RAPID IMPRESSION ESSAY EVALUATION BY SUITABLY PAIRED MARKERS

by

Gwendoline Pilkington

THE PROBLEM OF SUBJECTIVITY IN MARKING IN ENGLISH COMPOSITION AND THE EFFECTS OF USING A RAPID IMPRESSIONISTIC EVALUATION PROCEDURE BY SUITABLY PAIRED MARKERS

Thesis Submitted
by
Gwendoline Pilkington

as partial requirement for a Master of Arts degree in Education, Faculty of Education, McGill University

August, 1967

McGill University, Montreal,
Canada.

ABSTRACT OF THESIS SUBMITTED
by
Gwendoline Pilkington

as partial requirement for a Master
of Arts degree in Education, Faculty
of Education, McGill University.

THE PROBLEM OF SUBJECTIVITY IN MARKING IN ENGLISH COMPOSITION AND THE EFFECTS OF USING A RAPID IMPRESSIONISTIC EVALUATION PROCEDURE BY SUITABLY PAIRED MARKERS

This thesis investigated subjectivity in marking and suggested a technique whereby composition grades could be derived which would reflect minimum effects of subjective influences.

The hypothesis tested was that suitable pairing of markers will effectively minimize subjective tendencies, thus producing a 'truer standard' of assessment than can be derived by individuals.

The experiment involved eight Freshman composition lecturers who marked eighty essays, first as individuals and later as selected pairs. Markers were paired according to opposite marking tendencies - severity or lenity (means), and timidity and recklessness (standard deviations). A 'true standard' of marking was represented by the mean of the average scores for eight individual markings. Statistical results derived from analysis of variance (F-ratios) indicated that selected pairs produced significantly less variation in means and ranges, and a closer correlation with the 'true' scores, evidencing that effects of individual idiosyncrasies and biasses were lessened and quality of marking was improved.

FOREWORD

In his Memorandum I to The Marks of Examiners, Cyril Burt penetrates to the very core of the problem dealt with in this thesis. He speaks about the difficulty of arriving at a "true" evaluation of a student's performance, and he observes that there is no "external or objective criterion available" for this purpose. As a rule, there is only the single mark assigned to the report by the individual examiner. Yet, behind this figure, or letter-grade as the case may be, a number of real if subconscious factors have exerted a powerful influence on the examiner's decision. These are identified as "limited influences, personal influences, and accidental influences". Burt's observations touch on every aspect of the marking process so succinctly that the liberty was taken to quote them in full, rather than using a paraphrase which might prove both clumsy and less lucid.

> In most examinations the irrelevant factors that are likely to bias two or more examiners in the same direction will be fairly obvious. In essay papers, such things as spelling, grammar, handwriting, verbal expression, literary style, may count more with some examiners than with others. In subjects that involve questions of taste or doctrine rather than of fact or logical deduction - in art, literature, philosophy, for example, as distinct from languages, sciences, and mathematics - the particular school of thought towards which two of them share, may make one examiner's marks agree unduly with a second's, and seem positively antagonistic to those awarded by a third. But these are not the only tendencies that are likely to bias ... marking.

There are other influences more elusive and less easy to detect, because they are peculiar to each single examiner. In the main these are likely to be a matter of personal feeling or emotion rather than of intellectual attitude or taste. Generally, it may be said that every influence inducing a given examiner to swerve from the true mark operates, like other irrelevant and irrational influences, more or less unconsciously. But the less unconscious influences - those that are "fore-conscious" to borrow a term from Freud - are for the most part those which the examiner may share with other members of his group: they can, with a little effort and self-understanding, be consciously allowed for. The more personal influences are so deeply unconscious (in the psychoanalytic sense) that the plain man, no less than the psychoanalyst, realizes that it is always unsafe to trust to the judges' own powers of adjustment. As a surgeon is expected never to operate on his own relatives or even to diagnose their more serious complaints, so, instead of accepting the estimates of a master or tutor who knows his pupils or his students at first-hand and is therefore bound to have his prejudices and his favourites, we call in an external examiner or appoint an external examining body. Much the same holds true of subjects: if an examiner has taken some special problem for his own private research or his personal writings, he will tend to be unduly interested and influenced by the extent to which a candidate reproduces his teaching, quotes his books, or prefers the view of an opponent.

Finally, except in the most elementary of the abstract subjects - mechanical arithmetic, for example - there must, in every examiner's marking, be inevitably an ingredient of chance. By chance I understand the sum total of a very large number of very small influences, all irrelevant to the main purpose of the examination, and for the most part inseparable if not indefinable. Such miscellaneous influences as fatigue, lapse of attention, accidental changes of standard while working through a long series of scripts, will affect the marking quite irregularly if the order in which the papers are marked is unconnected with their merit (e.g., alphabetical order). A competent examiner will usually adopt some expedient for neutralizing these effects - for example, by going through the same scripts twice in a different order. But, even with the best precautions, the same examiner,

> unless he is guided by a retentive memory, will seldom give precisely the same mark on two successive occasions to precisely the same script. These fluctuations of the individual examiner about his own general estimate we may describe as his "random variation." 1

These points which Burt makes concerning subjectivity will provide the nucleus for the following discussion and will be examined in the particular light of the investigation undertaken.

---

[1]P.J. Hartog, E.C. Rhodes and C. Burt. The Marks of Examiners. International Examinations Enquiry. London: Macmillan and Co., Ltd., 1936, pp. 263-4.

PREFACE

The theory that every child should, as his birthright, have the opportunity to climb the academic ladder as far as his capabilities and aspirations allow has gained slow but steady acceptance over the centuries. In our own time, especially in the Western world, the progress has been perceptibly increased to the extent that the schools and universities are finding it difficult to cope with the influx of students. There are not enough classrooms, not enough adequately trained teachers, not enough competent administrators. One of the areas most vulnerable to this pressure of numbers is that of marking tests and examinations.

The number of individuals who can be successfully taught in a classroom is reasonably flexible, and modern technology has provided some new methods in the way of visual aids, closed circuit television etc., so that even larger classes can be accommodated. Help has been forthcoming as well in the way of computerized marking of objective examinations, at least in the fields of social and physical science. However, the problem of how to evaluate fairly and efficiently the mounting numbers of essay-type papers which are a necessary adjunct to the humanities has stubbornly defied a solution.

This failure cannot be attributed to a lack of educational researchers delving into the formidably complex area and, if not conquering it, at least publicizing the grave dangers of unfair or inaccurate assessment that it harbours. In fact, both Starch and

Elliott (1913) and Hartog (1936) concluded that even the marking of mathematics is not immune to the ills of subjective evaluation. However, such a subject does lend itself to scientifically designed objective tests, whereas there is still doubt as to whether ability in English literature or composition can be satisfactorily measured in this way. Efforts have been made both in England and the United States to design objective tests for those subjects, and these have enjoyed a rather mixed reception. Over the years the pendulum for and against them has swung from far right to far left to centre.

However, the one point that researchers seem to be agreed upon is that the traditional type essay examination, as marked by individual assessors for whom the candidate is a mere cipher on the brown envelope, is too prone to the fallibility of human bias and error to be accepted as valid.

Investigations in the 1930's in England, United States and France have clearly indicated that in too many cases students whose whole future was at stake were failed who should not have been, nor might they have been had their papers been read by another marker. Furthermore, where the awarding of scholarships on the basis of final examination results is concerned, the traditional system of marking is particularly iniquitous. As Professor C.W. Valentine pointed out some thirty-five years ago, public money directed towards scholarships very often is neither fairly nor wisely expended.[2]

---

[2]C.W. Valentine. The Reliability of Examinations. London: University of London Press, 1932, p. 37.

Recently, Morris A. Shirts, chairman of the education division of the College of Southern Utah, writing in the College Board Review, described the traditional assessments of final grades as the "seance" and the "meatgrinder" methods. He defines the former as one in which a "final grade is arbitrarily assigned through some mysterious, occult mental process little understood by modern science and even less by students." The "meatgrinder" technique "involves adding grades from tests, papers, class participation, and exams, then averaging and scaling them to a nice orderly curve." With these sentiments and with his sincere desire of finding "a grading system that would be fairer and more accurate,"[3] most educators could but echo a heartfelt amen!

This thesis will describe an attempt to find such a marking system. The investigation has focussed on the special problem of subjectivity in grading essays for a Freshman English composition course. This is of personal concern to the writer whose work as a lecturer and as an administrator of such a program involves the training of new instructors, especially in grading techniques, the establishing of general marking procedures and the setting of standards for the course.

In order to seek a solution to the problem, an experiment was carried out in which eighty essays were marked first by individual markers, and then again by the same markers, this time working in pairs. The objective was to determine whether the paired marking could produce an assessment of the students' work that

---

[3]Morris A. Shirts. "When College Students 'Contract' for Their Grades." College Board Review, No. 63, Spring, 1967.

would be less vulnerable to the "limited, personal, and accidental influences" of subjectivity as expounded by Burt in the Foreword.

The possibility of setting up such an experiment occurred to the writer in the winter of 1966-67 when participating in a pilot group investigation into the problem of assessment of High School examinations led by Dr. Norman France at McGill University. The idea of "Rapid Impressionistic Marking by Suitably Paired Assessors" (RIMSPA) was explored by the graduate students, and a series of marking experiments were carried on throughout the term. Some valuable insight into the problems of subjective evaluation was gained, and reference will be made to the results of the experiment in this thesis. Although the present research investigation will differ radically from the RIMSPA approach, it is in a sense an extension of the groundwork laid down by Dr. France and his class. Their very large contribution to this present study is gratefully acknowledged.

In order to carry out this experiment, the writer approached a group of seven highly qualified, experienced colleagues who, purely out of professional interest and self-development, and without remuneration, agreed to join with her in the marking sessions. Without their loyalty and support this venture would not have been possible.

To these, my confrères, who, with unstinting grace, patience and good humour contributed their precious time, energy and expert opinion; to my advisor, Dr. France, under whose initial inspiration the investigation was conceived and who steered my course of action; to Dr. J.K. Harley, who read and criticized the manuscript; to my husband, whose patience surely makes Job's pale by comparison and without whose help and encouragement the

project could not have come to fruition; and to his willing girl-Friday, who expertly typed the seemingly never-ending series of revisions and the final copy, I submit my humble but bountiful appreciation.

Gwendoline Pilkington
August, 1967.

TABLE OF CONTENTS

LIST OF TABLES

ILLUSTRATIONS AND MATERIAL OF SPECIAL INTEREST

CHAPTER I

STATEMENT OF THE PROBLEM, DEFINITION OF TERMS AND SOME RAMIFICATIONS

The problem which has been examined for this thesis is how to minimize the ill effects of subjectivity in marking essays and thus ensure that students will suffer very small chance that their grades will be unduly prejudiced by the particular biasses and random variations of a single marker. The meaning of the terms "biasses and random variations" will correspond to those used by Cyril Burt - "accidental, personal, and limited influences." Since these have been fully described in the Foreword, they should need no further definition. In order to reduce ambiguity to a minimum, the way in which the expressions "reliability and validity" are used throughout the discussions must be established. One commonly accepted definition is that "Reliability has to do with accuracy and precision of a measurement procedure....the extent to which a particular measurement is consistent and reproducible." Validity, on the other hand, "refers to the extent to which a test measures what we actually wish to measure."[4] A working definition, and one which can be accepted as applicable to this research, is the "Evidence of the reliability and validity of a written paper indicates that the marks

---

[4]Robert L. Thorndike and Elizabeth Hagen. Measurement and Evaluation of Psychology and Education. 2nd ed. New York: John Wiley & Sons, Inc., 1964, p. 160.

arising from it represent a more or less stable order of merit of candidates and that this order of merit is based on those aspects of a candidate's performance which are related to the objectives of the course of study which the candidate has taken."[5] One further qualification is that generally speaking wherever the terms "true or accurate assessments" are used, they are similar in meaning to "valid or reliable assessments."

