# THE USE OF COMPOSITION SCALES FOR THE MEASUREMENT OF QUALITY IN ENGLISH COMPOSITION (1903-1963)

by

JON F. LOVE

B. A., Southwestern College, 1962

---

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

School of Education

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1966

Approved by:

Charles M. Reeves

Major Professor

TABLE OF CONTENTS

# CHAPTER I

## THE ORIGIN AND DEFINITION OF EDUCATIONAL SCALES

Thesis I. Whatever Exists At All, Exists In Some Amount.
Thesis II. Anything That Exists In An Amount Can Be Measured.
Thesis III. Measurement In Education Is In General The Same As Measurement In The Physical Sciences.[1]

With these three theses, set forth in 1918, E. L. Thorndike best expressed the position and opinion of the scientifically oriented educators who were attempting to apply the popular empirical methods of the early twentieth century to every aspect of human behavior. It is in such an atmosphere that we begin an examination of one such attempt to describe an highly complex human behavioral characteristic in terms of statistically expressed value judgements. This characteristic was the quality of an individual's ability at English composition and the attempt to describe it empirically is known as the composition scale movement in education.[2]

### I. DEFINITION OF AN EDUCATIONAL SCALE

A need now arises to establish an operative definition of the term "scale." In educational circles, a scale is taken to be:

> . . . a series of objective forms of exercises or definite samples of products of different quality which, by means of . . . statistical procedure, have been arranged in a definite order or position, usually in ascending order of

---

[1]E. L. Thorndike, The Use of Educational Tests and Measurements, Seventeenth Yearbook of the National Society for the Study of Education (Bloomington, Illinois: Public School Publishing Company, 1918), p. 16.

[2]G. M. Wilson and Kremer J. Hoke, How to Measure (New York: The Macmillan Company, 1921), pp. 1-4.

> difficulty or merit. In a scale, each exercise or group represents as much greater value or merit than an exercise or group just below it on the scale.[3]

It is with this definition in mind that we now view a short history of educational scales and their application to the judgement of the quality of composition in English themes.

## II. ORIGIN OF EDUCATIONAL SCALES

The origins of the scale movement were closely related to the development of standardized tests. The earliest American pioneer in this field which related education to psychology was J. M. Rice. In the year 1894, Rice first focused attention on the need to measure the achievement of pupils in subject matter areas in order to best determine the effectiveness of the school's approach to teaching in these areas. It was in this same year that Rice constructed a list of fifty words and a companion test in sentence form to determine the adequacy of drill in teaching students spelling skills. Thus the measurement movement was begun.[4]

The modern educational measurement movement, however, did not receive its impetus until the year 1904.[5] In this year appeared the first book devoted to statistical and scientific measurement

---

[3] Harry Andrew Greene and Albert N. Jorgensen, The Use and Interpretation of Elementary School Tests, (New York: Longmans, Green and Company, 1935), p. 15.

[4] Walter Scott Monroe, An Introduction to the Theory of Educational Measurements (New York: Houghton Mifflin Company, 1923), pp. 3-4.

[5] Charles Watters Odell, Educational Measurement in High School (New York: The Century Company, 1930), p. 32.

of human attributes, written by E. L. Thorndike. This book quickly initiated the objective measurement of achievement in subject matter areas and served as a unique guide for all students in the field.[6] Thorndike was long a leader in the application of statistical procedures to the materials of educational psychology and it is not surprising that, under his tutelage or influence, a number of the prominent leaders in the field of educational measurement were led to formulate such procedures into objective tests and, finally, scales.[7]

## III. HISTORY OF EDUCATIONAL SCALES

The scale movement began when Thorndike published his Scale for Handwriting of Children in 1909.[8] This scale was an attempt to gather a broad sample of the actual handwriting of students and others more competent in penmanship and to distribute these specimens on the basis of general merit as determined by a group of judges schooled in the field. After the initial distribution, a number of representative samples judged to be equally distant from each other in terms of general merit were arranged as a kind of "yardstick" for judging any and all other samples of handwriting

---

[6]E. L. Thorndike, An Introduction to the Theory of Mental and Social Measurements (New York: Teachers College, Columbia University, 1904).

[7]Odell, Op. cit., p. 34.

[8]E. L. Thorndike, "Handwriting," Teachers College Record, XI (March, 1910), pp. 1-3.

that might be compared with the scale.[9]

The pattern just described was to be copied and modified by the followers of Thorndike in later years. In 1909, the same year that Thorndike had constructed his handwriting scale, S. A. Courtis published a scale for judging the four fundamental operations of arithmetic. This measurement used the same scales in all grades from the third up through the elementary school.[10] This was to be followed shortly by Ayres' Scale for Measuring the Quality of Handwriting in Young People, a measurement quite similar in statistical derivation to the earlier comparable work by Thorndike.[11] Later, in 1913, Thorndike constructed a second type of scale concerned with the "general merit of children's drawings," again on the same pattern as his earlier work with handwriting.[12] Numerous scales by a variety of the members of the educational field were to follow the initial efforts of Thorndike and his colleagues, including the Buckingham Spelling Scale in 1913 and a second scale by Ayres con-

---

[9] Ibid.

[10] S. A. Courtis, Manual of Instructions for Giving and Scoring the Courtis Standard Tests in the Three R's (Detroit: Department of Cooperative Research, 1910).

[11] L. P. Ayres, Scale for Measuring the Quality of Handwriting of School Children (Russell Sage Foundation Bulletin, No. E-113. New York: Russell Sage Foundation, 1912).

[12] E. L. Thorndike, "A Scale for Measuring Achievement in Drawing," Teachers College Record, XIV (November, 1913).

cerning ability in spelling published in 1915.[13,14] All of these scales were similar since they attempted to measure the general merit of an individual's products in a given school subject by comparing the products with those listed at equal intervals of merit on a scale constructed of samples taken from the work of other students.[15] It is the judgement of English themes and their general quality that now leads to an examination of the events that surrounded the publication of the Hillegas Scale for the Measurement of Quality in English Composition.[16]

---

[13] B. R. Buckingham, Spelling Ability: Its Measurement and Distribution (Teachers College Contributions to Education, No. 59. New York: Teachers College, Columbia University, 1913).

[14] L. P. Ayres, A Measuring Scale for Ability in Spelling (Russell Sage Foundation Bulletin, No. E-139. New York: Russell Sage Foundation, 1915).

[15] Odell, Op. cit., p. 35.

[16] Milo B. Hillegas, "A Scale for the Measurement of Quality in English Compositions by Young People," Teachers College Record, XIII (September, 1912).

## CHAPTER II

## THE THEORETICAL BASIS FOR COMPOSITION SCALES

With the scale movement in education came a growing awareness that teacher grade judgements were highly subjective and far removed from any generally accepted standards of performance and quality.[1] It was such an atmosphere of subjectivity that accelerated the growth of scales and other means of making more objective the grading of school products in all subject matter areas.[2] It is in the area of English, however, that some of the more interesting developments of the scale movement occur and, more specifically, in the use of educational scales for the measurement of quality in English compositions.

As early as 1911, E. L. Thorndike had proposed a scale which would measure the Merit in English Writing by Young People. Thorndike suggested that an ideal scale for measuring merit in English writing ability would consist of a series of compositions ranging from zero to the greatest possible quality whose degree of merit were known and which, taken as a body, would be easily comparable with other compositions.[3] Under the tutelage of E. L. Thorndike,

---

[1] C. W. Stone, Arithmetical Abilities and Some Factors Contributing to Them (Teachers College Contributions to Education, No.19. New York: Teachers College, Columbia University, 1921), p. 86.

[2] Milo B. Hillegas, "Scale for the Measurement of Quality in English Composition by Young People," Teachers College Record, XIII (September, 1912), 332.

[3] E. L. Thorndike, "A Scale for Merit in English Writing by Young People, The Journal of Educational Psychology, II (May, 1911), 361.

