Translation and composition were eschewed, in favor of discrete chunks of language--individual words or short sentences. The true/false scoring method was most popular initially, but as the 1920s wore on, multiple-choice, matching, completion, correct the error, rearrangement and other formats were added to the tester's repertoire. The required response was in all cases brief: a check mark, a number, a word, or at most a few words. Sampling was wide, since many topics could be introduced. The examination could be comprehensive in scope and in level of complexity, since perhaps fifty or more questions were asked where previously there had merely been a handful. The element of chance was thus diminished, with the promise of higher reliability levels. The distorting power of subjective factors, such as prejudices of examiner towards either form or content of answers, was greatly decreased by the mechanical nature of scoring. This ease and economy of grading for the "new-type" tests was of course offset by the much greater complexity involved in their preparation--in this respect they were the opposite of the old-type examinations. Rather than, for

instance, permitting examiners the vagueness of asking students to write an essay on any topic they wished--quite a common practice at this time--the "new-type" tests required careful and time-consuming elaboration. But in an era of rapid increase in numbers of students, and consequently of papers to be marked, the consideration of ease of scoring was predominant. Hence the new tendency in testing reflected the evolving structures in society at large.

Wood's New York Experiment

In the prevailing enthusiasm for testing in all subjects it was not uncommon for enormous numbers of students to be tested on the one day. One vast project in modern languages was undertaken in New York City (Wood 1927). Availing of funds from the Carnegie Foundation, the Board of Education of the city and the Regents of the State University of New York administered "new-type" tests to the city's high-school students. These tests were based to a significant extent on pilot work that had been carried on in placement testing at Columbia College in New York. Ben Wood, who was Associate Professor at Columbia College and director of its Educational Research Bureau, was the principal designer of the tests. Wood undertook to test all the modern language pupils in the junior high schools of New York city for two successive years, 1925 and 1926, and in the case of the high schools to test all those who took the 1925 Regents examination. One test, an experimental form of the Regents examination given at the senior level, was taken by over 31,000 foreign language students in 1925. The other test, that for junior high school students, was

administered in both 1925 and 1926. Some 19,000 students of French took this test each of the two years, about 6,500 taking the Spanish test in 1925 and about 4,000 in 1926. The format of the new-type tests made it for the fist time possible to compare achievements between schools and between classes in the same school.

Contrary to traditional practice, the tests were not fine tuned by year of study. The junior test was given to junior high school students, regardless of whether they were in first or second year of study. The senior examination was given to students at any point in the Regents cycle. Each test took 90 minutes. The junior tests, given in French and Spanish, followed almost identical formats in the two languages. They were composed of three parts. Part I consisted of 100 target language words, each followed by five English words. The words were chosen, as so often, from the most common words in contemporary word lists. The student had to choose the "equivalent" English words for the foreign.

French: CHIEN
1. CHIN 2. CHINESE 3. DOG 4. SHINE 5. CAT

Spanish: ZAPATO
1. LEATHER 2. SPADE 3. SHOE 4. CLUB 5. STRIKE

Part II was a Reading Comprehension test. 60 incomplete statements were given in the target language, with five alternative endings. The students' task was to pick which of the endings supplied was "coherent or true".

French:

S: ON SE SERT D'ENCRE POUR
R: 1. MANGER 2. BOIRE 3. COURIR 4. NAGER 5. ECRIRE

Spanish:
S: EL PERRO NORMAL TIENE
R: 1. UNA CABEZA 2. CINCO PIES 3. UN RELOJ 4. TRES OREJAS
5. CINCUENTA ANOS

Grammar was the focus for Part III, which was composed of 60 English sentences, each followed by an incomplete target language translation. The student was required to complete the translation.

French:
S: HER DRESS IS WHITE
R: SA ROBE --- ---.

Spanish:

S: I AM 12 YEARS OLD.
R: ------- DOCE ANOS.