The hypothesis to be tested was that the pooled judgment of two suitably paired markers, employing a rapid, impressionistic method, will render a much closer approximation to a "true" evaluation of a student's performance on an essay than could be obtained from an individual judgment.

The expressions, 'suitably paired markers', a 'rapid, impressionistic method', and 'true evaluation' must be made clear. They will be examined consecutively in some detail, and an attempt will be made to present sound bases for their use.

There is ample evidence recorded in research which suggests that part of the answer to the problem delineated is in having a number of persons read the same set of papers, thereby producing a composite judgment. This can be achieved either by taking a simple average of the marks given by different people to the same paper as the closest approximation to the true evaluation; or it can be derived by the group of markers getting together and somehow arriving at a consensus as to what the final mark should be. The latter suggestion has some

---

[5] Examinations Bulletin No. 3, "The Certificate of Secondary Education: An Introduction to Some Techniques of Examining." London: Secondary School Examinations Council, H.M.S.O., 1964, p. 20.

obvious limitations. Nevertheless, P.E. Vernon states that, "It was recognized in other fields of psychology that the random errors occurring in the estimates of a single judge are largely cancelled out when the judgments of a number of judges are pooled."[6] In educational research many experiments have been conducted in which the pooled judgment of several markers has been proven to render a mark which has both high reliability and validity. In fact, as a result of conclusions of these studies both in Britain and in the United States, the essay, which had been discarded as a test item on school leaving assessments because of difficulty in assessing it, was eventually returned to favour. Using the "pooled" or average mark, then, is one approach to reducing the ill effects of subjectivity. However, it is not always practical or even possible to have papers evaluated by a number of examiners. And if the second interpretation of the use of pooled judgment is taken - that of the group meeting and discussing what the final mark should be - it becomes even less practicable.

More recently the theory has been explored that instead of having several individuals read the same papers, equally reliable and valid results can be obtained by pairing markers whose standards of marking represent quite divergent patterns of severity or lenity and timidity or recklessness. Once these characteristics are identified and the markers have been paired, each one will mark the papers (using a rapid impressionistic method) and then the pairs will come together and compare results. They will re-mark any papers where there is a

---

[6] P.E. Vernon, ed. Secondary School Selection. London: Methuen and Co. Limited, 1957, p. 121.

wide divergence of opinion, and will discuss such papers carefully, each reconsidering what subjective elements have gone into the decision. It is recognized that each marker's impression of the essay will have been purely subjective and will have arisen from whatever main criterion is uppermost in his scale of values. Marker A perhaps has looked mainly for intelligently conceived ideas; Marker B, on the other hand, is more concerned with a well-organized precise statement of the idea, however weak its conception may be. By mutual discussion, the two markers must try to identify what biasses have seriously interfered with an objective appraisal of the overall performance, and one of them must be prepared to put these aside and reconsider his decision.

This, in essence, is what is meant by "suitable pairing." The statistical details of the process are discussed fully in the design of the experiment*. It must be emphasized here that where the pairs have not been able to agree on a final mark, they should not merely resort to averaging the two extremes. This would tend to cancel out the benefits derived from the technique of selecting only suitably paired markers. Rather it is suggested that in this event further opinions on the paper should be sought, making certain that the outside markers are not aware of the marks in dispute. The mark assigned finally to any highly contentious paper should be the average mark, representing the "pooled judgment" of all the assessors.

As Cyril Burt enunciated in the quote in the Foreword of this paper, there are many subjective influences at work when an examiner sits down to mark papers. It is usual in large-scale examining procedures such as matriculation, that all the candidates remain anonymous. This helps to eliminate some of the less desireable influences. But in a smaller situation such

---

* See pp. 59, and 65-68.

as that of a Freshman English Department in a university, for example, the instructors usually mark their own students' papers. In this case, the student faces all the barriers to accurate assessment - those that stem from personal idiosyncrasies and set standards of any marker, and those that arise from the marker's personal involvement with the student. In this experiment it was possible to examine only the effects of the first mentioned. It is suggested, however, that if an instructor and one equally experienced colleague can be suitably paired, we could assume that the student would benefit from a combined judgment of the lecturer who has guided him and is more personally and totally involved in his progress, and that of an outsider who is interested in his performance on that particular test. Furthermore, each marker would tend to check the other's unwarranted idiosyncrasies. At final examination time, this should ensure every student of as accurate an evaluation as present knowledge can permit.

The idea of using "suitable pairs" of markers has been explained and its applicability to the stated problem has been established. At least one further ramification of it has been suggested. The second term to be defined in the hypothesis is that of "rapid impressionistic marking".

A rapid impressionistic method of marking implies that the marker views the student's work as a whole, making no attempt to divorce the component parts of mechanics, content or style by giving them separate values. Examiners must school themselves (if they are not already in the habit of doing so) to make a very rapid appraisal of the essay and try not to dwell any longer on one part than another. The overall worth of the paper should be rapidly gauged, and a brisk pace* of marking should be maintained.

---

*For elaboration see p. 44, Wiseman's Instruction No. 2 to his markers; and p. 64, Instruction No. 1 given to markers in this experiment.

Evidence to support the practicability of this kind of marking is forthcoming from research into essay evaluation both in Britain and in the United States. Stephen Wiseman, in his 1949 study, had found that by having four markers give a rapid impressionistic evaluation of student essays, the results were sufficiently valid and reliable to warrant re-establishing the essay as a test device for grammar school selection purposes. Similarly in the United States, the College Entrance Examination Board re-instated the essay on their testing program after the results of research initiated by them indicated that if read impressionistically and independently by three readers the essay could give a reliable and valid estimate of the students' ability to write. These studies by Wiseman (1949), Godshalk, Swineford and Coffman (1966), and Myers, Coffman and McConville (1966) will be referred to in the review of the literature.

Recent research on the subject of marking by general impression indicates that it not only results in reliable and valid assessments, but even allowing for the fact that more than one reading of a script is required, it takes no longer than it would to do the same amount of marking by individual examiners using the detailed method.

The last term to be clarified is that of "true evaluation." The word true as Cyril Burt has asserted, represents only "an abstract or hypothetical concept".[7] However, it may be defined as a mark which has been arrived at by averaging a set of marks given to the same

---

[7]Hartog, Rhodes and Burt, p. 252.

paper by a group of experienced individual examiners marking under controlled conditions.

E.C. Rhodes, in his contribution to The Marks of Examiners, made the statement that "the average verdict of a number of examiners is better than any of the single verdicts; we might, therefore, use the simple average of the examiners' marks as an approximation to the ideal."[8] Earlier it was pointed out that the idea of diminishing the effects of random error which arise in a number of single judgments by pooling those single judgments is one that has been accepted in other areas of psychology. The word "true", then, as used in this paper will describe the average mark given to each script by the eight markers, this being the closest approximation to an accurate assessment of the paper that present knowledge of testing can permit. It assumes that the mean mark of several markers is more acceptable than any of the individual marks, and it also assumes that in taking that average mark as a true mark the extreme personal idiosynraasies of individual markers are largely cancelled out.

Although it is hoped that this experiment will point the way to a method of marking students' compositions that will significantly reduce the amount of undesirable variation in marking that stems from subjectivity and differing standards, it is recognized that perfect justice for the student is an unattainable ideal. Perhaps its attainment would not even be a worthy one; for, as the sages have taught, justice must always be tempered with mercy. Allowances must always be made for extraneous circumstances that quite unexpectedly arise and deter any

---

[8]Hartog, Rhodes and Burt, p. 186.

student from performing at his best on an examination. So that although the result may be correctly adjudged as being poor, if a plea is made for other consideration it should not be ignored or discounted.

In Bulletin No. 5, of the Schools Council series in Britain, the search for agreement on reliability and validity of internal examinations and the distribution of grades are discussed. The authors state that, "The problem is the same at all organizational levels.... Perfect reliability and perfect validity are unattainable in the present state of knowledge about methods of educational assessment." It is emphasized in the Bulletin that the only way to come to agreement is through discussion because, "...all examining is an exercise in human judgment about human behaviour in which there are no law-givers and no prophets; there is only a consensus of opinion patiently built up through the sharing of experience by many different teachers and other educators."[9]

Discussion; consensus; patience; and sharing of experience; and perhaps one should add a sense of proportion and humour; surely these are requisites for any group of markers, no matter what method of assessment they may employ.

Problems which loom large when closely examined in relation to the immediate present often can be placed in a truer perspective if set against related events of the past. For that reason it may be as well to glance at the historical background and the wider implications of the subject under study before reviewing the more contemporary aspects of it. This will be the concern of the next chapter.

---

[9]Examinations Bulletin No. 5, "The Certificate of Secondary Education, School-based Examinations, Examining Assessing and Moderating by Teachers." London: The Schools Council, H.M.S.O. 1965, pp. 1,2.

CHAPTER II

HISTORICAL BACKGROUND AND GENERAL CONSIDERATIONS

The origin of the idea of putting man to a test in which he must either succeed or accept the penalty of failure would be difficult to date with any exactitude. Certainly ever since Adam and Eve were expelled from the Garden of Eden for failing to act in accordance with the injunctions of their Creator, man has been subjected to tests of one kind or another, the results of which have either impeded or furthered his progress. Albert R. Lang reports that "as early as 2200 B.C., China had an elaborate national system of examinations, for the purpose of selecting public officials." The skill and fitness of youthful Athenian and Spartans were put to severe tests, both mental and physical, and the most famous of all teachers, Socrates, left us his pattern of question-response method of teaching and examining pupils. The reciting of the Christian catechism in the form of responses to set questions was a kind of recurring examination. In fact, says Lang, examinations have their roots in the ancient past and have been a continuing part of man's cultural heritage.[10]

With the advent of written communication, the art of testing an individual's worth or ability, especially in the academic sphere, became more sophisticated. It is

---

[10]Albert R. Lang. Modern Methods in Written Examinations. New York: Houghton Mifflin Company, 1930, p. 2.

possible that the first examination at university level, to be written rather than orally presented, was at Cambridge in 1702.[11]

The element of rigidity of procedure and standards in examining, and the role of the educator as a judge ruling upon the fate of candidates required to demonstrate their knowledge of endless facts learned by rote may be traced to the Jesuits, who began using written examinations to test their pupils as early as the 16th century.[12]

By the 19th century, rote learning had become a well-established educational practice. The gloomy task of memorizing and then verbalizing facts, either orally or in writing, has been graphically portrayed in many novels and biographies. Testimony to grim methods in 19th century English schools is provided in Charles Dicken's novel, Hard Times, where we meet schoolmaster, Thomas Gradgrind, "a kind of cannon loaded to the muzzle with facts" which he discharged into his hapless pupils with sublime confidence that if they could repeat them verbatim it was proof that they had learned something.[13]

This was fact veiled in fiction, but in Winston Churchill's biography, My Early Life, he records his first personal encounter with memorization and regurgitation of unintelligible subject matter as a criterion of academic worth. When, on his first day at school, his Form Master set him to work to learn by heart the First Declension of the noun, Mensa, he obediently complied. Then, in answer to

---

[11]Lang, p. 3.