Milo B. Hillegas was soon to produce just such a scale.[4]

## I. JUDGEMENT OF THEMES ON SINGLE GROUND OF MERIT

The publication of the Hillegas scale in 1912 seemed to offer new hope in the grading of English compositions.[5] It was believed that, with this new scale, English teachers would no longer have to rely upon their personal and subjective standards in marking a paper for errors in spelling, punctuation, grammar, syntax, and content. The individual teacher no longer need consider all of these various and variable components of a composition to arrive at a final grade. With the use of the new scale constructed by Hillegas, this would prove unnecessary, or so it was reasoned by the scale's proponents. The Hillegas scale was to be unique among the devices for the grading of English compositions since it was to denote the worth of writing on the bases of the single factor called "merit." The reasoning behind this approach and the method used to implement it are clearly outlined by Thomas Henry Briggs when he stated:

> Before the scientist there were two possible modes of procedure. First, he could in theory analyze effective writing into its elements and count the improvement in spelling, punctuation, choice of words, sentence structure, and the like: or, second, he could provide means of measuring the composition as a whole, considering the impression tout ensemble. The former plan was clearly impractical in that the same opportunities for error or effectiveness in details do not ordinarily appear in any two

---

[4] E. L. Thorndike, "A Scale for Measuring the Merit of English Writing," Science, XXXIII (June, 1916), p. 937.

[5] Hillegas, Loc. cit.

other on the scale on the basis of their general "merit." This was to be accomplished through the use of the Cattell-Fullerton Theorem of Significant Difference which states that a difference perceived by seventy-five percent of a given group of judges may be taken to be a unit of significant difference.[7] E. L. Thorndike himself first postulated this method in June of 1911 when he wrote:

> One inch may be said to be equal to another inch from any one of three lines of evidence. If the two are compared by a hundred experts, (1) the experts will report the two as indistinguishable; or (2) if some of them do, by microscope, micrometer or the like, find a difference of a trifle plus or minus, the number finding the first inch plus will equal the number finding it minus; or (3) if each man is forced to report a difference, half will find the first inch plus and half minus.
>
> One specimen of English may be said to be equal to another from the second or third lines of argument, the only logical difference between equating the two lengths and equating the two specimens of writing being that the variability of expert judges in the latter case is so great that we never find all of them, and rarely find many of them in agreement, as to the indiscernability of any difference.[8]

Milo B. Hillegas further explained the implementation and use of the Cattell-Fullerton Theorem applied to English themes when he noted:

> Any standard or scale should be based on a unit such that equal units may be derived independently of the scale. The unit in this scale has been defined as that difference which seventy-five percent of the judges are able to distinguish. All that is required to derive this unit is a set of samples that vary from each other by small degrees

---

[7] J. M. Cattell and G. S. Fullerton, On the Perception of Small Differences (Philadelphia: University of Pennsylvania Press, 1892).

[8] Thorndike, "A Scale for Measuring the Merit of English Writing," Op. cit., 935-936.

> in quality. When two samples are found such that seventy-five percent of the judges agree in calling one better than the other, the difference is just that difference used on the scale.[9]

E. L. Thorndike summarized the sentiments of the early scale-makers in their approach to English composition when he wrote:

> The "composition-meter," or scale for merit in English writing . . . consists . . . of a zero-point and of points at various exactly determined distances above this zero. Thus, quality 77 is as far above quality 67 as quality 47 is above quality 37. A composition that is regarded by impartial judges as being of the same merit as the specimen representing 93 is twice as good as a composition of quality 47. Wherever this scale was used, a mark of 40 or 60 or 80, if given without bias, would mean a known degree of excellence in paragraph-writing, just as 80 pounds means a known degree of weight wherever the avoirdupois scale is used. By using such scales, the absolute gain which any pupil can make . . . could be measured in the same way as his gain in height, weight, wages or pulse-rate.[10]

## II. ASSUMED ADVANTAGES OF COMPOSITION SCALES

Thus, with the assumption that scales for the measurement of quality in English composition were established on a sound basis of theory, educators were eager to reap the numerous benefits thought to be forthcoming from the practical application of these scales.

The first of these assumed advantages of writing scales to be anticipated was that of the objectification of standards for composition. The principle of massed value judgements upon which the

---

[9] Hillegas, "Scale for the Measurement of Quality in English Composition by Young People," Op. cit., 21.

[10] E. L. Thorndike, Education (New York: The Macmillan Company, 1912), pp. 213-214.

scales were based had appeared early in the history of education in the United States when the Commissioner of Education, in his report of 1897-98, stated:

> The difficulties of estimating intellectual ability in a quantitative way are well known, yet when there is an agreement in the reports of, say, more than ten teachers as to twenty or more pupils, there is a strong probability as to the general truth of the teachers' judgement. In questions where there is difference of opinion, the agreement of ten or more teachers is more trustworthy than the opinion of any single individual who is liable to have some cherished theory.[11]

Such an attitude as that expressed above was responsible for the enthusiasm for the initiation of composition scales. A large measure of this enthusiasm for the objective characteristics of early composition scales is demonstrated in the words of Breed and Frostic when they write:

> Among the various school subjects, English composition is generally conceded to be one of the most difficult to measure. The best-known means that have been prepared for the measurement of this subject is the Hillegas-Thorndike scale . . . The Hillegas-Thorndike scale measures the "quality of English compositions by young people," and by quality is meant general merit ....
>
> This kind of scale is not original with scientific students of education. It is well known in other scientific fields. Some readers may be familiar with the scale of hardness used in mineralogy . . . The scale of hardness is a series of mineral specimens ranging through ten steps of hardness. The hardness of each of specimens of the scale has been carefully determined. The hardness of any unknown material may be tested by comparing it with the scale. When a mineral on the scale is found that resists abrasion to about the same degree as the unknown, the hardness of that member of the scale is assigned to the mineral tested . . . .
>
> In the same manner in this composition scale we have a series of samples whose values have been care-

---

[11] Report of the Commissioner of Education, (Washington: U. S. Government Printing Office, 1898), pp. 1041-1042.

> fully determined by the average judgement of over sixty individuals, nearly all of whom were experienced teachers, twenty of them specialists in English, and the remainder students of education . . . .
>
> Just as the student of mineralogy can determine more accurately with the hardness scale . . . so the teacher can determine the merit of a sample of English composition with greater precision by comparing it with a graded series of evaluated compositions such as the present scale.[12]

The enthusiasm for the objectification of English composition standards was matched only by eagerness on the part of educators to use the tool of scales for quantitative measurement of school and class efficiency. An indication of this eagerness is given by the rhetorical question that closes the following statement made by D. C. Bliss in 1912:

> Dr. Thorndike published in The Journal of Educational Psychology for September, 1911 his article dealing with a scale of merit in English writing . . . Superintendents and principals are now asking themselves the question: Is it feasible to make use of standard tests in my school in such a manner as to determine relative classroom efficiency?[13]

An answer to this question is firmly offered by Guy Mitchell Wilson in his book, How to Measure, when he writes:

> In addition to the diagnosis of the language abilities of her class and a comparison of the results with standards of attainment in other cities, the teacher, with the use of the composition scales, should also be able to compare her class with other classes in the same school or in other schools of the same city; the supervisor should likewise be able to know the strong

---

[12]Fredrick S. Breed and F. W. Frostic, "A Scale for Measuring the General Merit of English Composition in the Sixth Grade," Elementary School Journal, XVII (January, 1917), 308-309.