Wood believed that these represented a fundamental improvement over old-style tests. They were the products of sustained work in pretesting and norming over a long period, and thus shared none of the "casualness" he associated with earlier approaches to testing. The vocabulary of the "new-style" tests, based as it was on word counts for elementary French and Spanish, was, he felt, more rationally selected than heretofore, and constituted a broad sample of the most common words in each language. The multiple-choice format was economical, in Wood's view, for it meant that particular likely problem areas could be addressed, with little time lost in "irrelevant activities, such as writing out translations of whole sentences or paragraphs". The new exams were comprehensive, proficiency-based rather than rooted in particular courses of instruction. They sought to measure outcomes, how much the student had learned, "regardless of whether he learned it from the teacher of that course, or in

spite of the teacher". And they took no account of "time-serving", the great bete noire of foreign language teachers even at this time, the practice of rewarding students for amount of time spent in class or number of courses they have taken, rather than giving credit for attained ability irrespective of time spent in attaining it.

One of the striking elements of the early literature on foreign language testing is the degree to which discussions overlooked the crucial relevance of grading and scoring decisions to the reliability and validity of tests. Wood, however, was unusual for the attention he devoted to problems of scoring. In the case of Parts 1 and 2 of the junior high school tests scoring was quite mechanical, there being only one acceptable answer. Part 3 was less so, but here directions to scorers were quite stringent: "In Part III absolute correctness must be rigorously enforced; no answer is to receive any credit if it is any respect--spelling, punctuation, capitalization, etc.,--deficient or incomplete." Reliability figures for the junior high-school test were high, the reliability coefficient (split-half) for the French test as a whole being .97, with similar figures for each constituent part.

Wood was a precursor of later foreign language testers in the fact that he had to face the lack of universally accepted external validating criteria for his tests. He believed that it was unnecessary to even argue for the validity of Parts I and III of the tests, since "vocabulary and grammar tests of this sort have been so thoroughly tried out and proved that no scientist or

teacher who has kept pace with recent developments can doubt their quality". (The empirical basis for the new tests was of course much more tenuous that might be gathered from such statements). As scores on the more innovative Part II correlated well with scores on Parts I and III, generally at a coefficient of around 0.80, Wood argued that this validated Part II. In discussing intercorrelations within a battery and using these as an argument for test validity, Wood showed a tendency to have it both ways, to read into the figures what he wanted them to show.

This was the first airing of a topic which has even today not been fully solved by foreign language testers, namely, the credence to be placed on correlations between tests. Wood's attitude towards statistics reappears again and again in the history of foreign language testing. The correlations, he wrote, were "high enough to vindicate Part II for general measurement purposes, and low enough to show that it is not a mere duplication of Parts I and III". In other words, separate components of a language test or battery should intercorrelate well, but not too well. For the Regents examination Wood showed external validity in the form of fairly high correlations between scores on the test and teachers' grades or scores on the "old-type" form. He did not address the logical inconsistency involved in validating a new and supposedly superior test against other measurements, often unreliable, which it sets out to supplant. However, Wood did put forward further evidence for the validity of the tests he employed. He showed that the great majority of questions discriminated well, being more likely to be answered correctly by good students than by

bad. Wood's statistical treatment was the first published use of detailed item-analysis in evaluating a foreign language test.

Wood also compared his "new-style" design of the French Regents examination with the "old-style". In the traditional format there was the customary large element of translation, to and from the target language, and of both paragraphs and particular difficult phrases, such as "four months ago" and "while reading his lesson". In addition, the French examination asked students to supply specific verb forms, both in context and out; further, they had to fill in the missing prepositions in sentences such as IL M'AIDE ___ FAIRE MES DEVOIRS. Finally--and this section had been added fairly recently--students were asked to write a 75 word composition on one of the following:

a/ LA FRANCE b/ LES FRANCAIS c/ LA LANGUE FRANCAISE
d/ L'ECOLE (le batiment, les matieres, les professeurs).

The "new-type" Regents examination designed by Wood had a lot in common with the test he gave to the junior high schools. Section II was different, however. In the new Regents examination, this was composed of 75 true/false statements in the target language "of an obvious truth or obvious fallacy ... easily within the knowledge of any high school student intelligent enough to study a foreign language". Most of these look unexceptional now, with items that appear to have met this criterion, though a few may have had a rather significant non-linguistic component:

EL GOBERNADOR DE UN ESTADO ES SIEMPRE LA PERSONA MAS INTELIGENTE DEL ESTADO.
LA ALHAMBRA ES UN CELEBRE PALACIO EN GRANADA.
PARA LEER BIEN HAY QUE APRENDER DE MEMORIA ALGUNAS POESIAS.