[12]"Examinations," Encyclopaedia Britannica, VIII (1963) 931,32.

[13]Charles Dickens, Hard Times, Chapter II.

the Master's query whether he had learned it, the young Churchill replied with the simple honesty and wisdom so often mistaken for naiveté in children, "I think I can say it, sir." After apparently pleasing the Master by successfully repeating the exercise, he then incurred the wrath of the man by insisting on being told what it all meant.[14]

Churchill's comments on examinations are particularly sagacious, and echo the sentiments of countless pupils who have shared a similar traumatic experience at examination time:

> I had scarcely passed my twelfth birthday when I entered the inhospitable regions of examinations, through which for the next seven years I was destined to journey. These examinations were a great trial to me. The subjects which were dearest to the examiners were almost invariably those I fancied least. I would have liked to have been examined in history, poetry, and writing essays. The examiners, on the other hand, were partial to Latin and mathematics. And their will prevailed. Moreover, the questions they asked on both these subjects were almost invariably those to which I was unable to suggest a satisfactory answer. I should have liked to be asked to say what I knew. They always tried to ask what I did not know. When I would have willingly displayed my knowledge, they sought to expose my ignorance. This sort of treatment had only one result: I did not do well in examinations.
>
> This was especially true of my Entrance Examination to Harrow. The Headmaster, Mr. Welldon, however, took a broad-minded view of my Latin prose: he showed discernment in judging my general ability. This was the more remarkable, because I was found unable to answer a single question in the Latin paper. I wrote my name at the top of the page. I wrote down the number of the question '(1)'. But thereafter I could not think of anything connected with it that was either relevant or true. Incidentally there

---

[14] Winston S. Churchill. My Early Life, Chapter I, "Childhood".

> arrived from nowhere in particular a blot and several smudges. I gazed for two whole hours at this sad spectacle; and then merciful ushers collected my piece of foolscap with all the others and carried it up to the Headmaster's table. It was from these slender indications of scholarship that Mr. Welldon drew the conclusion that I was worthy to pass into Harrow. It is very much to his credit. It showed that he was a man capable of looking beneath the surface of things: a man not dependent upon paper manifestations...15

Fortunately, today, many of the more inhumane and futile school exercises (which passed for education) have gone the way of "Trial by Ordeal" (which passed for justice); and yet the haunting question does remain, How many promising candidates have not had the benefit of a "judge" who was able to "look beneath the surface of things: (one) not dependent upon paper manifestations"? The future of scholars is still too often dependent upon their ability to pass two or three hour examinations in subjects ranging from mathematics to history, sometimes under most inhibiting conditions. The obstacles to success, as represented by not just a passing grade but also by a high enough overall percentage of marks, remain quite formidable. The teacher, or an officially appointed marker, is still acting in the capacity of a judge whose quite arbitrary and possibly inaccurate decision can have far-reaching effects on the candidate's future. Such hurdles as Junior Matriculation, Senior Matriculation, Eleven-plus examinations have been accepted by the community, both lay and academic, with a sublime but not necessarily well-founded faith as being both necessary and requisite to the process of sifting out those worthy or unworthy of the opportunity to pursue higher academic studies.

---

[15] Winston S. Churchill, My Early Life, Chapter II, "Harrow".

But the sacredness of such testing apparatus has not gone entirely unchallenged. Certainly since the theory of "individual differences" among children was promulgated, searching questions have been raised about traditional testing procedures. When scanning the literature on investigations of examination and marking practices one can note a progression of disquiet among educational researchers which began even before the opening decades of the 20th century. P.E. Vernon reports that:

> Discussion of the inconsistency and element of change in teachers' marks came in the first place from the developing science of statistics, and for early evidence one must turn to the publications of the Royal Statistical Society. Edgeworth (1888, 1890), for example, writing of results obtained by Bryant and himself, drew attention to the errors attributable both to the idiosyncrasies of examiners and to the limitations of their sensitivity to differing degrees of merit. 16

These assertions did not go unnoticed. Among those who first took up the challenge in the United States were Starch and Elliott in 1912 and 1913, who pointed out grave inconsistencies in the marking of both English essays and mathematics papers,[17] and Hudelson in 1923, who expressed the urgent need for a more valid and reliable method of assessing English composition.[18]

In 1930, Lang discussed "The Traditional Essay Examination" and stated bluntly that "numerous...experiments could be cited to show that essay examination grades depend

---

[16] P.E. Vernon, ed. Secondary School Selection. London: Methuen and Co. Ltd., 1957, p. 115.

[17] Daniel Starch & Edward C. Elliott. "Reliability of the Grading of High-School Work in English." The School Review, XX (1912) 442-457. XXI (1913) 254-259.

[18] Earl Hudelson. English Composition Its Aims, Methods & Measurement. Bloomington, Illinois: Public School Publishing Company, 1923, p. 30.

more upon the scorer than upon the persons taking the examination." Furthermore, when the same person marks the same paper after an interval of time he will most likely assign a quite different mark.[19]

Two years after Lang's work was published, C.W. Valentine, in the The Reliability of Examinations, emphasized the need to discuss "where the strengths and weaknesses of examinations lie," and particularly to ascertain "how far we can rely on examination results."[20] He found in his own investigation that, in the marking of essays especially, extraordinary variations occur between the marks of different examiners, and he reports that "Even at university entrance stage there are suggestions made that the present tests are unreliable, in that they let through some who are merely crammed and 'spoon-fed', but who are lacking in general intelligence and especially in initiative and independent thought."[21]

The Marks of Examiners by Hartog, Rhodes and Burt, reflects the general concern of all the countries (England, France, Germany, Scotland, Switzerland, and U.S.A.) that took part in an International Conference on Examinations held in May, 1931. The investigations into examination marking on such a grand scale stemmed from the earlier findings of Professor Edgeworth in England, Starch and Elliott in the United States, and M. Laugier and Mlle. Weinberg in France.

In the preface to The Marks of Examiners, the writers stress the importance of putting the examination problem

---

[19] Albert R. Lang, p. 71.

[20] C.W. Valentine. The Reliability of Examinations. London: University of London Press, 1932, p. 9.

[21] Ibid., p. 26 and p. 37.

in its true perspective and they clearly define its dimensions:

> No element in the structure of our national education occupies at the present moment more public attention than our system of examinations. It guards the gates that lead from elementary education to intermediate and secondary education, from secondary education to the universities, the professions, and many business careers, from the elementary and middle stages of professional education to professional life...
>
> The examination system has grown to be an important element, not only in our education, but in the whole social system of our country; and the interest of many other countries in this matter is not less than our own. 22

The statement was directed to the English scene, but it contains the essence of universality, and, if anything, it is more pertinent in 1967 than it was at the time it was written over thirty years ago.

Although there has been only slight evidence of progress generally in making students' assessments more realistic, the work begun by these 1931 investigations has continued in England and elsewhere. In Britain, researchers such as Stephen Wiseman, P.E. Vernon, G.D. Millican, R.L. Morrison, and more recently D.R. Mather, N. France and G.T. Sare have penetrated the entire field of examining candidates and have made positive contributions by publicizing glaring weaknesses in methods formerly accepted with implicit faith, and in suggesting means of eradicating them. Much work has been done in the United States by educationists sponsored by the College Examination Board, for example, and in Canada progress is seen in some areas with the instituting of the

---

[22]Hartog, Rhodes and Burt, Preface, pp. ix,x,vii.

ungraded school and the trend towards accepting students into first year university on the basis of the school assessment and the various standardized achievement and aptitude tests.

It becomes more probable that Professor Valentine's prediction made in 1932 that "A revelation of great unreliability of examinations should tend to lessen the weight attached to examinations and so decrease their dominance in the general scheme of education,"[23] will finally be realized.

That the human species always resists change is a truism universally acknowledged; that educators are sometimes the worst offenders is not always so openly admitted. In 1965, when describing the background of the "written public examination system" peculiar to Great Britain, Mather, France and Sare discuss the various reports on matriculation examinations dating back to 1911 and culminating in the Beloe Report of 1960, and they somewhat wryly suggest that:

> In 1911 concern had been shown for the multiplicity of examinations serving the needs of a small minority. In 1960 concern had been shown for the multiplicity of examinations serving the needs - potentially - of the much greater number of average pupils.
>
> If the pattern is not to repeat itself in the year 2011 machinery must be set up which is in permanent and close touch with the schools and society, and which can react swiftly to the need for change and development in a dynamically changing society.
>
> Unless such machinery is devised it is at least theoretically possible that some young man now in the cradle may be called upon to prepare a report on the multiplicity of tests available to those pupils

---

[23] C.W. Valentine, Chapter one.

> of secondary school age now known as the Newsom pupils, in honour of the Newsom Report, Half our Future, prepared for the Central Advisory Council for Education." 24

This brief discussion of the serious general problem of student evaluation at all levels of education is but the backdrop for a narrower treatment of one crucial segment of the total area of concern - that of the grading of essays in Freshman English composition.

As has been stated earlier, the special purpose of the present investigation was to examine the ramifications of subjectivity in marking and to experiment with a technique for evaluating student essays which would reduce the chances of grading being prejudiced by the subjective judgment of a single marker. As large a segment as possible of the whole field of examinations and marking has been researched, but particular attention has been paid to any studies which might throw light opon the present investigation. To this review of the literature we now turn.

---

[24]D.R. Mather, N. France and G.T. Sare. The Certificate of Secondary Education. A Handbook for Moderators, London: Collins, 1965, p. 19.

CHAPTER III

REVIEW OF RELATED LITERATURE

In this resume, three distinct but nonetheless related areas of the subject of examinations and marking will be surveyed:

A. the historical development of the controversy between those who have favoured the use of subjectively evaluated traditional essay-type examinations, and those who advocated mechanically scored scientifically constructed objective tests.

B. the debate over a strictly detailed and rigidly structured marking scheme as opposed to a rapid impressionistic appraisal of papers.

C. the suggestion that assessment can be made more accurately and efficiently by always using more than one marker - preferably a suitably selected pair using the rapid impression method.

Since each of these topics impinges on the others, it is impossible to treat them as completely separate entities. Generally speaking, they will be discussed in the order enumerated with references being made to both British and American studies as they either paralleled or preceded one another.

A. The Objective versus The Essay-Type Examination.

The difficulties in arriving at a fair evaluation of students' abilities as reflected by their performance on examinations have long plagued educators - even prior to the advent of staggering enrollments in higher education facilities. Attempts to find a more satisfactory solution have ranged from judging students on the basis of a single essay-type examination to the use of purely objective testing procedures, depending upon the subject matter.

Albert R. Lang states that the idea of human testing has been evolutionary, dating back to ancient times.[25] A concomitant idea has been that, no matter what the test, the examiner has acted in the role of a judge meting out some form of punishment to the candidates who fail to meet required standards. From the point of view of students, Lang suggests that testing procedures "have a very personal and crucial meaning in the way of promotions, failures, conditions, scholastic standing, admission to high school or college, scholarship awards, school honors, and esteem by others."[26] These words of Lang's were written in 1930. How much more meaningful they are for the contemporary student who dwells in such a keenly competitive and much more densely populated academic world.