[13]Thorndike, Education, Loc. cit.

and weak points in her organization.[14]

Yet another hope derived from the initiation of composition scales was that they might speed the grading of papers submitted to the teacher of English. This had long been a hope of all English instructors and it was felt that, with samples of ranked compositions before them, those responsible for the grading of English themes would have a much easier time of performing this arduous task.[15]

The final expectation for the early composition scales was that they might serve as a motivational device for students of English. This hope was clearly expressed in the following passage written by Hudelson:

> A knowledge on the part of the child of how much progress his last theme shows over the one previously scored by the use of the same scale is an incentive to still further improvement. It is advisable, therefore, either to put a copy of the scale into the hands of each pupil or to exhibit it where all may consult it, to study together the scale and the reasons for assigning the various samples their values, and so take the pupils into one's confidence to such a degree that they will not look upon teacher's marks as mysterious symbols which they are not supposed to understand and which, if they knew the truth, the teacher can neither fully explain nor justify.
>
> This practice has also a distinct social justification. Encouraging a pupil to attain higher and higher steps on the scale substitutes in part for the old and sometimes envious group rivalry a more salutary

---

[14]Guy Mitchell Wilson, How to Measure in Education (New York: The American Book Company, 1924), pp. 362-364.

[15]Marion Rex Trabue, Measuring Results in Education (New York: The American Book Company, 1924), pp. 362-364.

and progressive competition with himself.[16]

Thus, with so much enthusiasm being generated by the composition scale at the time of its initial formation, it would be wise to turn to an examination of the historical development of such scales and an attempt to discern their relative success or failure in practical application.

---

[16]Earl Hudelson, English Compostion: Its Aims, Methods, and Measurements (Bloomington, Illinois: Public School Publishing Company, 1923), p. 39.

## CHAPTER III

## THE HISTORY OF THE COMPOSITION SCALE MOVEMENT

A delineation of the composition scale movement should begin with the initial efforts of J. M. Rice in 1903. In that year Rice, a pioneer in educational testing, began work on reducing the variability in scoring mechanical and structural aspects of compositions. Rice began his study by having a story read to more than eight thousand pupils in various schools and then asking these students to reproduce the story. The first drafts of these themes written in class were then graded into five piles: Excellent, Good, Fair, Poor, and Failure. From this distribution, class averages were computed and samples of the reproduced story were selected as a guide to further scoring. Although unsophisticated in approach and focused upon the mechanics of writing, The Rice rankings served as a progenitor for the more elaborate statistical scales that were to follow.[1]

### I. HILLEGAS SCALE

The first true scale in the movement to measure ability at English composition in terms of general merit was produced by Milo

---

[1]Earl Hudelson, English Composition: Its Aims, Methods, and Measurement, (The Twenty-Second Yearbook of the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Company, 1923), pp. 42-43.

B. Hillegas in 1912.[2] Following the example set by the earlier efforts of Rice, Hillegas collected from various sources samples of the writing of young people from the grades through college.[3] Some seven thousand compositions written by the youth mentioned were then sorted roughly into ten classes ranging in quality from the poorest to the best. Seventy-five samples were then selected from these ten classes by Hillegas and his staff as being representative of the original seven thousand. To these seventy-five compositions were added artificial samples written by experts to represent zero merit and the writings of adults to represent superior merit.[4] The eighty-three samples thus selected were then subjected to the judgements of almost five hundred persons to arrange them in an order of merit. These judges included men of outstanding literary ability, master teachers, and a number of psychologists.[5] From the rankings of these judges, the number of samples was reduced to twenty-nine and again submitted to more than one hundred similar judges for further evaluation. When the final ratings were examined,

---

[2]Milo B. Hillegas, "Scale for the Measurement of Quality in English Composition by Young People," Teachers College Record, XIII (September, 1912).

[3]Marion James Van Wagenen, A Teacher's Manual in the Use of the Educational Scales, (Bloomington, Illinois: Public School Publishing Co., 1928), pp. 268-269.

[4]Eugene Mark Hinton, An Analytical Study of the Qualities of Style and Rhetoric Found in English Compositions (Teachers College Contributions to Education, No. 806. New York: Teachers College, Columbia University, 1940), pp. 2-3.

[5]Rollo LaVerne Lyman, Summary of Investigations Relating to Grammar, Language, and Composition (Chicago: The University of Chicago, 1929), pp. 137-138.

Hillegas and his staff selected ten specimens representing all types of writing except poetry to make up the finished scale.[6]

The judgement of the samples collected for the Hillegas study is interesting in terms of the fact that it employs the collective statistical method. Briefly stated, the main assumption of the method is as follows, "Differences that are equally often noticed are equal, unless the differences are either always or never noticed."[7] The unit of difference that separates the compositions listed on the scale is:

> . . . that difference which seventy-five per cent of the judges are able to distinguish . . . When two samples are found such that seventy-five per cent of the judges agree in calling one better than the other the difference is just the difference used as the unit on the scale.[8]

It is also worthy of note that Hillegas made no attempt to establish objective standards of merit in English composition.[9] In separating the sample compositions into their respective categories, Hillegas relied on a definition of general merit in composition which was taken to be, "just that value which competent persons commonly consider as merit."[10]

As to the use of the scale, Hillegas commented:

> The value of any English composition may be obtained by placing it beside the samples constituting

---

[6]Hinton, Loc. cit.

[7]Hillegas, Op. cit., p. 18.

[8]Hillegas, Op. cit., p. 9.

[9]Ibid. p. 13.

[10]Ibid.

> the scale and determining to which it most nearly corresponds.[11]

Just how the person making the comparison is to make this determination was not indicated by Hillegas.

The author of this first true scale of English composition recognized, upon publication of the scale, that the instrument was hardly perfect. Hillegas noted:

> No claim is made that the values given in the scale are absolutely perfect. Variation among the judges was very great, and to make a perfect scale would require the services of more judges than it was possible to secure for this study. The scale is accurate enough to be of very great practical value in measuring the merit of English compositions written in the upper grades of the elementary school and in the high school. The scale will also serve as the basis of future efforts in this direction, and it can be refined and perfected part by part.[12]

It is toward the refinement and perfection of the composition scales that we now look to discover their subsequent development.

## II. BALLOU SCALE

One of the initial shortcomings of the Hillegas scale was that the sampling of themes taken to provide raw material for the scale made no account of the nature of the compositions in regard to the four forms of discourse.[13] This step on the part of Hillegas met with widespread criticism from those attempting to use the scales for measurement purposes. It was held that this shortcoming limited

---

[11] Ibid.

[12] Ibid., p. 56.

[13] Hillegas, On. cit., p. 14.

the accuracy of the Hillegas scale and made comparison of themes written in forms of discourse different from those on the scale difficult if not impossible.[14] As a direct result of failure on the part of the teachers of Newton, Massachusetts, to successfully apply the Hillegas scale in grading compositions, Frank W. Ballou was led to formulate his own composition scale in 1914.[15] Ballou's Harvard-Newton scale attempted to overcome the weakness of the Hillegas scale by arranging all sample compositions gathered for the new formulation into the four classic forms of discourse: description, exposition, narration and argumentation.[16]

The Ballou scale also employed the collective statistical method used by Hillegas in much the same manner. A large number of samples of the writing of eighth-grade students was collected by Ballou and his staff and reduced to a representative group of twenty-five themes for each of the types of discourse mentioned, thus making a total of one hundred themes. These themes, in turn, were graded into piles representing 95%, 85%, 75%, 65%, 55%, and 45% merit respectively by the elementary teachers and principals of the Newton, Massachusetts, public schools. The accuracy of these gradings was predicated on the assumption that those themes judged to fall into a particular percentile category by at least seventy-five percent

---

[14]James Fleming Hosic, "The Essentials of Composition and Grammar," School and Society, XVII (April, 1915), pp. 583-584.

[15]F. W. Ballou, Scale for the Measurement of English Composition (Harvard-Newton Bulletin, No. 2. Cambridge: Harvard University Press, 1914).

[16]Hosic, Op. cit., p. 584.

of the judges must, of statistical necessity, be of that value. From the twenty-five themes sorted in the manner described one theme was selected to represent each of the four categories of discourse, thus yielding four separate scales comprised of six themes each. Comments from the twenty-four judges were compiled by Ballou and affixed to back of each of the scale themes to explain its respective rank in terms of merit as compared to each of the compositions above and below it on the scale.[17] Judges for this scale were drawn from the ranks of the Newton schools in the belief that those who instructed the students in composition were best qualified to rank the themes produced according to a fixed percentile rating.[18] This necessity for local adjustment of composition scales was to have great importance in the later criticisms of the scale movement.