TOUS CEUX QUI HABITENT LA CAMPAGNE SONT DES PAYSANS.
ON NE VOIT PAS LES ETOILES QUAND IL Y A PLEINE LUNE.
LES MONUMENTS NE SERVENT A RIEN.

The practice of intercorrelating scores on the "new-type" tests with those on the old Regents format was tenable only if reliability figures on both were high. Yet no statistical treatment of scores on the "old-type" examinations existed. Wood therefore analyzed the scorer reliability of a number of old style Regents examinations from previous years. Of about a thousand scripts in each of French and Spanish, he calculated reliabilities between .5 and .78. This compared with reliabilities for the "new-type" examinations ranging between .8 and .9. The ratings for the old type Regents varied tremendously from school to school. As Wood pointed out, reliabilities for the "old-type" examinations were really unacceptable, and made any correlation between them and the "new-type" examinations almost meaningless. However, this did not stop him from citing these very correlations as evidence for the validity of his new-type tests. Wood did carry out intercorrelation calculations, and found that the "new-type" correlated with the "old-type" at values around .6. Wood considered this figure "satisfactory", and offered it as evidence for the validity of the "new-type" tests. This figure concealed some wide differences in patterns on the two type s of Regents examinations. Students failed the 'old-type' examination who did well on the 'new-type' and vice versa. Wood used this finding to refute the commonly held belief that new-type examination formats were able to measure the most elementary outcomes of learning, but were not suited to assessing more advanced levels.

What information was yielded from such a vast study as that undertaken in New York ? The fundamental finding was that

the standardization which the Regents examination was supposed to impose on the New York school system was totally absent. Efforts at testing and assessment, inasmuch as they existed, were not doing what needed to be done. Wood showed that a vast amount of "overlapping" or misplacement was in effect. Thousands of students were misplaced by at least a semester, being nearer in achievement to the average of the class above or below them than the class they were presently taking. Wood called this "the sacrificing of bright students on the altar of mediocrity" and bemoaned its concomitant "wasted energies and frayed nerves". Wide disparities were visible in the achievement of the 44 individual schools that took part in the junior high school study, and even within particular schools. There were violent fluctuations in this. A fourth semester class in one school showed average achievement little better than a third semester class in another school. A so-called third semester class in one school might be better termed fourth-semester, because the average of its scores met or exceeded the city-wide fourth semester average. THe reverse was equally applicable--a so-called third-semester class being actually a misnomer for a true second-semester class, based on city-wide averages.

Such disparities could also exist within the same school. Only about 40% of second semester students of French were closer to the average of their own class than to some other class average. The position was even worse for Spanish. Because of haphazard placement, Wood (1927, 27) observed that teachers did not "have <u>classes</u> to teach but heterogeneous aggregations of unhappy students"

No improvement as to homogeneity of levels of language instruction was noticeable between 1925 and 1926. If anything, there was deterioration all around; misplacement had become even more rife, particularly in Spanish, and those schools or classes scoring well in 1925 had in many cases deteriorated. The relation between first and second year outcomes for each particular class was, in Wood's words, "very nearly one of pure chance".

The senior level examined in 1925 proved out to be no better than the junior in terms of articulation and homogeneity. Indeed the New York Regents system, an unusually centralized one when compared to other states, had utterly failed to produce the standardization that was one of its goals. More than 60% of high school students were misplaced by at least a semester, approximately 30% of all French students were misplaced by a year or more. The situation was "very near to chaos as far as classification for instructional purposes and as far as educational guidance are concerned".