In the preface to Secondary School Selection, the problem as it pertained to the post-war British educational scene, is stressed. The editor, P.E. Vernon, states: "...in Britain, with its complex social history, the post-war conditions, and the very natural desire of parents to ensure within their means the best opportunities for their children, have helped to make the process of selection for secondary education a matter of genuine concern and, in many cases, of anxiety."[27]

As a result of this widespread "concern" and "anxiety", and because of the "spate of misleading and often emotionally-toned writing on the topic," an inquiry into its various facets was launched by the British Psychological Society with the intention of providing a

---

[25]Lang, Modern Methods In Written Examinations, 1930, p. 2.

[26]Ibid., p. 15

[27]P.E. Vernon, Secondary School Selection, 1957, p.7.

basis for "better-informed" discussions leading to educational reform, and also to indicate the direction of past and present research.[28]

In the resulting publication, a summary is presented of the evolutionary process of pupil selection for secondary schooling in Britain and of the origin and progress of the 'tug-of-war' between the use of the traditional essay-type examination and the "new-type" of objective testing as prescribed by such educational pioneers as Godfrey Thomson. The latter was anxious to prove that for various reasons many capable pupils were not being considered for the Junior Scholarship Examinations. In 1919 he devised an intelligence test (the Northumberland Mental Test) to be used henceforth in Northumberland schools in conjunction with the usual examination procedures in order to select pupils for free places in the county grammar schools. The new test was designed to eliminate the unfair advantages that pupils had who attended larger schools where they received special attention and training in the art of writing the scholarship examinations.[29]

By 1925, thanks to Cyril Burt who had been commissioned by the Northumberland Education Committee to devise new testing procedures, the ability of county children was being measured by the use of standardized attainment tests in English and Arithmetic. Teachers' acceptance of these tests was gained quite readily, but it was not until 1932 that the "new-type" examinations completely supplanted the traditional examinations.[30]

---

[28] P.E. Vernon, p. 8.

[29] Ibid., pp. 23, 24.

[30] Ibid., p. 24.

In a more general treatment of this subject of measuring attainment, P.E. Vernon states:

> Discussions of evidence as to the relative merits of new and old (testing methods) passed through various phases, and tests once described as 'new-type' are now more often called 'objective' or 'standardized', to distinguish them from the older examinations which were subjective in the sense that the questions asked and the answers accepted were determined subjectively by the personal decision of their author, and unstandardized in the sense that evidence was available as to the relative degree of success or failure in the answers of large samples of pupils of known age or ability. In the words 'objective' and 'standardized' there is thus epitomized much of the history of the testing movement. 31

Even before the scepticism of British researchers like Thomson was voiced, late 19th century criticism of traditional testing methods had centred on the fact that they were too prone to the ill effects of subjectivity and they did not test a sufficiently wide range of ability.[32]

Among the earliest American voices of protest were those of Daniel Starch and Edward Elliott. The results of their investigations into the "Reliability of the Grading of High-School Work in English" caused consternation and dismay among educationists and touched off a wave of widespread interest in researching the problem which has not subsided to this day. It was Starch and Dearborn who had exposed the wide discrepancies between marks awarded by various assessors within the same school system, and also between grades assigned to identical classes by different teachers. Starch and Elliott stated that:

---

[31] P.E. Vernon, p. 114.

[32] Ibid.

> ...The recent studies of grades have emphatically directed our attention to the wide variation and the utter absence of standards in the assignment of values. Dearborn pointed out in his investigation the large inequalities in the standards of grading employed by different teachers. Of two instructors in the same department one gave 43 per cent of his students the grade of "excellent" and to none the grade of "failure", whereas the other gave to none of his students the grade of "excellent" and to 14 per cent the grade of "failure". The difference is mainly attributable to varying standards in marking rather than to different abilities of candidates. 33

In these experiments the same papers were graded independently by several teachers and the test papers had been reproduced exactly as written, by photographing. The authors remarked that "The first and most startling fact brought out by this investigation is the tremendously wide range of variation....It is almost shocking...to find that the range of marks given by different teachers to the same paper may be as large as 35 or 40 points."[34] It seems even more shocking to realize that this study was done in 1912 and yet in 1967 there is still very little general recognition that a problem even exists.

A second study done by these researchers the following year produced evidence to disprove the old theory that marking in mathematics was more accurate than in English. It was shown, for instance, that a sample geometry paper was given an even wider spread of grades than had the two English papers used in the previous

---

[33] D. Starch and E.C. Elliott, "Reliability of the Grading of High-School Work in English," The School Review, 20 (1912) 442.

[34] Ibid., p. 454.

year's study.[35] It was concluded that whether or not a pupil was promoted in a subject or in a grade was largely dependent upon the teacher's personal whims. An interesting and amusing bit of evidence is added to this discussion by the writer's parent who attended a Grammar School in England around this time. He recalls that when the class of 11-year old boys were due for promotion, the only one who passed was the son of a candy merchant who used to bring candy to the teacher.

As a result of these and other investigations, a movement towards more scientifically standardized testing began. In Britain, the early work done by Thomson and Burt in standardized testing in Arithmetic and English for the purpose of grammar school selection finally had led to the formulation in 1925 of the Moray House Tests of Intelligence which, by 1954, were being utilized by three-quarters of the Local Educational Authorities. It is pointed out by Vernon that these tests "needed nothing beyond the competence of every teacher. Marking was in fact automatic and demanded no judgment on the part of the marker...."[36]

The trend in Britain on into the 1930's continued towards dropping the use of essays in English and long problems in Arithmetic, and substituting the Moray House standardized tests in English, Arithmetic and Intelligence. It was generally accepted that the marking of traditional essay-type examinations was "grossly unreliable", and there was a widespread feeling that in the use of standardized tests the answer to accurate selection of secondary school candidates was indeed solved for all time.[37]

---

[35]Starch and Elliott, "Reliability of Grading High School Work," School Review, 21 (1913) 257-8.

[36]Vernon, p. 25.

[37]Ibid., pp. 26-7.

It was inevitable that the pendulum would swing back, and gradually blind acceptance of the standardized testing panacea shifted towards sporadic rumblings of dissatisfaction and disquiet on the part of the teachers. The major complaints were that the use of standardized procedures had resulted in too much stress being placed on preparatory coaching and drilling of pupils in order to ensure success, and that it was exceedingly risky and unfair to judge the ability of a child on the basis of a one-stand test. Some educational psychologists were inclined to agree.[38]

Gradually, Local Authorities began to place "increasing weight on non-quantified data, particularly for children in the border-zone, thus implying that, while objective tests are of great usefulness in making decisions on the 'clear-accepts' and the 'clear-rejects', they are less useful in differentiating those in between." There was growing awareness of the importance of assessing "the whole child as a person, and, especially at the border-zone, to take account of individual quirks and circumstances."[39]

Undoubtedly one of the major reasons for these changes in educational attitudes in Britain was the 1944 Education act which dramatically raised the numbers of children staying on in school and entering the grammar schools. The situation that this created for the hitherto favoured and largely complacent middle and upper strata of society, whose children had long enjoyed wider educational opportunity, caused wide dismay and, in some cases, panic. The attendant problems have by no means been resolved.

---

[38] Vernon, p. 33.

[29] Ibid.

It is evident that over the years since written examinations became a well-established routine, the fallacy of blindly accepting one subjective judgment of a student's worth has been recognized. In an attempt to overcome this difficulty, especially as it applied to the marking of essays, various kinds of measurement scales were invented and tried out. In 1923, Earl Hudelson published a comprehensive American study which advocated a more "scientific" method of evaluating student themes. He stressed the fact that English composition is a very complex subject to assess and is more prone than any other to purely subjective interpretation of the examiner. Therefore, some form of objective measurement, such as scales, must be used if justice is to be meted out.[40]

Hudelson traces the origin and development of scales for judging writing ability, describing and discussing in some detail those of Rice, Haggerty and Van Wagenen, Hillegas, Breed and Frostic, Thorndike, and his own. Commenting on the use of diagnostic, analytical-type scales, in which separate elements of expression are measured, he states: "Specific qualities can and often should be measured separately; but when the general effect of written expression such as society is usually concerned with is to be judged, it must be considered in its entirety. In matters of appreciation the sum of all the parts does not necessarily equal the whole." He suggests that just as one cannot judge a painting by separately appreciating the pigments, the design, the frame, the canvas, so it is impossible to judge a piece of writing

---

[40]Earl Hudelson. English Composition Its Aims, Methods and Measurement. Bloomington, Illinois: Public School Publishing Company, 1923, p. 30.

by looking at its component parts. Like the painting, the composition "must be seen singly and seen as a whole....Imagination, which, after all, renders the final verdict upon art, defies mere analyzing; and composition is an art."[41]

In his discussion as to how scales should be used, Hudelson favours those which measure general merit, such as Hillegas, Thorndike and his own, rather than the diagnostic, analytical type which measure separate elements of writing. Nevertheless, he admits that as far as the teacher's rating is concerned "the results are virtually the same with or without the use of a scale for measuring general composition merit."[42] This does not seem to weight the argument in favour of using scales. Furthermore, he concludes that objective scales for general merit are not useful in "discriminating between sincere and pretentious composition. Neither do they materially affect a teacher's estimate of the relative importance of the various elements of composition."[43]

If this is so, and if the analytical scales fail to measure the worth of an essay as a unified piece of work, then what is the place of scales in evaluating composition? It is suggested that there is a definite need to "judge general merit objectively" for practical purposes required by spheres outside of the academic or school environment. On the other hand, Hudelson says, teachers require "devices for analyzing composition and diagnosing merits and defects for the purpose of improving instruction."[44]

---

[41] Hudelson, p. 55

[42] Ibid., p. 29.

[43] Ibid., p. 30.

[44] Ibid., p. 57.

He had stated earlier that "Objective devices for measuring composition merit are not, then, in spite of assumptions to the contrary, designed to improve writing directly." Rather their use is justified as being the best way for a teacher to assess improvement in expression, to test various ways of teaching, and perhaps improve his own methodology.[45] Thus scales are not designed to help the student write better, but rather to help the teacher teach better. And for this purpose, Hudelson recommends the diagnostic-type scale. However, he says that scales like the Van Wagenen Minnesota English Composition scale, "renders judgments confusing and difficult if, as is customary with teachers, the separate evaluations are combined into one general score." He had previously shown that the reliability of marking essays is reduced considerably when the assessment is analytical and he feels that neither Van Wagenen's nor a General Merit scale will likely yield completely satisfactory results. The remedy lies in using a scale which measures only one writing factor at a time, but it is then suggested that teachers experimenting with available scales and trying to rate composition elements in separate units find such a procedure confusing and exasperating.[46] This seems to be a rather circular argument for the use of scales.

What Hudelson seems to be trying to do in advocating a more "scientific" method of essay evaluation by using scales, is to find a way of reducing the appraisal of a work of art to some general formula which could be universally accepted. This is indeed a task for the Olympian gods.

---

[45] Hudelson, pp. 39, 40.

[46] Ibid., pp. 52,3.

It does not appear, to this writer at any rate, that the case for scales has been sufficiently justified by Hudelson. Support for this view is given by W.S. Monroe, who maintained that some of Hudelson's claims about the reliability of certain scales were exaggerated.[47]

While Hudelson may not have presented a clear-cut case for the use of scales in marking composition, his research did emphasize the difficulties in assessing the essay and the need to find a way of overcoming them.