## III. THORNDIKE SUPPLEMENT TO HILLEGAS SCALE

The next phase of the composition scale movement was the attempt by Thorndike to supplement the original Hillegas scale.[19] This supplement was published in 1915 and consisted of a number of sample themes substituted for those contrived by adults on the Hillegas scale. It also included an increased range of composition

---

[17] Ballou, Op. cit., pp. 28-37.

[18] Lyman, Op. cit., p. 138.

[19] Edward L. Thorndike, An Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People (New York: Teachers College, Columbia University, 1915).

examples at the middle of the scale to give more adequate and numerous models for the type of composition that a majority of a class could be expected to write. This addition improved the range of possible theme comparisons, but lengthened the Hillegas scale to an extent that required extensive practice for accuracy in its use.[20]

## IV. TRABUE'S CORRECTION OF HILLEGAS

In the spring of 1916, Marion Rex Trabue undertook the task of making a number of corrections in the Hillegas scale of 1912. Trabue had noted that the compositions that made up the Hillegas scale were lacking in uniformity of theme and form. On the basis of this notation, Trabue had 5,500 themes collected from the elementary school children of Nassau County, New York. All of these themes were written on the topic, "What I Should Like to Do Next Saturday" and Trabue urged that the children be encouraged to write in narrative form.[21] In ranking the themes thus collected, Trabue attempted to overcome another of the shortcomings of the Hillegas scale by training the judges in the use of this instrument until the discrepancy in their judgements on any given single theme was quite small. This eliminated some of the weakness inherent in Hillegas' initial assumption that the variability of judgements on one composition will be exactly equal to the variability of judge-

---

[20] Hinton, Op. cit., pp. 3-4.

[21] Ibid.

ments on another composition.[22] The original 5,500 themes were reduced to a set of twenty-eight by two judges trained in the use of the Hillegas scale. These themes were then submitted to 139 different judges who reduced their number to seven. The eighth and ninth specimens were taken from the Thorndike supplement, and the tenth was an artificial sample from literature. These ten ranked compositions comprise the final scale by Trabue.

Trabue's scale also refined the Hillegas effort in a few other important respects. First, Trabue evaluated and substituted the compositions of children for the artificial specimens at the bottom of the Hillegas scale.[23] In fact, so successful were the efforts of Trabue in this line that he managed to collect a student theme that a large majority of the judges agreed in calling zero ability.[24] A final refinement of the work by Hillegas was Trabue's use of longer samples in his own scale so that the reader might more easily obtain an appreciation of their quality for comparative purposes.[25] In the opinion of at least one author, no better objective scale has been devised for measuring composition achievement than that of Trabue.[26] The first publication of this scale came

---

[22] Ibid.

[23] Lyman, Op. cit., p. 139.

[24] Van Wagenen, Op. cit., p. 269.

[25] Lyman, Op. cit., p. 140.

[26] Hudelson, Op. cit., p. 50.

in 1917.[27]

## V. BREED AND FROSTIC SCALE

At about the same time that Trabue was making his survey of the elementary schools in Nassau County, Breed and Frostic were collecting compositions from a similar survey of ten Michigan cities.[28] The first part of a story, "The Picnic," was read to children in the sixth grade. They were then asked to finish the story in any way they wished.[29] The children were given twenty minutes to complete this task and were asked to do so in narrative form.[30] In selecting and evaluating their scale samples these authors followed the methods devised by Hillegas. In order to approximate actual pupil production, all features of the compositions were reproduced except the pupil's own handwriting.[31] This was the only composition scale which attempted to reproduce the physical characteristics of the written compositions which compose it. Its chief distinction, however, was the homogeneity resulting from its narrow range of merit. Its concentration on the adolescent period,

---

[27]Marion Rex Trabue, Nassau County Supplement to the Hillegas Scale (New York: Teachers College, Columbia University, Bureau of Publications, 1917).

[28]Fredrick W. Breed and F. W. Frostic, "A Scale for Measuring the General Merit of English Compositions in the Sixth Grade," Elementary School Journal, XVII (January, 1917).

[29]Hinton, Op. cit., p. 4.

[30]Hudelson, Loc. cit.

[31]Hinton, Loc. cit.

its indicative topic, and its having been devised under controlled conditions make the Breed and Frostic scale a sound instrument for measuring composition ability. This scale, unfortunately, received little or no attention on the part of educators.[32]

## VI. WILLING SCALE

In 1918 there appeared a scale for the measurement of composition ability that attempted a separation of merit into the components of content and mechanics. This scale was the work of Matthew H. Willing and was the result of compositions written by pupils in grades four through eight in the Denver and Grand Rapids public schools on the topic, "An Exciting Experience."[33] Willing selected sixty-three samples which he thought represented fairly well a cross section of all the compositions written. The author then submitted them to twelve judges with the request that the samples be ranked from the poorest to the best on purely rhetorical grounds. From the ratings thus made, the author selected eight specimens which appeared to represent the total range of quality obtained.[34] Without the cooperation of the judges, Willing then constructed a scale somewhat more crude in a scientific or statistical sense than the Hillegas and Thorndike scales. The scale ranges from the value of A to H. Accompanying each sample was a statement of the number of

---

[32]Hudelson, On. cit., p. 51.

[33]Matthew H. Willing, Scale for Measuring Written Composition (Bloomington, Illinois; Public School Publishing Company, 1918).

[34]Hinton, Loc. cit.

errors in spelling, punctuation, and grammar per hundred words; such errors increased in number from five per hundred words for the B theme to thirty per hundred words for the H theme.[35] A teacher using this scale assigns two scores, one for errors as indicated and one for content value, for which no criteria are provided. If a composition ranks high in content and receives a low mark in form, a score between the two is approximated, but "no paper is marked above 70 which does not have good story value and technical excellence; nor is a paper marked below 40 which does not lack both of these qualities."[36] The confused nature of this scale was obvious and it was this lumping together of the scores for style that brought Willing his most severe educational criticism.[37] Nonetheless, the Willing scale and its separation of content and mechanics points to the fact that educators were now beginning to doubt the value of scales based on the measurement of a single quality called "merit."

## VII. VAN WAGENEN SCALE

Following the lead set by Willing, Marvin James Van Wagenen decided to further subdivide the elements to be judged by a composition scale into three areas: thought content, sentence and paragraph structure, and mechanical errors.[38] Using the judgement of

---

[35] Lyman, Op. cit., p. 144.

[36] Willing, Op. cit., p. 198.

[37] Hudelson, Loc. cit.

[38] Van Wagenen, Op. cit., p. 15.

forty-one experts in the field of composition and a statistical operation identical to that of the Hillegas scale, Van Wagenen reduced 600 themes to three scales of exposition, description, and narration. Each of these three scales were then judged three times on the basis of thought, structure, and mechanics thus yielding nine scales in all.[39] Van Wagenen's rationale for this operation was contained in the following statement:

> Each quality . . . must be considered as a distinct scale in itself. Only as a matter of accident would 80 in thought content, for instance, be the same distance from the arbitrary zero point selected as 80 in mechanics . . . Hence, simply adding together the three values assigned for the three qualities of a theme to get the general merit is not much, if any, more accurate than would be the measurement of a liquid obtained by adding together its density, its temperature, and its volume, and then dividing the result by three.[40]

Thus, separate values had been assigned to each specimen in each scale for thought content, structure, and mechanics. The three qualities were not evaluated in equivalent terms in the same scale, but each quality in each scale furnished practically an equivalent scale for the same quality in either of the other two discourses. Therefore, a 72 in thought content was not equal to a 72 in either structure or mechanics within the same scale or in either of the other two scales; but a 72 in thought content on any one of the scales was practically equal to a 72 in thought content on either of the other two scales.[41]

---

[39] Lyman, Loc. cit.

[40] Van Wagenen, Op. cit., p. 275.

[41] Ibid.