The tests devised by Wood and his colleagues were later published in the form of the Columbia Research Bureau and American Council Alpha and Beta tests. These tests were widely used and influential for perhaps two decades afterwards. Though his tests were entirely based on reading and writing, Wood was not oblivious of the need to measure speaking and auditory skills. He felt that the best way to measure these was "by means of conversations with students, one at a time, using carefully prepared sets of questions and conversational materials" (1927, 96). However, Wood felt that such tests, though valid, would be subjective; hence he

did not include them with his avowedly objective tests. He failed to see that all tests are subjective, in the sense that some person or persons has had to create them. What Wood meant by objective tests might be better termed "mechanical" or "easily scored". In fact, though Wood had a long subsequent career in testing he never attempted to produce speaking or listening tests. For several decades direct oral testing remained unattempted in the academic environment.

Though now forgotten, the work carried out in New York by Ben Wood and his colleagues remains monumental today, almost seven decades later. It was a major success for Wood--and an indication of the spirit of inquiry of the times--to have the Board of Regents, which had a venerable history going back to 1865, commit its resources to evaluating the claims of the new testing movement, even when they remained unconvinced that the new should supplant the old. The effort needed to administer the tests to the tens of thousands of students involved, and the statistical work involved in providing item analyses and measurements of central tendency for all the different tests could only with difficulty be replicated even in today's computer-assisted times. Wood's report was thoughtfully and forcefully argued and came buttressed with an array of statistical charts and tables. If he erred, it was in failing to reflect more profoundly on the problem of how the validity of language tests might be established. He did not speculate on the lack of fit between his measurement instrument and the kind of instruction pupils were actually receiving in foreign languages in New York, even though new-type tests were undoubtedly more divorced from contemporary classroom procedures than were the

old-type. All his arguments against the current situation were contingent upon his tests being valid. Wood never succeeded in demonstrating beyond question the validity of the sample that his tests constituted. It was a common criticism of the "new-type" tests that they did not tap into what was actually going on in foreign language classrooms. If Wood's test did not properly measure what students were actually being taught in the New York schools, then these tests were no better than the ones they supplanted, and perhaps worse, since the old type tests for better or worse did reflect many contemporary practices. Wood's tests were not at all as "objective" as he claimed, since they represented a rational sample chosen by him or by this colleagues. The kind of scoring systems that were compatible with an "objective" format precluded fine discrimination as to the quality of performance.

Wood carried out the business of testing within what he wished to be a rigorous and scientific methodology. He realized that learning goals are not abstract--they msut be concretely set within an educational context and a time frame. In the numbers of students tested Wood's New York experiments have still not been surpassed--a high percentage of the high school foreign language students in New York city participated in one or other of Wood's studies. For this contribution, as being one of the first to see testing as dependent on techniques and concepts drawn from the scientific method and subject to verification in the real world, Wood deserves to be remembered, perhaps even to be dubbed the father of foreign language testing in America. The strong

psychometric element in his work has ever since been an important part of the United States language testing tradition.

The Modern Foreign Language Study

Wood's research was but part of an even wider and more ambitious project. Between 1924 and 1927 a massive inquiry into the state of foreign language education was carried out in the United States and Canada. Working with funds provided by the Carnegie Foundation, the U.S. Modern Foreign Language Study Committee and the Canadian Committee on Foreign Languages implemented what even today remains the most comprehensive single survey ever carried out in this field. The initial goals of the committee were to study a wide range of issues including enrollment, achievement, methodology, and teacher training. As the Report of the Committee put it (Henmon 1929 vi.) it was in some aspects "an undertaking which may be said to be unique in the history of secondary education". For the first time ever, an effort was made to devise national norms for high schools and colleges, and to statistically compare results achieved by different methods of instruction.

The publications of the Study offer an invaluable snapshot of the state of the art in many areas of foreign language teaching in the mid-1920s. Vivion Henmon's 350-page Report provided a comprehensive articulation of what the most informed and advanced language teachers and testers considered to be the issues facing them in the 1920s. According to Henmon at least nine separate foreign language skills existed and needed to be tested. These were: vocabulary, reading with comprehension, translation

into English, translation into the foreign language, free composition, grammar, auditory comprehension, pronunciation and speaking. The Committee only succeeded in developing tests for four of these: vocabulary, reading, grammar and composition. A fifth category, listening comprehension, was the subject of research inspired by the Study but published a little later.