This challenge was taken up in the United States by Albert R. Lang, in his investigations into marking procedures in the 1930's. He was particularly thorough in presenting authentic cases where "essay examination grades depended more upon the scorer than upon the person taking the examination."[48] He cites instances in which teachers being examined by an Ohio County Board received wildly divergent evaluations on their essays, from different sets of examiners. For example, "The arithmetic paper was graded by 55 examiners who gave it marks ranging from 60 to 99 per cent. The geography paper was graded by 52 examiners with marks ranging from 41 to 90 per cent. The theory and practice paper was graded by 52 examiners with marks ranging from 55 to 94 per cent." Perhaps one of his most disquieting (if somewhat wryly amusing) examples concerned the 1920 grading session of a group of professors at Columbia University. A model paper had been devised by one of the group, to be used to formulate his own general marking standards. Unfortunately, it became mixed in with all the other scripts, and when it was located it was found to have been awarded marks ranging from 40 to 90 per cent.[49]

---

[47] W.S. Monroe. "The Unreliability of the Measurement of Ability in Written Composition." Yearbook of the National Society for the Study of Education. 22(1923) 169-171.

[48] Lang, p. 70.

[49] Ibid., p. 71.

In proving quite conclusively that "unaided human judgment is fallible", Lang hastens to point out that such wide variation in individual standards (even in the case of the marker assigning a quite different grade to a paper the second time he assesses it) is not surprising. It is simply attributable to the fact that markers are human beings with the usual quota of human bias, and therefore there should be appended "no discredit to the scorer in their shortcomings."[50] But he does suggest that this factor must be admitted and he emphasized that the knowledge of students cannot be accurately gauged by a grade received on an examination which more likely reflects the leniency or severity of the marker or the degree of difficulty of the test, or even the physical or emotional state of the candidate at the time he was being examined.[51] In Britain, Mather, France and Sare struck the same chord when discussing the "Causes of instability in candidates' performances" on examinations. They suggest that "health, environmental stress, and temperament" are very real factors which might distort the picture a test gives of a candidate's ability.[52]

What Lang's extensive investigations underscored was the absolute necessity for a reappraisal of the whole method of examining and marking. He stated in the beginning of his 1930 publication that twenty years had already elapsed since a similar plea had been made, and in the meantime more evidence had been amassed to attest to the unreliability of traditional marking methods.[53] This cry resounds in the aforementioned 1965 British publication of Mather, France and Sare who lament that despite researchers consistently pointing out the hazards of objective assessment

---

[50] Lang, p. 72.

[51] Ibid., p. 77-9.

[52] Mather, France and Sare, p. 80.

[53] Lang, p. 13.

caused by "personal idiosyncrasies of examiners especially when examining creative work such as English essays, Art, Music....public examinations are still being conducted as if this research had never been done.[54]

Although Lang also made abundantly clear the weaknesses in the traditional kind of evaluation, he did not share the growing enthusiasm of some people of his time for scientific or objective-type methods of testing, especially in composition. He stated flatly that "Composition, by its very nature does not lend itself to objective measurement.[55] His faith in the value of an essay-type examination was not shaken by the revelations of studies such as Starch and Elliott's and others. He maintained that those who believed that the essay-type test was no longer a useful instrument and should be replaced by a standardized objective test were wrong.[56] He did, however, make several suggestions for improving essay-type examinations which included such comments as "A clearly stipulated marking scheme should be adhered to which allows set values for grasping main ideas, half-right answers, tidiness, organization etc. These should be marked one at a time."[57]

In order to maintain consistent standards, Lang suggested that a model paper containing correct answers be compiled. It is hoped that if this were ever carried out it might not share the fate of the model examination of the aforementioned history professor at Columbia.

Another major study of this same decade which dealt with "examinations" was that done in England by C.W. Valentine. Like Lang, Valentine felt that the examination

---

[54]Mather, France and Sare, pp. 134-5.

[55]Lang, p. 64.

[56]Ibid., p. 63.

[57]Ibid., p. 81.

system was "an inevitable part of the educational scene,[58] and he was also concerned about the unreliability of the marking procedures - especially of the essay.[59] He describes an incident in which a group of his own post-graduate students in Education in 1924 marked essays written on the same topic by seventeen children aged eleven and twelve. Eleven of the essays received a first class mark from some examiners and a failing grade from others. This type of inconsistency in results on identical papers is what undermines one's confidence in examination awards, Valentine asserts.[60] What particularly disturbed him was that the traditional system of marking very often allowed the wrong people into and kept the wrong ones out of universities.[61] He suggested that the student should have demonstrated his ability to write well consistently on all examinations, and not just on the English essay. If the markers of other subjects "will not pay adequate attention to English or if their estimates cannot be equated, it may be considered whether a special examiner in English composition may not see the papers in all or most subjects, to assess the candidates' capacity for writing English."[62] Whatever else such a practice might have to recommend it, one could hardly call it a very practical solution to the problem of subjectivity.

The general concern about unreliability of examination assessment was again voiced by John M. and Ruth C. Stalnaker at the University of Chicago in their 1934 study. They emphasized that "The significant criterion of a test item, whether the item is in essay or objective form, is not its reliability but its validity -

---

[58] C.W. Valentine, p. ix.

[59] Ibid., p. 30

[60] Ibid., pp. 26, 27.

[61] Ibid., p. 37.

[62] Ibid., p. 167.

the fidelity with which it measures what it is intended to measure." However, they point out that reliability is much simpler to measure than validity and the former is "important chiefly because its improvement may raise validity."[63]

The idea that essay tests, in comparison with objective tests, were highly unreliable was discounted by the Stalnakers. It is a false assumption, they claim, and one that has injudiciously discredited the essay-type examination.[64] Furthermore, an essay test, if carefully constructed and marked can be reliably read. The secret lies in "formulating essay-test questions so that a definite, restricted type of answer is required." They give several examples of the kind of question which will do this. For instance, one does not ask the student to "compare the writings of Corneille and Racine," but rather to make the comparison as to "(a) modernity, (b) use of action, (c) observance of unities."[65]

The Stalnakers maintain that whether questions are "general or restricted" the following marking practices must be observed if reliable assessments are to be made:

1. An objective scheme of scoring must be used.
2. The readers must first agree closely what the question is to be marked for; they must then analyze the ideal answer, assigning a certain number of points to each significant part of it.
3. Several papers must then be read independently by each of several readers to determine whether the scoring scheme is workable. Differences in scoring will lead to discussion and further elaboration of the marking scheme.

---

[63] John M. Stalnaker & Ruth C. Stalnaker. "Reliable Reading of Essay Tests," School Review, 42:(October, 1934),599.

[64] Ibid.

[65] Ibid., p. 601.

4. The official reading should not be started until close agreement is reached among readers and once this has been done it is still essential for readers to check at intervals with one another or with a standard set of papers in order to make certain that comparable standards are being maintained. 66

They further stipulate that there should be no corrections placed on papers, no discussion of them until all individual markers have assessed the papers. Students' identities should not be known to the markers. It is stated that a great deal of hard work and long hours will be required before a degree of conformity is reached and the more so if untrained personnel are involved and if the test has not been properly constructed. When agreement cannot be reached, some compromise would have to be effected.[67]

Certainly one can see how such an elaborate marking system would improve reliability of assessment, but, as with Professor Valentine's proposal that all types of examinations be evaluated by one composition teacher, the question arises as to its practicality, particularly where multitudes of assessments are involved.

With respect to the marking of composition examinations, the Stalnakers share the sentiments of other researchers that here is one of the most treacherous areas of academic assessment. They assert that, "Not only will several readers disagree widely in their judgments of such intangible qualities as originality, interest, and organization, but a single reader, judging the same paper on two different occasions, will not give it the same rating." They add that reliability and validity of such reading

---

[66]John M. Stalnaker & Ruth C. Stalnaker, p. 602.

[67]Ibid.

(i.e. general impression) is very low and its "worthlessness has been well demonstrated."[68]

More current research on the whole idea of "general impression" marking has clearly refuted this scathing condemnation of it by the Stalnakers. These studies will be discussed in Part B of the Review.

The Stalnakers' suggestion that a pupil's ability "to subordinate and coordinate material properly" can be better tested by a specially designed exercise has some merit, but perhaps such tests should be reserved for measuring progress during the term rather than on entrance examinations for higher education. One is reminded, too, of Hudelson's assertion that "Composition is, after all, an art," and as such cannot be satisfactorily judged by looking at the parts that make up the whole. The whole is much more than the sum of its parts. Nevertheless, the contribution of the Stalnakers' study is well recognized and their admonition that examination questions, whether of objective or essay-type should be carefully phrased is worthy of the attention of all examiners.

The debate over the essay versus the objective question is most crucial where selection of candidates for higher education is at issue. Philip E. Vernon, in his comprehensive work, Secondary School Selection, has outlined this problem as it affected Britain's grammar school selection. The American scene has been similarly served by Edward S. Noyes in his introduction to the 1966 Godshalk, Swineford, Coffman study, The Measurement of Writing Ability. Noyes states that the American College Examinations Board

---

[68]John M. Stalnaker & Ruth C. Stalnaker, p. 603.

has been faced with the problem of how to measure writing ability since 1901, when its first examinations were offered.[69]

He briefly traces the history of the problem as it evolved in the United States, suggesting that the difficulty inherent in assessing composition originated in 1910, at which time an hour-long essay was included in the three hour comprehensive examination in English. This test was used until 1940. The Board dropped the three hour comprehensive exams after World War II, and substituted for them a series of one-hour achievement tests.

The English test consisted only of a composition. The reliability of scoring this essay became a highly contentious issue, and in 1947 researchers such as Noyes, Sale, and the Stalnakers concluded that "the candidate's grade in English composition tended to depend far too much on which reader happened to have scored his essay."[70] In fact, the reliability of marking the tests was below that required to meet College Board standards.[71] After several unsuccessful attempts to find a reliable method of scoring an essay test, the Board examiners moved entirely to objective testing in composition.[72]

Subsequently, just as happened in Britain, the protests of American teachers against such a policy gradually brought forth a change so that finally "the examiners in English composition devised a semi-objective section called the interlinear exercise, in which students were asked to discover and amend by writing between the lines, errors

---

[69]Fred I. Godshalk, Frances Swineford and William E. Coffman. The Measurement of Writing Ability. Introduction by Edward S. Noyes. New York: College Entrance Examinations Board, 1966, p. iv.

[70]Ibid.

[71]Ibid., p. 2.

[72]Ibid., p. iv.

deliberately introduced into a passage of prose."[73] This device was accepted with considerable reservation from some quarters. It was said to be very poor teaching practice since students were presented with errors they might not even be likely to make themselves. Furthermore, the evaluation of such an exercise did not in any sense measure the candidates' capabilities in composition. As a result of the growing dissatisfaction, the College Board examiners, who were still inclined to believe in the new-type test item, instituted a series of studies to try to refute the criticism, and the experiment described in The Measurement of Writing Ability by Godshalk, Swineford and Coffman, is one of the important outcomes.[74]

In their own statement of the problem, the authors outline the various earlier studies to which they had recourse, tracing back to the 1921 findings of Hopkins who had "demonstrated that the scores a student made on a College Board examination might well depend more on which year he appeared for the examination, or on which person read his paper than it would on what he had written."[75]

They relate that after the 1945 Noyes and Sale and the 1947 Stalnaker revelations, the trend was towards objective or semi-objective tests. In the 1950's both Edith Huddleston and P.B. Diederich attempted to justify objective testing procedures. Huddleston's thesis was that if you measure verbal ability you are implicitly measuring writing ability and she felt that if SAT-verbal scores could be combined with objective English marks,

---

[73] Godshalk, Swineford and Coffman, p. v.