While the Van Wagenen scales represented a worthy attempt to analyze composition writing for diagnostic purposes, they rendered judgements confusing and difficult if, as was customary with teachers, the separate evaluations were combined into one general score.

## VIII. HUDELSON SCALE

In 1923, the same year that saw the publication of the scales by Van Wagenen, Earl Hudelson constructed his Maximal Composition Ability Scale.[42] This scale differed little from the work of Hillegas and, in fact, used the Thorndike extension of the Hillegas scale to rank a judgemental reduction of 800 narrative themes into twenty steps that represented .5 values of the original ten-sample scale.[43] Hudelson must be credited for his emphasis on narrative form and his attempt to make the steps of the scale more uniform. The length of the resulting scale, however, cancelled the assumed benefits and rendered the work useless without extensive practice.

## IX. FINAL ATTEMPT BY LEONARD

Other than an unsuccessful attempt by S. A. Leonard to measure composition ability apart from mechanics, the scale by Hudelson marked the close of an era in which educators, having found the new tool of statistical measurement, felt confident that anything could

---

[42] Hudelson, Op. cit., p. 52.

[43] Ibid., p. 53.

be measured.[44] The hope that composition ability, or "merit" as it was often called, could be measured scientifically as a single operational entity had ended. Diversity in approach and method now marked the field of testing English composition ability as evidenced by Lewis' scale for measuring the ability at composing letters.[45] Other so-called "scales" by a variety of authors showed no confidence or interest in the statistical operations carried out by Hillegas and his successors. This turning-point marked the close of the scale theme movement.[46] Perhaps Hudelson himself best expressed the sentiment of the times concerning composition scales when he noted that:

> It is doubtful whether we shall get much further either by Van Wagenen's scheme or with general-merit scales. It is likely that progress will be made in the future with scales designed to measure only one composition element at a time, such as clearness or capitalization.[47]

The fragmentation of purpose notable in this statement is the key to the ultimate death of the composition scales in the manner of Hillegas and leads now to an investigation of the criticism leveled against the attempts to measure writing ability in English on the sole basis of merit.

---

[44]S. A. Leonard,"Building a Scale of Purely Composition Quality," English Journal, XIV (September, 1925).

[45]E. E. Lewis, Scales for Measuring Special Types of English Composition (Yonkers, New York: The World Book Company, 1926).

[46]Hinton, Op. cit., pp. 8-13.

[47]Hudelson, Op. cit., p. 52.

## CHAPTER IV

## THE FAILURES OF THE COMPOSITION SCALES

To understand the decline and failure of attempts to measure writing ability on the single basis of quality it is necessary to examine some of the criticism evoked by the composition scales. Such criticism of the composition scales falls into general categories and these may be enumerated as follows: (1) the scales were criticized on the basis of their scientific validity; many critics questioned the ability of a measuring device based on opinion to yield objective results, (2) objections were made by critics concerning the sample themes on the scales in regard to their artificial nature and a lack of their clear division into the classic forms of discourse, (3) composition scales were accused of stripping the student theme of its individuality and of providing no analytical suggestions for remedial work with the individual student, and (4) the scales were found difficult and time-consuming in practical application because they attempted to measure too complex a product in one operation.

### I. SCIENTIFIC VALIDITY

The first specific criticism to be leveled against the composition scales was one concerning their scientific validity. Isadore Kayfetz, writing in the Pedagogical Seminary of December, 1914, made this penetrating comment on the supposedly "scientific" methods of the early scale-makers:

> Hillegas did not study the composition work of school children under normal conditions. He studied the opinions of "expert" judges as to the relative merits of pupils' compositions. He tells us nothing of the conditions under which the compositions that formed the material of this study were written. Since he loses sight of this important requirement we are warranted in assuming that the conditions were not uniform nor fully controlled.[1]

With these words Kayfetz opened the floodgates of criticism concerning the soundness of the empirical methods used to formulate the composition scales.

## II. COLLECTIVE STATISTICAL METHODS

The next criticism of the composition scales was one directed to the scientific method employed in their derivation. If the premise was accepted that some benefit was to be gained from ranking a number of English themes in an order corresponding to their relative general merit, what then was to be the basis of this ranking? Many educators took issue with the "collective opinion" scheme proposed by Hillegas and his followers shortly after the publication of the first composition scales.[2]

The first of these educators to take issue with the collective statistical method used to derive the composition scales was James Drever, and he did so on the grounds that the scales could never be truly objective since they relied upon subjective judgements for a large measure of their content. Drever made this issue clear in

---

[1]Isadore Kayfetz, "A Critical Study of the Hillegas Composition Scale," Pedagogical Seminary, XXI (December, 1914), p. 569.

[2]Kayfetz, Op. cit., p. 570.

the following statement:

> Now the fundamental objection to such a scale is that it can never be in any real sense objective, nor can its use by any individual give an objective determination of the merit or value of any specimen of writing. The scale is not objective, because it is simply the average of a number of individual opinions of merit, a composite portrait of a number of subjective opinions . . .[3]

W. F. Tidyman also made a clear assessment of the problem of using collective statistical techniques to formulate a composition scale when he wrote the following criticism of Rice's early efforts to measure spelling efficiency:

> The statistical method is limited to facts of quantitative determination. Qualitative facts are beyond its sphere. It is concerned with the what, not with the causes and conditions underlying phenomena. Because of these limitations of nature and purpose it cannot settle pedagogical questions. Many questions it cannot answer at all.[4]

Finally, Matthew H. Willing delivered perhaps the strongest charge against the composition scales and their supposedly objective nature when he noted:

> No composition scale yet published is objective either in its derivation or in its use. All are the product, more or less, of massed opinion, and in their application are at the mercy of the special intelligence and experience of those who use them.[5]

Yet another problem encountered with the use of a large number

---

[3]James Drever, "Notes on the Experimental Study of Writing," Journal of Experimental Pedagogy, II (March, 1913), p. 28.

[4]W. F. Tidyman, "A Critical Study of Rice's Investigation of Spelling Efficiency," Pedagogical Seminary, XXII (September, 1915), p. 397.

[5]Matthew H. Willing, "The Measurement of Written Composition in Grades Four to Eight," English Journal, VII (March, 1918), p. 198.

of opinions of expert judges to determine the merit of a particular piece of writing is mentioned in the words of F. W. Johnson when he criticizes the Hillegas scale:

> The scale represents the composite judgment of some five hundred individuals more or less expert as teachers of English, writers, and experts in fields that render their individual judgments worthy of respect. But the scale represents the judgment of no single individual, and the judgment of a single individual or of any group of individuals whose judgment entered into the formation of the scale such as might reasonably be expected to apply the scale to any large body of material necessary to an adequate practical test would not represent the aggregate judgment of those whose arrangement of the material formed the basis of the scale. But the scale cannot be used in practice by all the persons whose judgments may be expected to vary as widely as did those in the making of the scale. As there is no such person as the average pupil, so there is no average judgment which can be applied to a test of school products.[6]

This revealed the fact that not only were the scales questionable in terms of their scientific validity but that the very judges whose opinions were collected statistically to formulate such scales would have difficulty in making a practical application of these measuring devices since they reflected the judgment of no one individual.

## III. SAMPLES USED ON THE SCALES

Apart from the difficulties encountered with the scales in regard to the method used to formulate them, such measuring devices met with further criticism in terms of the materials they attempted

---

[6] F. W. Johnson, "The Hillegas-Thorndike Scale for the Measurement of Quality in English Composition by Young People," School Review, XXI (January, 1913), p. 47.

to incorporate for illustrative and measurement purposes. As early as 1911, William H. Dall had noted the difficulty of comparing any two given paragraphs so long as they were written on different topics.[7] In 1913, F. W. Johnson made objection to the sample compositions contained in the early scales on the grounds that such samples were stilted and artificial, thus making comparisons between the work of students and the writings on the scale difficult. Johnson states:

> It cannot be supposed that composition subjects in any school were chosen exclusively from such a barren field as is represented by the material on these scales. Certainly no school should be expected to furnish any large amount of material on subjects so far removed from all present-day human experience. It is altogether impossible to compare compositions on subjects that offer opportunity for originality of thought and expression with the formal material found in the scale which depends for its content largely upon the memory of books read or discussed in class. The scale differs from the material to be measured. It is like using a yardstick to determine the weight of material in the physical laboratory.[8]

P. M. Watson reiterates the charge made by Johnson and adds the thought that the arbitrary division of composition samples into the four classic forms of discourse will not help remove the problem of artificiality, since these divisions of composition do not normally occur in student writing.[9] This critical position is well

---

[7]William H. Dall, "Measuring the Merit of English Writing," Science, XXXIV (June, 1911), pp. 115-116.