The Study produced sixteen standardized foreign language tests, more than all those previously available. Of the sixteen, nine were in French, the rest spread between German, Spanish and Italian. The American Council tests were the best known products of the MFLS efforts in testing. The first modern language tests to be standardized on significant numbers, they came in two forms, an Alpha form for upper high school and college students, created by Vivion Henmon of the University of Wisconsin in collaboration with Algernon Coleman of the University of Chicago. The Beta tests--those drawn up by Ben Wood of Teachers College, have already been discussed.

Those devising vocabulary tests for the Modern Foreign Language Study were faced with the problem of on what basis to select those words to be tested. A difficult problem at any time, it was made more acute by the many different texts and methodologies used throughout the country, ranging from a highly oral and colloquial focus to literary and grammar-translation. As Ben Wood put it "our textbooks are a veritable Tower of Babel" (1927, 98). Indeed, Wood's comparative word-count of sixteen elementary French textbooks had shown that only 134 words were common to all sixteen books, and that were a list of the 1000 most common words to be drawn up, they would be found in only 9 of the

16. It seemed that in many cases what was being drilled in the classroom was not what was of greatest importance in the language. Hence the Vocabulary Tests devised by the Modern Foreign Language Study were created on the basis of frequency counts and word lists. Formats for testing vocabulary were particularly context-impoverished, not even the article being provided with the French nouns. As in Wood's New York study, the test was multiple-choice. Given five English words, students were asked to mark which one was "a correct translation" of the French word. Two of the distractors were chosen to create "confusion"; two were chosen at random.

S: MAIS
R: HAND MORE BUT MONTH DAY

Henmon defended the rather bare look of these tests on the grounds that "time of administration and cost of printing" were important factors in mass testing of the kind being implemented. In answering possible objections to the lack of context of the test items, he offered evidence of a high correlation between scores on this type of "column" test and on those tests where words were contextualized. He also offered evidence that while scores on a "recognition" test were consistently higher than on a "recall" test, the correlations were so high that the former more economical format could safely be used. Here were early instances of a theme which runs through the history of foreign language testing, namely that of the weight that can be legitimately attached to correlations. If Test A correlates highly with scores on Test B, but is much cheaper and easier to run, why replace A with B, regardless of A's patent imperfections ? This issue also

emerged in the case of the Reading Test. In pilot work at the University of Iowa various formats were trialled, such as foreign language passages with questions in English, passages with questions in the foreign language, true/false answering, multiple-choice. The test designers preferred the paragraphs with multiple-choice format if only because it was the one most familiar to teachers, but for administrative reasons, the study found in favor of the True/False scoring method, in which students had to mark True or False to a series of statements in English about the foreign language prose reading. The principal rationale offered was that since the true/false answering method was easy to score and yielded scores that correlated highly with other more cumbersome scoring methods there was no reason not to use true/false. Henmon here perhaps betrayed the influence of his own background. He had no formation in foreign language, being trained in psychology. He was perhaps too concerned with mechanics and statistics, and insufficiently wary of the impoverished use of language called for by true/false responses.

The Henmon data were gathered from throughout the United States, being taken from tests administered to about 5000 secondary students of French, 3300 of German, and 4800 of Spanish. The general findings replicated those of Wood in New York. According to Henmon (p.146) "The cumulative evidence is very strong that 50% of the students tested are erroneously classified, and should be a semester or more above or below the classifications in which they are found. A similar analysis will show that 25% are erroneously classified by a whole year. The situation in the colleges is quite as bad".

Throughout its work the Modern Foreign Language Study was confronted with the problem of validation. How could test users be sure that the test was doing what it purported to do ? What evidence was there that the tests offered valid measurements of student ability ? Various attempts to confirm this were tried. Scores on tests were correlated with years of study of the foreign language. These correlations turned out to be low, though it could not be determined whether this was because of deficiencies in the tests themselves or because of the invalidity of the criterion. Other criteria, such as correlations with course grades or teachers' marks, were not too high, the best tests in this regard being those for Grammar. Intercorrelations--between, say, a Vocabulary test and a Grammar test, were fairly but not spectacularly high, generally in the .50 to .60 range. This tended to be offered as evidence for validity, though it also posed the classical dilemma of how to treat the constituent scores yielded by a battery of tests. If tests intercorrelated at high levels did this mean that they were measuring the same thing ? If measuring the same thing was either of them redundant ?