[74] Ibid., p. v.

[75] Ibid., p. 2.

together these would give the most reliable and valid prediction of ability presently possible.[76]

Diederich's 1950 study gave somewhat similar results to Huddleston's but he found that the English composition test provided slightly more weight to the over-all assessment than did the SAT score.[77]

A number of other studies cited by Godshalk and his colleagues (1966) had shown strong support for the validity of the objective and semi-objective English composition tests; but educators were not convinced. They still wanted an essay included in order to have students demonstrate their ability to write at college level. Therefore, the problem to be faced was how to evaluate an essay so that the results would be reliable and valid. This particular aspect of marking is inseparably linked with the topic of concern in Part B of the review.

B. Detailed— Analytical versus Rapid-Impression Evaluation of Essays

The debate over whether objective or essay-type questions are more acceptable testing devices continued both in Britain and in the United States, and the answer seemed to hinge largely on whether or not a way could be found to read the essay with some degree of reliability and validity. A very important sub-problem was raised by the research of Hudelson and others into the use of scales for measuring merit in composition. The question thus posed was, in order

---

[76]Edith Huddleston. "Measurement of Writing Ability at the College-Level: Objective vs. Subjective Testing Techniques." Journal of Experimental Education. 23(1954) 165-213.

[77]P.B. Diederich. "The 1950 College Board English Validity Study." Research Bulletin RB-50-58, Princeton, N.J.: Educational Testing Service, 1950.

to gain valid and reliable essay assessment, is the use of a rigidly structured detailed marking procedure preferable to that of purely subjective impressionistic evaluation?

Although the Stalnakers (1934) favoured retention of essay-type tests, they had presented a strong case for strictly objective detailed marking, and had stated that marking themes on general impression is "worthless". Earlier, Hudelson (1923) had advocated the use of scales as the only accurate way to measure "achievement and improvement" in English composition. In 1936, Steele and Talman had proposed a very complex detailed analytical marking procedure which was subsequently discounted, some five years later, by a Scottish inquiry conducted by R.L. Morrison and P.E. Vernon.

The latter two researchers concluded that even in as detailed a system of marking as the one prescribed by Steele and Talman, "Different markers interpret its rules very differently, and though many of the discrepancies cancel out when the total marks for an essay are summed, there are still considerable divergencies between markers." They further suggest that while the Steele-Talman objective evaluation process is highly reliable, it has not been proven to be valid. It tends to stress "the efficiency with which pupils express themselves" in the way of grammar and mechanics, and in the process ignores other more important facets of expression which many teachers consider paramount. For example, it overlooks the "more general aesthetic qualities which are admittedly intangible and difficult to assess."[78]

---

[78]R.L. Morrison & P.E. Vernon. "A New Method of Marking English Composition." British Journal of Educational Psychology, 11 (1941) 109-19.

Hence, after a very carefully set-up experiment in which the results of the Steele-Talman detailed marking apparatus were measured against those of a combined impression and analytic technique, the following points were concluded:

1. Those examiners using the Steele-Talman method found it "more trouble than conventional methods," and although it was easy enough to learn, they "doubted if any great increase in efficiency resulted from practice." Only one of the examiners was interested in using it later in his own work.

2. The method did not prove to be completely objective since the various rules set down were subjectively interpreted and applied.

3. Its reliability proved to be no higher than the impressionist-analytic evaluation and its validity could not be substantiated. 79

The marking of English essays in Special Place Examinations in England was one of the major areas investigated by members of the International Institute of Examinations Enquiry. According to Hartog, the goal was to compare "The discrepancies between the marks awarded by ten different examiners, all experienced in the marking of the Special Place Examination English Essay scripts", when an impressionistic method was used, with "the discrepancies which occur when they are marked in accordance with a detailed marking scheme."[80]

In this investigation, great care was taken to reach maximum agreement on marking procedure, and then typewritten copies of the 150 scripts were distributed to the examiners, who were asked to mark, with 100 as the maximum, 75 papers

---

[79]Morrison and Vernon, p. 117.

[80]Hartog, Rhodes and Burt, p. 117.

using impression method only. These papers were to be completed before the remaining 75 were graded according to a very detailed marking scheme. Test papers were chosen completely at random in each case. This assured that any discrepancies in the results stemmed from the method of marking rather than from any difference in quality between the two sets.[81]

It was found that the detailed marking produced higher average marks than did the marking by general impression, and that the former method seemed to bring the markers' assessments closer together. What they achieved, in effect, was to show that there was a larger margin of difference between the marks derived from impressionistic marking than those from a detailed assessment. In other words, impressionistic marking reflected more clearly the different standards of the markers. However, there was "no appreciable difference between the random variations of the examiners in the two methods of marking." (italics added) Furthermore, it was shown that "The averages of the two standard deviations for the two methods of marking are (thus) the same; in other words, the method of marking by impression and the method of marking by details produce on the average the same degree of discrimination between the merits of the different candidates, the same "spread" of the marks."[82]

In further experiments carried out in 1941, Hartog confirmed his earlier findings that there was considerable variation in individual standards of markers, but that analytic marking did not show any particular superiority over impressionistic marking.[83]

---

[81] Hartog, Rhodes and Burt, pp. 117-120.

[82] Ibid., p. 124.

[83] P. Hartog. The Marking of English Essays. London: Macmillan, 1941.

Another publication stemming from the International Examinations Inquiry is that of C.E. Smith's, The Marking of English Essays, published in 1941. This investigation was also concerned with the comparison between detailed and impressionistic marking, and the results of the experiment indicated that even when values for parts of answers are well defined, it is hard for markers to agree on the final grade. Furthermore, the fact that the personal bias of the marker greatly influences his final decision indicates that all marks are very much weighted by "general impression", and this occurs just as often in the marking of objective type questions as in the essay. Because of the impossibility of devising an equitable marking scheme for essays, the report recommended that such questions be removed from School Certificate Examinations.[84] Thus, partly as a result of mounting criticism and partly as a result of these findings by Smith and other members of the subcommittee, the essay tended to be dropped from the eleven-plus examination.[85]

An article published by B.D.M. Cast in 1940 threw some light on this controversy of impression versus detailed marking. In the experiment described by Cast a total of 40 scripts were assessed by twelve markers using 4 different types of assessment: (1) The individual's personal method (2) general impression evaluation (3) Burt's analytic method and (4) Hartog's achievement method.[86] It is beyond the scope of this thesis to give a detailed study of these marking schemes, therefore only results will be summarized here.

---

[84]C.E. Smith. The Marking of English Essays. London: Macmillan, 1941.

[85]P.E. Vernon and G.D. Millican. "A Further Study of the Reliability of English Essays," Part II. The British Journal of Statistical Psychology, 7 (1954) 65-73.

[86]B.D.M. Cast, "The Efficiency of Different Methods of Marking English Composition," Part II. British Journal of Educational Psychology, 10 (1940) 49-60.

Cast concluded that there is no way to ensure complete reliability in any method of marking English composition, therefore there should always be standardized instructions for the examiners who should be carefully schooled in their task. Her experiment indicated that Burt's analytic method of evaluation was the best according to the criteria used, but it was considered "laborious and unpopular" with the markers. She would not necessarily advocate its use over any other method. Rather, she shares Cyril Burt's own contention that although the analytic method is one of the most useful ways to train the novice examiner, nevertheless, "the intuitive or impressionistic method corrects many faults to which a crude, mechanical, quantitive dissection might inevitably lead....It allows us to judge the candidate's work by its general form or Gestalt, i.e., as a whole rather than as a mosaic of disconnected items; and thus permits us to grant full value to elusive and organic qualities that could scarcely be catalogued, or decomposed into separate portions."[87]

It was stated earlier that as a result of mounting criticism of its unreliable evaluation, the essay was discredited for use in eleven-plus examinations. P.E. Vernon and G.D. Millican report, however, that after Wiseman's 1949 experiment using a combined judgment of four markers employing a rapid impressionistic marking, it was considered that the results were favourable enough to suggest reinstating the essay as a gauge of the students' performance on the eleven-plus.[88]

---

[87]Cyril Burt. Mental & Scholastic Tests. London: King, 1921, cited by B.D.M. Cast, op.cit. 10(1940) 60.

[88]Vernon and Millican, p. 65.

Wiseman's attitude towards these two methods of marking is summed up in his comments in the 1949 study:

> Among teachers of English a constant battle is waged between supporters of analytic marking and those who believe whole-heartedly in general impression. Therefore, many researchers have attempted to compare these methods. Cast...showed a slight superiority of the analytic method...but this is more than offset by the results of Hartog's 1941 experiment. Since the time and labour expended on analytic methods gives no appreciable return in higher consistencies, might we not be better employed in using general impression, and in selecting markers who show themselves to be self-consistent? [89]

The necessity for obtaining maximum self-consistency in markers is stressed by Wiseman, but he feels that as long as experienced teachers are doing the marking, it is not important that their standards be consistent. In fact, he feels that this is not even a desirable situation, because "lack of high inter-correlation...points to a diversity of viewpoint in the judgment of complex material, i.e., each composition is illuminated by beams from different angles, and the total mark gives a truer "all-round" picture."[90]

It must be pointed out that Wiseman is talking about marking by "teamed impression" and not by individuals. His study did pave the way for the reinstatement of the essay in the eleven-plus examinations and the results of his experiment are worth enumerating:

1. Previous investigations have shown low reliabilities for essay marking.

2. There appears to be little difference in reliability between general impression and analytic marking, but it is important to note that the former is much quicker.

---

[89] S. Wiseman. "The Marking of English Composition in Grammar School Selection," The British Journal of Educational Research, 19 (1949) 204-5.

[90] Ibid., p. 206.

3. It is believed that general impression marking is more likely to yield valid results than will analytic methods.

4. By using four independent markers for 11+ compositions no more time and effort is required than for one analytic marking.

5. The efficiency of markers should be judged primarily by their "self-consistency."

6. Results show that a total mark re-mark reliability of over .9 may be achieved by this method. 91

Some of the instructions to the markers who took part in the Wiseman 1949 study are also worth reporting because they are similar to those given to the markers in the present experiment.

1. You are not asked to give a mark to the composition as a piece of English.

2. You are to give your mark on your impressions of the whole performance. Sub-totals for spelling, vocabulary, etc., are not to be used. You are expected to make up your mind quickly, keeping to a rate of about 50 per hour.

3. You must not look at the composition expecting certain things, and penalizing their absence.

4. If, when you are marking some such thoughts as "I'm giving rather high marks," or, "I haven't given many 13's lately," come to your mind, you must stop and exclude all such general ideas from your mind before you again begin to make your judgment. The judgment on each child must be an individual event, a placing of this child against a scale. Marks of other children are quite irrelevant.

5. Record your marks on the sheets provided. 92

---

[91] Wiseman, "The Marking of English Composition in Grammar School Selection," p. 208.