[8]Johnson, Op. cit., p. 48.

[9]P. M. Watson, "The Harvard-Newton Composition Scale," Educational Administration and Supervision, I (January, 1915), p. 58.

summarized in the words of Ernest J. Ashbaugh, writing in the Journal of Educational Research, when he states:

> . . . anyone who has read many children's themes will instantly recall that children seldom write themes which are wholly narrative, descriptive, argumentative, or expository. They have a very disconcerting way of mixing two or more forms into one inglorious whole.[10]

Finally, in regard to the sample compositions listed on the scales for measuring writing merit, Isadore Kayfetz again affirmed the need for more empirical information concerning the background and events which produced these sample works of students. He commented that composition scales were defective since:

> . . . they do not give us full and detailed information regarding all the objective and subjective conditions under which the compositions were written. We cannot judge a composition properly from the pedagogical point of view unless we know the following facts concerning the writer of the composition: 1. age, 2. grade, 3. sex, 4. intelligence, 5. socio-economic background.[11]

## IV. DIFFICULTY IN PRACTICAL APPLICATION OF SCALES

The criticisms concerning the methods and procedures used for the formulation and construction of the composition scales were second in number only to those elicited by the attempts at practical application of these measuring devices.

One of the earliest educational concerns that resulted from the attempts to apply composition scales to classroom use was that

---

[10]Earnest J. Ashbaugh, "The Measurement of Language," Journal of Educational Research, IV (June, 1921), p. 32.

[11]Isadore Kayfetz, "A Critical Study of the Harvard-Newton Composition Scale," Pedagogical Seminary, XXIII (September, 1916), pp. 337-338.

for the individual student. It was felt that the composition scales, with their emphasis on national and group standards, would deprive the student of the attention he deserved in terms of his individual ability at expression and stifle attempts to measure his progress against his personal capability for writing in terms of progress or decline. This sentiment is clear in the words of F. N. Scott when he states:

> The student's composition, as the teacher should look at it, is an expression of the student's life. To evaluate it is to evaluate life itself in one of its most delicate manifestations. When, however, applying to it a scale . . . we strip it of its individual character and reduce it to an abstraction, we excise at one stroke the most significant and essential features.[12]

The next criticism to result from the attempt to make practical application of the composition scales was one again directed to helping the individual student with his particular writing problem. The Presseys, in their book on standard instruments of measurement used in the classroom, criticized the composition scales for leaving the teacher few remedial suggestions with which to improve the writing of her students. They stated:

> The scales so far constructed investigate ability in written English in general; they do not analyze the situation so that the teacher will know just what to emphasize in her corrective instruction.[13]

A final criticism of the composition scales in regard to their adaptability to the individual concerns a need to coordinate such

---

[12] F. N. Scott, "Our Problems," English Journal, II (January, 1913), pp. 4-5.

[13] L. C. Pressey and S. L. Pressey, An Introduction to the Use of Standard Tests (New York: World Book Company, 1922), p. 101.

measuring devices with the grade level of the particular student in question. James Flemming Hosic, writing in School and Society, made this necessity clear when he wrote:

> It should be remarked that such scales are intended to provide a fixed objective standard. They do not indicate what may be expected at any particular point in the school course. Supposing the scales to be a just estimate of excellence in composition, we should still be in doubt as to whether a boy in fifth grade ought to be expected to write a composition as good as some particular example on the scales, and if so whether on first attempt or after careful revision.[14]

Apart from considerations of the individual student, other difficulties arose with the practical application of the composition scales. This second area of difficulty centered in the teachers using the scales and the problems encountered by them in making actual theme comparisons with the composition scales. The previously mentioned problem of dissimilar subject matter among the various samples on the scales is here reiterated by William E. Stark in terms of teacher difficulty encountered in such practical applications. Stark noted:

> The chief difficulty in using the scales seemed to be that the character of the material in the scale samples was so different from that of the compositions to be measured that the reader was often uncertain as to which of three or four steps should be regarded as representative of the merit of a given composition.[15]

Not only did teachers experience difficulty in making practical application of the writing scales because of the diffuse composition

---

[14]James Flemming Hosic, "The Essentials of Composition and Grammar," School and Society, I (April, 1915), p. 583.

[15]William E. Stark, "Measurement of Eighth-Grade Composition," School and Society, II (August, 1915), p. 209.

topics they contained, but also because the scales, even if constructed of themes written on a single topic, failed to account for the difference in ability to write on a single given topic caused by a student's geographical location. Earl Hudelson clarified this difficulty when he related that:

> No standard, probably, will ever be made that will be equally suited to all schools. Just as we found it advisable to substitue writing topics in our tests at Bloomington, so I believe each school should have essentially its own composition standard, in order that compositions typical of the locality might be found.[16]

Even if the composition scales were able to overcome their weaknesses of subjective derivation, unproved statistical assumptions, and difficulty of practical application, there would still be the need for extensive training and practice on the part of the classroom teacher in order to make the instruments workable as measuring tools.[17] In fact, without this adequate and rather extensive preparation, teachers using the composition scales might well be criticized for the variability of their grades on a given set of themes at different intervals of time. An example of such criticism was apparent in the comments of Fredrick James Kelley concerning his study of the reliability of teacher judgements of compositions based on the use of several writing scales. Kelley noted:

> It is thus seen that the distributions obtained in this study show rather more than normal variability for the unpracticed (normal being determined by the

---

[16]Earl Hudelson, "Some Achievements for the Measurement of English Composition in the Bloomington, Indiana Schools," English Journal, V (November, 1916), p. 596.

[17]Willing, Op. cit., p. 200.

> variability of the judges whose ratings entered into the makeup of the scale). The very effort to define general merit in so complex a thing as a composition by a single example seems to make great variation possible.[18]

Perhaps the last sentence of this statement by Kelley best summarized all of the criticism leveled against the effort to measure the worth of a composition on the single basis of general merit.

## V. DECLINE IN USE OF SCALES

As we seek to find the reasons for the death of the composition scale movement, it is quite possible to see the role played in this decay through the immediate and numerous criticisms leveled at the scales and the difficulty encountered when attempts were made to put these scales to practical use. One or two other factors, however, must be mentioned in order to complete the background necessary to an understanding of the failure of this movement.

It might be assumed that the central reason for the composition scales failing to achieve widespread popularity would be a basic ignorance of their existence on the part of classroom teachers. Clarification of this thought was offered in the following words of M. J. Van Wagenen:

> Although scales for measuring general merit in English composition have now been in existence for several years it is astonishing how few teachers of English make use of them or even know about them.

---

[18] Fredrick James Kelley, Teachers' Marks: Their Variability and Standardization (New York: Teachers College, Columbia University, 1914), p. 130.