The MFLS produced few auditory tests. M.A. Buchanan of the University of Toronto did produce a Spanish Audition Test for the Study, It was originally hoped to put audition tests on phonograph, but this fell through, leaving the teacher to read a script. The test consisted of two sets of twenty-five questions. In the first set, the examiner read a one-word stimulus once. In Spanish, students had to mark the word corresponding to the word they heard.

S: AGUA
R: BEBER NOMBRE LUZ AYER

Twenty-five such stimuli were given, followed by a further twenty-five items of a somewhat more complicated nature. This was quite an imaginative format; it called for a rather high order of language use, was not based on the single-word level, and made no use of English. The second set followed the principle of answering by association, often creating quite complicated tasks.

S: CUANDO NOS CORTAMOS EL DEDO SALE SANGRE
R: CONOCIMIENTO HERIDA CAPITAL PIMIENTO

S: POCO A POCO SE VA LEJOS
R: FRANQUEZA ECO PACIENCIA MIRAR

The type of language processing required for answering of questions such as those in the second set was quite sophisticated, though to what extent non-linguistic elements entered is moot. The test was very much oriented towards vocabulary, but was noteworthy for its willingness to mix skills, to have answering dependent on a reading element as well as auditory comprehension. Experimental work on an auditory French test was undertaken by Agnes Rogers at Bryn Mawr University. Though the test was developed too late for the work of the Study, it was published in 1933 as the American Council French Aural Comprehension Test. In contrast to the Spanish test, the French testers were unwilling to test more than one thing at a time, to mix the auditory and reading recognition skills. Thus the examiner provided the spoken question, students marking an answer in English.

S: AVEC QUOI ECRIT-ON ?
R: BRUSH PEN PAPER KNIFE

Rogers reported satisfactory validation, using teachers' marks as criterion. A more important auditory test, though also produced too late for the work of the Study, was the Lundeberg-Tharp Audition-Pronunciation Test in French, initially reported in 1929. The auditory section consisted of three parts. In the first, on phonetic accuracy, the teacher/examiner read aloud a list that contained fifty sets of similar-sounding words or phrases.

S: IL SAIT TOUT R: IL S'EST TU IL CEDE TOUT IL SE TUE
IL SAIT TOUT

The sentences were produced in isolation, allowing no use of context on the part of the listener. The focus was thus entirely on sound rather than meaning. Comprehension for meaning was called for, however, in Part II, in which the tester read a series of twenty incomplete statements in French and students generated the missing words in English.

S: LE PERE DE MA MERE EST MON
R: GRANDFATHER

In Part III twenty definitions in French wee heard and students supplied the answers in English

S: L'EAU QUI TOMBE DU CIEL EN GOUTTES
R: RAIN

The 1920s saw the first effort to tackle the hitherto insoluble problem of standardized testing of spoken speech. This came with Aurelio Espinosa's Stanford Spanish Test (1927) and Jeanne Greenleaf's French Pronunciation Test (1929). The latter is especially interesting for the fact that it made use of the technology of the dictaphone, a resource that had been available for decades, but had not been incorporated into the tester's

armory. The student made a recording of his speech by reading a short passage of French into the dictaphone. This was subsequently scored for pronunciation. Greenleaf was vague on how this scoring was carried out. She did, however, report that average time of administration was two minutes, and that she unaided had to administer and grade about 400 of such tests every semester. Greenleaf's untimely death put a stop to this work, however. To appreciate the modernity of Greenleaf's test it is only necessary to contrast it other methods of testing pronunciation. As late as the mid-1920s the New York Regents examination offered written pronunciation tests such as

INDICATE THE MAIN DIFFICULTY OR PECULIARITY IN THE SOUND OF ONE CONSONANT IN EACH OF THE FOLLOWING WORDS OR PHRASES:
absurde ... soixante ... nom anglais ... cent un ... grand homme

Though innovative, Greenleaf's work was characteristic of her time in isolating an element such as pronunciation from its function within the entire expression. Oral composition was seen as parallelling written composition. Speech was seen as reading aloud or uttering one-way discourse. It was not seen as interactive.