[92] Ibid., p. 208 (Appendix).

Wiseman gives credit to R.K. Robertson, whom he succeeded as Chief Examiner in Devon, for being the one who conceived of multiple marking by rapid impression in order to gain more reliability and efficiency in grammar school selection processes.[93]

The 1949 Wiseman study touched off a series of investigations over the next few years, some of which supported his conclusions and others which took issue with them. Douglas S. Finlayson in 1951, for example, disputed Wiseman's claim that the method of teamed impression marking in composition was just as reliable as the currently used objective tests.[94] Two years later H. Lamb conducted an experiment using both individual and teamed impression marking, and analytic-type marking. He stated that "By all criteria, teamed impression was best, analytic points second, and then individual impression."[95] However, another critic appeared in the person of J.D. Nisbet who discounted Wiseman's faith in the efficiency and economy of multiple marking by general impression. He suggests that using four markers (as Wiseman advocated) was perhaps overly costly.[96]

Three more studies were reported in 1956, under the title "Symposium: The use of Essays in Selection at 11+". D.M. Edwards Penfold was concerned particularly about the validity of essay examinations - do they measure what they

---

[93]Wiseman, "The Marking of English Composition in Grammar School Selection." p. 205 (appended note).

[94]Douglas S. Finlayson, "The Reliability of the Marking of Essays." British Journal of Educational Psychology, 21 (1951) 126-7.

[95]H. Lamb. "The English Essay in Secondary Selection Examinations: A Comparison of Two Methods of Marking." British Journal of Educational Psychology. 23 (1953) 131-3.

[96]J.D. Nisbet. "English Composition in Secondary School Selection." British Journal of Educational Psychology. 25 (1955) 51-4.

are purported to measure? She was not happy with conclusions drawn from previous studies in this regard, and tended to go along with Finlayson's criticism of Wiseman.[97] The second part of the Symposium concerned "The Predictive Power of the English Composition in the 11+ Examination." Its authors, E.A. Peel and H.G. Armstrong, found that the best test combination for prediction to aid in secondary selection was the use of English composition plus the Moray House Intelligence Test. They suggested that perhaps more than one composition must be written by pupils, and these must be marked by at least two markers. These results should then be combined with those of a Moray House Intelligence Test and then "we shall obtain a richer and fuller measure of each pupil's potentialities."[98]

The last word in the Symposium was given by Wiseman who defended his earlier findings, which had been criticized by Finlayson, Nisbet and Edwards Penfold. Wiseman repeated his assertion that teamed impression marking was not more costly than other methods, and it "seems quite clearly to be one way in which the essay can become 'respectable' as far as reliability is concerned, and its validity has been demonstrated under difficult conditions."[99]

Although a great deal of the more current research dealt with thus far has centred on the British educational scene, it was shown in Part A - the discussion of objective

---

[97] D.M. Edwards Penfold, "Symposium: The Use of Essays in Selection at 11+, I, Essay Marking Experiments: Shorter and Longer Essays," British Journal of Educational Psychology, 26(1956) 128-36.

[98] E.A. Peel and H.G. Armstrong. "Symposium: The Use of Essays in Selection at 11+, II, The Predictive Power of the English Composition in the 11+ Examination," British Journal of Educational Psychology, 26(1956) 163-171.

[99] S. Wiseman. "Symposium: The Use of Essays in Selection at 11+, III, Reliability and Validity," British Journal of Educational Psychology, 26(1956) 172-9.

versus essay testing - that the same problems and subsequent investigations into examining and marking have arisen in the United States. In Britain the question of grammar school selection, especially as manifested in the dissatisfaction with the eleven-plus examination, has been the prime concern of educationists. In the United States, researchers have been beset by a similar dilemma in finding the best way to sort out matriculation candidates seeking to enter various colleges and universities.

The Carnegie Corporation in New York has been instrumental in sponsoring research into all aspects of examining and grading techniques. An interesting point is made on the debate of detailed versus subjective evaluation of composition by Albert Kitzhaber, whose report of the Dartmouth Study of Student Writing was underwritten by the Carnegie fund. Kitzhaber described the great care that was taken to mark the essays analytically and to score them mechanically, using IBM machines etc. But in the final analysis, almost the only error which did not succumb to subjective interpretation was that of spelling. "With nearly all other errors and defects, the question of whether they are in fact errors or defects, and if so which particular kind, rested on the subjective judgment of individual English teachers."[100]

A recent investigation initiated by the American College Examinations Board is that of Godshalk, Swineford and Coffman which spans the controversy over objective versus essay-type questions and the analytical versus impression methods of marking. The authors of the study emphasize that they were aware even prior to starting their work that reliable and valid measurement of the essay

---

[100]Albert R. Kitzhaber, Themes, Theories and Therapy: The Teaching of Writing in College, New York: McGraw-Hill Book Company Inc., 1963, p. 45.

rested on (1) ensuring that there was a variety of topics available to the candidates so that they would not be faced with only one subject, for which they might have insufficient information, or no inspiration, and (2) making certain that the essays were appraised by more than just one marker.[101]

They point out that despite many attempts of research investigations in the 1940's to prove the case for strictly detailed analytical marking of essays, Coward, in 1950, had shown that the issues were by no means clearly resolved. In fact, it appeared as though "the efforts to improve reading reliability had been going in the wrong direction. The solution, it seemed, was in subjecting each paper to the judgment of a number of different readers. The consensus would constitute a valid measure of writing ability, assuming, of course, that the readers were competent." Further support for this theory was forthcoming from two more studies - one done in 1961 by Diederich, French and Carlton, and another in 1960 by Anderson. These indicated that the answer to the problem of accurate essay evaluation lay in the direction of multiple marking by impression. Both experiments proved that higher reliability could be obtained through using teams of markers, and Anderson's had called for "holistic"* judgments.[102]

It was this data that had provided the starting point for the experimental work of Godshalk and colleagues, in which marks from 646 test papers (made up of a short essay, 6 objective tests and 2 interlinear exercises) were analysed. The essay had been read "holistically" by 25 experienced markers, and a year later 2 of the essays were

---

[101] Godshalk, Swineford, Coffman, The Measurement of Writing Ability, 1966, p. 4.

[102] Ibid., pp. 4,5.

*The term "holistic" as used in the American studies corresponds with that of "rapid impression" used by British researchers.

re-read by 146 different markers. "The total of the 25 scores thus assigned became the criterion for evaluating the objective tests and interlinear exercises."[103]

A brief account of some of the results of this experiment is herewith presented:

1. The greater the number of essays written, and the more different readings they receive, the higher is the reliability of the grading. (This is in line with the findings of Vernon and Millican (1954) who concluded that "a combination of 7 essays, each marked twice provides a fairly consistent measure of English ability.")

2. The type of objective tests specially designed for the experiment had a high correlation with the established criterion - that provided by the aforementioned "total of the 25 scores."

3. The authors do not assert that this established criterion "was in any sense an ultimate one." They warn that students are usually under stress due to such things as imposed time limits which leave inadequate opportunity for revision, lack of dictionaries or other references etc. These factors "place a premium on fluency and ability to write correctly and with some style in a first draft. In actual life situations the writer is seldom under such sharp limitations." What the samples used in this experiment provided in the way of criteria was "as valid measure of writing under test conditions that can be obtained in a similar period of time."

4. It was concluded that, "The most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions or interlinear exercises in combination with objective questions.... When essay scores are combined with objective sub-test scores, they produce validity coefficients even higher than an interlinear exercise."

5. Finally, the authors suggest that their findings are very much in line with those of many British researchers, and that their "criterion measure" has high reliability and "permits relationships to be viewed in sharp focus." 104

---

[103] Godshalk, Swineford, Coffman, p. 39.

[104] Ibid., pp. 39-42.

What is particularly pertinent to this thesis is that in a later investigation, Myers, Coffman, and McConville (1966) verified the fact that the marks used to acquire the criterion in the Godshalk experiment were derived from strictly holistic reading of the essays.[105]

The more far-reaching implications of this important American study are summed up by Noyes in the introduction to the work.

> It is enough to say that checked against a criterion far more reliable than the usual criteria of teachers' ratings or school college grades, all but one of the item types currently used in the English Composition Test proved to be excellent predictors: that a very high correlation was achieved when for a typical one-hour test, two objective item types were combined with an interlinear exercise; and that a 20-minute essay - read, not analytically, but impressionistically and independently by three readers - contributed somewhat more than even the interlinear exercise to the validity of the total score. The combination of objective items (which measure accurately some skills involved in writing) with an essay (which measures directly, if somewhat less accurately, the writing itself) proved to be more valid than either type of item alone. This discovery may well have important implications for testing in subjects other than English composition. (italics added) 106

In actual fact, as a result of the Godshalk, Swineford, Coffman study, and of the subsequent one by Myers, Coffman and McConville, the essay regained favour and was certified for use on the College Board English Composition Test, in conjunction with other types of questions.[107]

---

[105] Godshalk, Swineford and Coffman, p. 40.

[106] Ibid., p. v.

[107] Ibid., p. 39.

Thus both in Britain and in the United States the use of the essay as a valid and reliable test of writing ability was vindicated, as long as it could be evaluated by a team of experienced markers using a rapid impression method.

There will always be voices raised against this kind of an appraisal of students' written work and the arguments against it can be forceful. The most recent publication of the Schools Council in England describes the Trial Examination marking experiment in written English in which the general impression method was employed, and some protests are recorded. Some markers "found that they could not themselves separate the marking of mechanical error from a general impression of the essay as a whole, nor could they understand how anyone could assess the essay without taking into account the degree of mechanical error so obviously present."[108] However, the final verdict of the experiment was that from the point of view both of reliability and validity, all three parts of the Trial Examination - that of the Essay, the Comprehension Test and the Literature Paper - had been very successfully marked by rapid impression. The Committee stated that with regard to the marking of Literature papers, the results of this experiment simply "...confirm the practice of many teachers who have used the method for many years."[109]

What we have been dealing with thus far is marking essays of students whose identity is unknown to the markers, and the idea of gaining substantial reliability and

---

[108] Examinations Bulletin No. 16. "The Certificate of Secondary Education Trial Examinations in Written English." London: The Schools Council, H.M.S.O., 1967, p. 10.

[109] Ibid., p. 15.

validity by using a team of markers and rapid impression methods has been supported and documented. That equally good results might be obtained by using suitably selected pairs of markers is a more recent concept, but one worthy of serious consideration.

C. The Method of Paired Marking

The idea of using pairs of markers was explored by Hartog in the second investigation of the International Examinations Enquiry. University Mathematical Honours Scripts were the focal point of the experiment. The goal was to "test the degree of consistency (a) of individual examiners, all experienced in the particular kind of examination, and (b) of pairs of examiners, similarly experienced, and acting conjointly but independently of the other pairs." Twenty three papers were marked by six different markers and then the same papers were "independently" revised by the same six people but this time working in pairs.[110]

It was found that although pairing markers did not greatly reduce the differences in the mean scores, nevertheless, it did have a significant effect on the ranges. For example, "the extremes for the pairs are 4 and 46 and the average 18.3, only a little more than half the average range for the six examiners (34.7)." When trying to place candidates in order of merit, there was less discrepancy when the markers were paired than on an individual basis.[111] Furthermore, the pairing of examiners did successfully reduce the amount of random variation introduced by individual

---

[110] Hartog, Rhodes and Burt, p. 148.