> In part this is undoubtedly due to the hesitancy and trepidity with which people give up old ways of doing things to learn new and better ways. To some extent it may be due to the teachers' questioning of the advisability of using scales for measuring general merit when more interest may be felt in measuring the various qualities that go to make up general merit, especially when the emphasis to be given to each of the qualities of thought content, sentence and paragraph structure, or mechanical perfection is nowhere clearly expressed.[19]

From this statement it may be concluded that, although teacher ignorance of the composition scales may have had some bearing on their failure, such scales met their downfall at the hands of some other more basic shortcoming. This shortcoming may well have been the initial attempt to measure the quality of a whole composition on the single basis of merit. The difficulty in composition measurement that gives rise to this shortcoming was readily apparent in the following statement by Earnest J. Ashbaugh:

> Any analysis of written expression (which of necessity must be the phase of language ability which will be most frequently measured) must recognize at least three groups of factors, namely mechanical, grammatical, and rhetorical. Each of these groups contains many factors or elements, and only as we separate these complexes into simpler elements will our measurement become truly helpful to the teacher and supervisor in the improvement of the work in the classroom.[20]

From this statement we may conclude that an attempt to ascertain the general merit of any composition will, of necessity, involve the measurement of a complex of skills. The failure of the

---

[19] M. J. Van Wagenen, "The Accuracy With Which English Themes May Be Graded With the Use of English Composition Scales," School and Society, XI (April, 1920), p. 441.

[20] Ashbaugh, Op. cit., p. 34.

composition scales to make an adequate assessment of these several skills was apparent in the following statement made by Green and Jorgensen in their text on educational measurements. The authors commented:

> The measurement of general merit of written composition, while dating well back into the history of educational measurement, has not responded to efforts to improve it in proportion to the attention it has received. This difficulty comes from the great complexity of the skills involved in producing merit in written language, and from the vagueness with which these skills have been recognized.[21]

Perhaps the best reason for this vagueness of standards was indicated in a statement made by Milo B. Hillegas, author of the first true composition scale:

> No attempt has been made in this study to define merit. The term as here used means just that quality which competent persons commonly consider as merit, and the scale measures just that quality.[22]

Because this lack of any objective definition of the standards constituting merit was a characteristic of all composition scales that attempted to define the general quality of writing, critics of the scales were prompted to make such statements as this one by Isadore Kayfetz:

> No person, no matter how competent he may be in judging merit in composition, has any absolute standard of merit. His standard must necessarily be relative and variable. A child's composition is one of the most complex pieces of school work. It is extremely difficult

---

[21]Harry A. Greene and W. A. Greene, Elementary School Tests (New York: Longmans, Green and Company, 1954), p. 357.

[22]Milo Burdette Hillegas, A Scale for the Measurement of Quality in English Composition by Young People (New York: Teachers College, Columbia University, 1913), p. 13.

> to analyze. It is made up of a number of factors--literary, psychological, and pedagogical--each one of which is in turn exceedingly complex.[23]

Thus, above all other difficulties encountered in the formulation and application of the composition scales, the failure of such scales was primarily attributable to their inability to measure merit--something far too complex to be graded by a single scale.

## VI. SUMMARY OF REASONS FOR FAILURES

A review of the factors responsible for the failure of the composition scales would encompass several areas of difficulty. First, the scales met with a flood of initial criticism which rendered their credibility as scientific measuring instruments questionable. Next, the composition scales were found to be difficult to handle in practical classroom application. Third, the scales met with difficulty throughout their history because they attempted to measure composition ability on the basis of merit, a quality difficult if not impossible to define in the concise terms necessary for the formulation of such scales. A fourth and final source of the failure of composition scales to gain widespread acceptance and use may well have been ignorance of their existence on the part of the teacher of composition. The combination of these factors served to make the composition scale movement an appealing, but unsuccessful venture aimed at making more objective an area of English instruction which still receives little aid from scientific measuring devices.

---

[23] Kayfetz, "A Critical Study of the Harvard-Newton Composition Scale," *Op. cit.*, pp. 570-571.

## CHAPTER V

## RECENT ATTEMPTS TO ESTABLISH STANDARDS FOR ENGLISH COMPOSITION

Although the attempts to measure writing ability solely on the basis of merit were abandoned, interest in objective standards for English composition is still strong today. This interest is the direct result of the growing awareness of a lack of commonly accepted standards of good writing among all teachers of English. Such awareness, fostered by the research pertinent to the composition scale movement, stands as a permanent contribution to the fields of education and English even though the scales were themselves a failure.

### I. EFFORTS RELATED TO PAST SCALES

Some of the interest in objective standards for English composition has taken the form of sample themes, much like those listed on the composition scales of the past, treated in a variety of ways to suggest points for remedial attention to teachers and instructors. These samples have been collected into booklets by colleges, state education associations, and various individuals concerned with the teaching of composition.

### II. STATE BULLETINS

One example of the evaluation booklets mentioned is that of the Association of English Teachers of Western Pennsylvania. This booklet consists of a number of principles listed for evaluating

junior high school themes, a selection of actual themes taken from the writing of junior high school pupils corrected with interlinear notations, and a bibliography to assist the teacher of composition in gaining more background for theme correction.[1] Other booklets on the same pattern are produced by teacher associations in California,[2] Indiana,[3] Illinois,[4] and Kentucky.[5]

## III. INDIVIDUAL EFFORTS

Another effort to increase standardization of English theme grades is the individual work of Ednah Shepard Thomas, an instructor at the University of Wisconsin. This booklet contains the composition efforts of entering freshmen at the university classified into three groups representing themes of Unsatisfactory, Middle, and Superior quality. Interlinear corrections of the themes are omitted in favor of remarks by the author at the end of each

---

[1]Lois M. Grose (ed.), Suggestions for Evaluating Junior High School Writing (Pittsburg: The Association of English Teachers of Western Pennsylvania, 1960).

[2]N. Field Winn, et. al., A Scale for Evaluation of High School Writing (Champaign, Illinois: National Council of Teachers of English, 1960), Sponsored by the California Association of Teachers of English.

[3]Robert Hunting, et. al., "Standards for Written English in Grade Twelve," Indiana English Leaflet, III (October, 1960).

[4]"Evaluating Ninth-Grade Themes," Illinois English Bulletin, XIV (March, 1953), and "Evaluating Ninth-Grade Themes," Illinois English Bulletin, XIV (April, 1953).

[5]William S. Ward, "Principles and Standards in Composition for Kentucky High Schools and Colleges," Kentucky English Bulletin, VI (Fall, 1956-1957).

selection as to just why each composition is classified in one of the three categories. This booklet is intended to be a guide to students and instructors of English alike.[6]

## IV. COLLEGE SYLLABI

The various college syllabi for composition are too numerous for listing in this work, but it is interesting to note that a number of these do contain sample themes scaled to some extent according to their respective "merit."[7]

## V. CONCLUSIONS

It may be concluded that the composition scales which failed to completely objectify the grading or judgement of themes did, however, arouse great concern about the lack of standards for English composition. These scales also illustrated the inequality of grade values among various teachers and pointed to the pressing need for the clarification of all phases of appraising writing done for academic credit on all levels from the grades through college.

---

[6] Ednah S. Thomas, Evaluating Student Themes (Madison: The University of Wisconsin Press, 1955).

[7] Harry A. Greene, Developing Language Skills (New York: The Macmillan Company, 1959), p. 451.

BIBLIOGRAPHY

Books


Cattell, J. M. and G. S. Fullerton. On the Perception of Small Differences. Philadelphia: University of Pennsylvania Press, 1892.

Greene, Harry A. Developing Language Skills. New York: The Macmillan Company, 1959.

_____, and W. A. Greene. Elementary School Tests. New York: Longmans, Green and Company, 1954.

_____, and Albert N. Jorgensen. The Use and Interpretation of Elementary School Tests. New York: Longmans, Green and Company, 1935.

Hillegas, Milo B. A Scale for the Measurement of Quality in English Composition by Young People. New York: Teachers College, Columbia University, 1913.

Hudelson, Earl. English Composition: Its Aims, Methods, and Measurement. Bloomington, Illinois: Public School Publishing Company, 1923.

Kelley, Fredrick J. Teachers' Marks: Their Variability and Standardization. New York: Teachers College, Columbia University, 1914.

Lewis, E. E. Scales for Measuring Special Types of English Composition. New York: The World Book Company, 1926.

Lyman, Rollo L. Summary of Investigations Relating to Grammar, Language, and Composition. Chicago: The University of Chicago, 1929.