Stimulated by the environment of inquiry fostered by the Modern Foreign Language Study, tests began to appear with somewhat bewildering frequency. For the first time ever, the topic of testing was assigned an important role in the agenda of meetings of professional organizations. A talk on modern language testing at a Conference of High School teachers at the University of Illinois in November 1926 "provoked enthusiasm and discussion" (Modern Language Journal, January 1927). Two years later, in November

1928, an entire session on testing was held at the meeting of the Illinois conference. The AATF annual meeting of 1927 offered a session titled "How shall we measure achievement in modern languages ?", probably the first such session ever held anywhere. The testing boom was not restricted to the production of tests for publishing. A large number of books and articles appeared on how to apply the methodology of the "new tests" to daily classroom use. There was even a nation-wide contest in 1929 in which participants constructed tests for French and Spanish. These were the years of the founding of important professional organizations, such as the AATG and the AATF, the first Summer Linguistics Institute (1928) the appearance of such journals as The French Review and the German Quarterly (both 1928). An atmosphere of optimism and dynamism prevailed in regard to the possibilities of a new "scientific" basis for foreign language teaching and testing. Researchers were anxious to get to grips with the diverse questions facing foreign language teachers, and, for a few years at least, many were confident that the new tests provided by the Modern Foreign Language Study would furnish an important tool for this purpose. In Henmon's words they were opening up "a field that would be carried on further".

However great the numbers of studies, however wide their sweep, the fact is that the vast majority of testing is done in semi-private, by class teachers. It would be prudent to look for resistance to the new-type tests at this level. Nevertheless, though one can speculate that many teachers were not convinced, published words of caution at this time were rare. Two

bibliographers were less than effusive in recording the plethora of "objective (perhaps one might even say mechanistic) measurement" (Welch and Van Horne 1928). Fife (1931, 99) wrote of the "formidable" opposition which the tests had met. He traced this to teachers' unwillingness to accept the statistical reasoning upon which standardization had been carried out and on the lack of fit between current teaching practice and the new tests. Yet in general Fife held that the reception of the new type tests by teachers had been "quite enthusiastic", and this seems accurate. The College Board examinations, however, persisted as purely reading and writing, with particular value being attached to translation to English. In one writer's phrase (Robert 1927) in these exams French was treated as "a dead language".

The Modern Foreign Language Study's great interest in testing was actually a by-product of its original goals. The Study's primary objectives were to gain data about achievement in foreign language, what helped or hindered it. It never truly got to the point of addressing this question, since it was from the beginning faced with the problem of how to measure the outcomes it wished to study. The various tests devised were seen as only constituting the first phase of the study, being no more than the required tools for carrying forward the real object of the investigation. Though devising measures that would be comparable between schools and across semesters was only to be a means to an end, the Study spent so much time on testing that it never reached its ends. The Modern Foreign Language Study experience thus showed the utter centrality of testing to educational research. Seven decades later, comparability of achievement in one school to

another, or one university to another, is not much higher than it was in the time of the MFLS. Those participating in the MFLS were not unwilling to gather data in the population as a whole by going out into the classrooms. They tried out their tests on large and unwieldy numbers of subjects. In this they still stand as an example to some of today's testers. The measurement instruments in use in the 1920s may not have been good tests in our eyes, but it is quite fallacious to label them as "pre-scientific". For the first time, those involved in teaching were consciously reflecting on the question of how to assess achievement, and in so doing were creating the very idea of testing as a discrete activity. Seeking to found their discipline on a scientific methodology, they incorporated all the then-available tools of statistics and technology. They faced many difficulties, some of which have been mentioned here.

As it turned out, however, little came of the great language testing boom of the 1920s. It was soon replaced by the Depression and the arid controversy over the Reading Method. Little of consequence in language testing was produced thereafter until the Second World War gave rise to dramatically new goals in testing. The efforts to reform foreign language teaching and testing in the 1920s largely failed. Leaving aside social and economic factors in the wider world, the great weakness of the testing pioneers of the 1920s lay in their failure to reflect on how the validity of language tests might be established. They never noticed the lack of fit between the new measurement instruments and the kind of instruction pupils were actually

receiving in foreign languages. Nor did they seek to develop any theory or construct of language which might have provided some coherence or logic in decisions as to which elements to test and how to test them. They were divorced from linguistics, though it is doubtful that linguistics as practiced at the time was in a position to supply a theory of language to justify particular testing formats.