Monroe, Walter S. An Introduction to the Theory of Educational Measurements. New York: Houghton Mifflin Company, 1923.

Odell, Charles W. Educational Measurement in High School. New York: The Century Company, 1930.

Pressey, L. C. and S. L. Pressey. An Introduction to the Use of of Standard Tests. New York: The World Book Company, 1922.

Thomas, Ednah S. Evaluating Student Themes. Madison: The University of Wisconsin Press, 1955.

Thorndike, Edward L. Education. New York: The Macmillan Company, 1912.

_____. An Extension of the Hillegas Scale for the Measurement of Quality in English Composition by Young People. New York: Teachers College, Columbia University, 1915.

_____. An Introduction to the Theory of Mental and Social Measurements. New York: Teachers College, Columbia University, 1904.

Trabue, Marion R. Measuring Results in Education. New York: The American Book Company, 1924.

Van Wagenen, Marion J. A Teacher's Manual in the Use of the Educational Scales. Bloomington, Illinois: Public School Publishing Company, 1928.

Willing, Matthew H. Scale for Measuring Written Composition. Bloomington, Illinois: Public School Publishing Company, 1918.

Wilson, Guy M. How to Measure in Education. New York: The American Book Company, 1924.

_____, and Kremer J. Hoke. How to Measure. New York: The Macmillan Company, 1921.

Books: Parts of Series

Ballou, F. W. Scale for the Measurement of English Composition. Harvard-Newton Bulletin, No. 2. Cambridge: Harvard University Press, 1914.

Buckingham, B. R. Spelling Ability: Its Measurement and Distribution. Teachers College Contributions to Education, No. 59. New York: Teachers College, Columbia University, 1913.

Hinton, Eugene Mark. An Analytical Study of the Qualities of Style and Rhetoric Found in English Compositions. Teachers College Contributions to Education, No. 806. New York: Teachers College, Columbia University, 1940.

Stone, C. W. Arithmetical Abilities and Some Factors Contributing to Them. Teachers College Contributions to Education, No. 19. New York: Teachers College, Columbia University, 1921.

## Publications of the Government, Learned Societies, And Other Organizations

Ayres, L. P. Scale for Measuring the Quality of Handwriting of School Children, Russell Sage Foundation, Bulletin No. E-113, New York: Russell Sage Foundation, 1912.

Courtis, S. A. Manual of Instructions for Giving and Scoring the Courtis Standard Tests in the Three R's. Detroit: Department of Cooperative Research, 1910.

Grose, Lois M. (ed.). Suggestions for Evaluating Junior High School Writing. Pittsburg: The Association of English Teachers of Western Pennsylvania, 1960.

Hudelson, Earl. English Composition: Its Aims, Methods, and Measurement, pp. 42-43. Twenty-Second Yearbook of the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Company, 1923.

Report of the Commissioner of Education. Washington: Government Printing Office, 1898.

Thorndike, E. L. The Use of Educational Tests and Measurements. pp. 1-178. Seventeenth Yearbook of the National Society for the Study of Education. Bloomington, Illinois: Public School Publishing Company, 1918.

Winn, N. Field, et. al. A Scale for Evaluation of High School Writing. Champaign, Illinois: National Council of Teachers of English, 1960.

## Periodicals

Ashbaugh, Earnest J. "The Measurement of Language," Journal of Educational Research, IV (June, 1921), 24-32.

Breed, Fredrick S. and F. W. Frostic. "A Scale for Measuring the General Merit of English Composition in the Sixth Grade," Elementary School Journal, XVII (January, 1917), 308-309.

Briggs, Thomas Henry. "English Composition Scales," Teachers College Record, XXIII (November, 1922), 424-425.

Dall, William H. "Measuring the Merit of English Writing," Science, XXXIV (June, 1911), 112-117.

Drever, James. "Notes on the Experimental Study of Writing," Journal of Experimental Pedagogy, II (March, 1913), 25-28.

"Evaluating Ninth-Grade Themes," Illinois English Bulletin, XIV (March and April, 1953).

Hillegas, Milo B. "Scale for the Measurement of Quality in English Composition by Young People," Teachers College Record, XIII (September, 1912), 332-336.

Hosic, James F. "The Essentials of Composition and Grammar," School and Society, XVII (April, 1915), 582-587.

Hunting, Robert, et. al. "Standards for Written English in Grade Twelve," Indiana English Leaflet, III (October, 1960).

Johnson, F. W. "The Hillegas-Thorndike Scale for the Measurement of Quality in English Composition by Young People," School Review (January, 1913), 34-50.

Kayfetz, Isadore. "A Critical Study of the Hillegas Composition Scale," Pedagogical Seminary, XXI (December, 1914), 551-570.

_____. "A Critical Study of the Harvard-Newton Composition Scale," Pedagogical Seminary, XXIII (September, 1916), 335-352.

Leonard, S. A. "Building a Scale of Purely Composition Quality," English Journal, XIV (September, 1925), 28-37.

Scott, F. N. "Our Problems," English Journal, II (January, 1913), 4-5.

Stark, William E. "Measurement of Eighth-Grade Composition," School and Society, II (August, 1915), 198-210.

Thorndike, Edward L. "A Scale for Measuring Achievement in Drawing," Teachers College Record, XIV (November, 1913), 19-37.

_____. "A Scale for Measuring the Merit of English Writing," Science, XXXIII (June, 1916), 930-948.

_____. "A Scale for Merit in English Writing by Young People," The Journal of Educational Psychology, II (May, 1911), 340-363.

Tidyman, W. F. "A Critical Study of Rice's Investigation of Spelling Efficiency," Pedagogical Seminary, XXII (September, 1915), 387-398.

Ward, William S. "Principles and Standards in Composition for Kentucky High Schools and Colleges," Kentucky English Bulletin, VI (Fall, 1956-1957).

Watson, P. M. "The Harvard-Newton Composition Scale," Educational Administration and Supervision, I (January, 1915), 53-60.

# THE USE OF COMPOSITION SCALES FOR THE MEASUREMENT OF QUALITY IN ENGLISH COMPOSITION (1903-1963)

by

JON F. LOVE

B. A., Southwestern College, 1962

---

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

School of Education

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1964

Attempts have been made to establish objective standards for English composition throughout the history of educational measurement. One such attempt took the form of composition scales intended to measure the general quality of composition.

Scales have been a part of educational measurement since the early 1900's, and they have generally been samples of student work arranged in an order of merit through a statistical procedure.

Thorndike and other leaders in the field of education adapted scales to the measurement of quality in English composition through the use of the Cattell-Fullerton Theorem of Significant Difference. This theorem proposed that a difference perceived by seventy-five percent of a group of judges might be taken to be a unit of significant difference. Under the influence of this theorem, the judgements of a large number of authorities were used to arrange student themes into scales which represented writing quality from zero to the greatest possible merit in equal steps. The major intent of these scales was to speed the marking of student compositions by eliminating the necessity of separate gradings for content and mechanics. It was further assumed that such scales would help standardize grading among the various teachers of any particular school and make the comparison of the compositions written in any two given schools much easier. A number of composition scales were devised by a variety of authors, but they followed closely the original pattern established by Thorndike.

Although the composition scales showed great initial promise, they met with many difficulties in being accepted by educators and teachers. A flood of criticism which questioned their scientific

validity, the difficulty of their practical classroom application, and their attempt to measure the rather nebulous entity called quality all combined to defeat the composition scales.

There have been recent attempts to establish objective standards for English composition. These have taken the form of sample themes, much like those listed on the composition scales of the past, treated in a variety of ways to suggest points for remedial attention. These latest efforts have been collected into booklets by colleges, state education associations, and various individuals interested in the teaching of composition.

Because of the early scales and the recent efforts to make concrete suggestions for grading themes, concern has been generated about the lack of objective standards for English composition. This concern has illustrated the inequality of grade values among various teachers of composition and the pressing need for the clarification of all phases of writing done for academic credit on all levels from the grades through college.