Were most people called on to name some of the great figures in foreign language teaching/testing history it seems probable that few would put men like Vivion Henmon and Ben Wood on any short-list. It is doubtful that the Modern Foreign Language Study would be often cited as one of the seminal events in our profession. Yet efforts such as those sketched here represent the beginnings of the United States language testing tradition, and perhaps the fact that these men and their times are almost totally forgotten today reflects more the amnesia of the foreign language teaching profession than their true place in its history. Surely no other community has so little awareness of its tradition and story. It is perhaps this lack of awareness of the continuity between various epochs of language testing that caused Spolsky to make the mistake mentioned at the start of this paper, and permitted the error to go uncorrected for so long.

References

American and Canadian Committees on Modern Languages. 1930. Studies in modern language teaching. New York: MacMillan.

Barlow, William M. 1926. Address of the president. Hispania 9: 31-8.

Buchanan, M.A. 1927. American Council Alpha Spanish Test.

Yonkers, N.Y: World Book Company.

Coleman, Algernon. 1929. The teaching of modern languages. Chicago, University of Chicago.

Decker, W.C. 1925. Oral and aural tests as integral parts of the Regents examination. Modern Language Journal, 9: 369-74.

Fife, Robert. 1931. Summary of reports on the modern foreign languages. New York, Macmillan.

Greenleaf, Jeanne. 1929. French pronunciation tests. Modern Language Journal 13: 534-7.

Handschin, Charles. 1919. Handschin modern language tests. Yonkers, N.Y: World Book Company.

Handschin, Charles. 1920. Tests and measurements in modern language work. Modern Language Journal, 217-25.

Hayden, Phillip, M. 1920. Experience with oral examinations in modern languages. Modern Language Journal 5.

Henmon, Vivian. 1929. Prognosis tests in the modern foreign languages. New York, MacMillan.

Henmon, Vivion. 1930. Achievement tests in the modern foreign languages. New York, Macmillan.

Heuser, Frederick, 1921. Regents examination in German. Modern Language Journal 5.

Kandel, I.L. 1936. Examinations and their substitutes in the United States. Bulletin of the Carnegie Foundation, No.28.

Kaulfers, Walter. 1940. Review of foreign language prognosis tests. In Oscar Buros ed., Mental Measurements Yearbook, 1340-1.

Lodeman, A. 1887. The modern languages in university, college and secondary schools. Modern Language Notes 1: 97-109.

Lundeberg, Olav K. 1929. Recent developments in audition-speech tests. Modern Language Journal 14: 193-202.

Meras, A. 1917. French examinations. Modern Language Journal 1: 285-301.

Modern Language Association. 1899. Report of the committee of twelve of the Modern Language Association. In Proceedings and addresses of the 38th annual meeting of the National Education Association, N.E.A.

Monroe, W.S. 1928. Ten years of educational research 1918-27. University of Illinois Bulletin, Urbana.

Otis, 1918. An absolute point scale for the group measurement of intelligence. Journal of Educational Psychology 9: 239-61, 333-48.

Rice, George A. 1930. A study of achievement in foreign languages in junior and senior high school. In Studies, 435-71.

Robert, Osmond T. 1926. College Entrance examinations in French. Modern Language Journal 11: 17-24.

Rogers, Agnes, and Frances Clark. 1933. Report on Bryn Mawr test of ability to understand spoken French. Modern Language Journal, 17: 241-8.

Spolsky, Bernard. ed. 1978. Approaches to language testing. Arlington: Center for Applied Linguistics.

Starch, Charles 1916. Educational measurements. New York, MacMillan.

Stone, C.W. 1908. Arithmetical abilities and some factors determining them. Teachers College Contributions to Education, New York.

Terman, L. 1916. The measurement of intelligence. Houghton-Mifflin, New York.

Wood, Ben D. 1927. New York experiments with new-type modern language tests. New York, Macmillan.

-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-