[111] Ibid., p. 150-1.

examiners, although some variation remained.[112] It will be remembered that in the previous experiment described by Hartog, in which detailed and impression marking were compared, there was no real difference in random variation between the two methods. However, here it seems that by pairing markers the amount of random variation was significantly reduced. Thus it was clearly shown that by using pairs of examiners a better general standard of marking is obtained and random variations due to individual examiner's personal idiosyncrasies are minimized. A warning is proffered by the authors that even when using pairs of markers, borderline cases should be given further consideration.[113]

In the previous Hartog experiment, it was English essays that were evaluated, and in this second investigation mathematics papers were assessed. Hartog concluded that it was no easier to find agreement in the assessment of mathematics papers than it was in any of the other subjects. This reaffirmed the very early findings of Starch and Elliott in 1913. However, the same conclusion could not be made with regard to the 1967 RIMSPA investigation which, as has been explained, provided the initial impetus for this thesis. In the RIMSPA experiment scripts from the subject areas of English literature and composition, history, modern languages, and mathematics were evaluated first by individual markers using rapid impression procedure and later by suitably paired markers using the same method. The results showed that the mathematics markers were able to demonstrate very high correlations both between individuals' set of marks and those of the group average, and those of the pairs of

---

[112] Hartog, Rhodes and Burt, p. 238.

[113] *Ibid.*

markers. They also showed a high correlation between marks of the pairs and those of the school assessor. (These were actual scripts which had already been marked by the class teacher.)

It is significant that one of the main differences between the RIMSPA experiment and that of Hartog's was that in the former the markers were paired according to their diverse marking patterns as derived from their means and standard deviations on a trial session.* There is no indication that the markers in the Hartog experiment were paired according to any particular rationale. It is possible that higher correlations in the RIMSPA pairs stemmed from the special pairing arrangement. Hartog showed concern that the averages were reduced so minimally: He comments that, "...the fact that in an examination of this kind two out of three pairs of examiners can differ by as much as they do in the case of Candidate No. 20, who is assigned 132, 123 and 169 marks, or of Candidate No. 4 who is assigned 186, 177 and 210 marks is remarkable."[114]

This may not really be so surprising if one stops to consider that in the pairing procedure it is quite likely that two markers were paired who represented the same type of bias, the same degree of leniency or severity in marking and similar degrees of recklessness or timidity. In this case, their combined mark would do nothing to bring them closer to their colleagues whose combined judgment might also be exhibiting a double manifestation of quite opposite marking idiosyncrasies. The point to be stressed here is that if pairing of markers is to have a salutory effect in the way of

---

[114] Hartog, Rhodes and Burt, p. 150.

*The pairing method is described on page 66 and 67 of this thesis.

providing a closer approximation to a true mark for the candidate, the selection must be carried out on some kind of controlled basis. Otherwise, the paired markers might simply confirm each other's prejudices.

Concerned with total problem of "Assessing the school differences" in the Certificate of Secondary Education examinations, Mather, France and Sare put forward the suggestion of pairing markers on the basis of preliminary training sessions in which sample papers would be corrected. Those whose characteristics of marking were farthest apart would be asked to mark as a team. "So by matching in pairs the most hard-hearted with the least... the differences can be limited and more consistent standard of judgment can be attained.[115] In this case it is only the levels of the markers, as measured by their mean scores, that are being contrasted, whereas in the RIMSPA investigation and in the present one, the characteristics of timidity or recklessness as measured by the standard deviations or fluctuations about the mean are also used in order to give a broader picture of the markers' standards. The question as to what constitutes these standards is not a simple one.

It was outlined in the Foreword of this paper that there are several influences at work when an individual sits down to read essays or examination papers. One stems from personal idiosyncrasies and set standards of any marker, and another arises from the marker's personal involvement with the candidate. Both are an integral part of subjectivity and both can operate either for or against the student's best interest. Where the candidate remains anonymous, the second factor mentioned is not present, but

---

[115] Mather, France and Sare, p. 140.

it is very much operative when a marker is evaluating his own students. Henry Meckel refers to this specific problem in his article dealing with "Research on Teaching Composition and Literature." He says,

> The teacher's treatment of his students' papers is related to questions about validity that arise in using essay tests to measure writing ability. Whenever the teacher grades a composition, he in a sense ceases to be a teacher and becomes a judge, often a harsh judge from the point of view of the student. Thus the teacher who reads his pupils' papers with an awareness of the needs of individual writers and the intent of guiding improvement in writing skill assumes a somewhat different role from that of the teacher who acts as an examiner seeking to measure composition skill with accuracy...
>
> The validity and reliability of measurement become of great importance in required courses in which students must earn a satisfactory grade.

A similar point is made in Bulletin No. 3 of the Schools Council in England. It states that "Marking for examination and marking for purposes of instruction serve two quite different ends. In the latter the aim is to provide a pupil with the kind of knowledge about his performance in a field of study which will enable him to improve his grasp of the subject and to make progress. For examination purposes, marking is aimed at producing a stable and valid order of merit...."[117]

Whatever the case, it is evident that the student is at the mercy of the multiplicity of "limited, personal,

---

[116]Henry C. Meckel. "Research on Teaching Composition and Literature," Handbook of Research on Teaching, N.L. Gage, ed. Chicago: Rand, McNally & Co., 1963, p. 987.

[117]Examinations Bulletin No. 3, "The Certificate of Secondary Education: An Introduction to Some Techniques of Examining." London: Secondary School Examinations Council, H.M.S.O., 1964, p. 20.

and accidental influences" expounded by Cyril Burt in the Foreword of this thesis. This is true in all examinations, but in a complex subject like Composition (which presupposes a reasonably wide range of general knowledge on the part of students, so that the main concern of the course is with technique rather than content) perhaps it is even more necessary that two viewpoints go into the assessments. Ideally these would be represented by two markers who had been suitably paired - one of them being the class instructor who in the words of Meckel, "had the role of guiding improvement in writing skill", and the other a fellow instructor who knows what the course goals are and "seeks to measure composition skill with accuracy," and is interested only in the performance on that particular test.

In Bulletin No. 5 of The Schools Council the idea of paired marking is discussed and the question is raised, "How are the pairs to operate? Should they work independently at first, only comparing their awards after they have all been given, or should they work together, discussing each script in turn?" In the present experiment, and in the RIMSPA one, the pairs worked together discussing each script which had been given a very divergent grade. The authors of Bulletin No. 5 say that "Experience may show that one or other of these methods is to be preferred, but at present there do not appear to be any strong general reasons for choosing one rather than the other: the matter is rather for personal choice."[118]

---

[118] Examinations Bulletin No. 5, "The Certificate of Secondary Education, School-based Examinations, Examining, Assessing and Moderating by Teachers," London: The Schools Council, H.M.S.O., 1965, p. 11.

Two final statements may serve as suitable notes on which to close this review of the literature. The first is a comment made by the authors of Bulletin No. 5 of The Schools Council who are discussing the pairing method used in their investigation:

> Although the original basis for pairing was that one was more severe and the other lenient, it has been found in the agreement trials so far held that the trial not only reveals lack of sufficient uniformity: in itself, it produces more agreement between moderators. In their subsequent work, the lenient became more severe, and the severe more lenient, not merely when they proceed by discussing each script in turn, but also when they first give separate awards and only discuss afterwards. In fact moderators, like the rest of us, live and learn. 119

The last word is from the March 10, 1967, News from the Schools Council which enunciates that English is almost impossible to evaluate objectively. On the other hand, "It is most in demand by society generally for a variety of reasons, as a test of literacy, as a qualification for a host of occupations, and a basis for further education."[120]

In this statement lies the justification for the years of painstaking research which have been outlined here, and the rationale for this thesis.

---

[119] Examinations Bulletin No. 5, p. 11.

[120] News from the Schools Council, London: H.M.S.O., March 10, 1967.

CHAPTER IV

DESIGN OF THE EXPERIMENT

The problem of how to gain a more accurate evaluation of a student's performance, especially in composition, has been revealed as one of considerable complexity. In this thesis the hypothesis was tested that if essays are evaluated by suitably paired markers using a rapid impressionistic method, the differences in standards between pairs of markers will become insignificant, whereas the differences between the individual markers will be significant at the 5% level. Thus, by using pairs of markers, thereby reducing the level of significance of differences, it should be possible to substantially minimize the chances of grades being prejudiced by the subjective judgment of a single marker.

The terms 'suitably paired' and 'rapid impression' have been defined previously, but the statistical rationale behind the pairing idea was not given at that time. It arises from the assumption that there are three main characteristics to the set of results of any marker. These are the level or measure of severity or lenity of the marker (mean); the range or measure of timidity or recklessness of the marker (standard deviation); and the conformity of the individual marker with other markers (correlation). It is with the first two characteristics that we are mainly concerned for the purpose of pairing. How these factors were measured, and how they fitted into the design of this experiment will emerge in the following discussion.

In order to test the hypothesis, it was necessary to assemble a group of markers who would be willing to undertake a series of evaluations of student themes, first as individuals and later as selected pairs. Experienced lecturers in a large compulsory Freshman composition course were canvassed for interest in taking part in such a venture, and, while many volunteered, the field was finally narrowed to just eight (including the author) - a group who were as homogeneous as it was possible to select. Briefly their qualities and qualifications are as follows:

1. All are mature, married women with a minimum of three years' experience teaching the course.

2. They teach at the same hours on the same days of the week, and as a result have ample opportunity to discuss teaching and grading problems.

3. They are all considered to be competent lecturers who are keenly interested in the course and in their students.

4. Their academic backgrounds range from a Master's degree in history to Bachelor's degrees in law, psychology and general arts. Two have followed all the course work required for a Master of Arts in Education. Some have High School Teacher's Diplomas and have taught at that level.

5. All of them expressed genuine interest in the experiment and were anxious to see how they would conform in their marking both with their colleagues and with themselves.

6. They are always concerned about giving their own students as fair an evaluation as possible, but clearly recognize many of the barriers to accomplishing this goal.

7. Their enthusiasm and interest are attested to by the fact that they gave of their valuable time and experience not only generously but also freely.

The selection of candidates was the next step. Since the students all write a number of class essays during the term, all that was required was to choose several classes that were writing such an assignment, and collect the papers for use in the experiment. It was decided to use four classes who met at the same time, and this would give approximately 120 papers from which to select a sample. In the end, due to absentees and to some scripts which did not reproduce well enough to be deciphered, there were actually 80 candidates whose papers were read.

These students represented a wide range of ability in written expression ranging from failing to potential A level, and they were from all faculties within the university - Arts, Commerce, Science, Engineering and Fine Arts. Some students were repeaters and a few were older than the average Freshman of seventeen, and therefore may have been more mature. This is a factor which could influence the marking of the instructor, but not the marker who had no knowledge of the student as was the case here.

The conditions of writing closely approximated those under which a normal class-assignment is written, the exception being that the students were told that the essays were to be used for experimental purposes. This was not expected to, nor did it, have any effect on the way in which they performed. They were asked to fill out a short questionnaire designed so that it was possible to remove the cover sheet on which their names appeared and to identify the assignment by a number code on the top right corner where the name would normally appear. Information concerning age, faculty and high school background was requested in case it might have some relevance or use to which it could be put at a later date.

The following instructions and considerations applied to the actual essays.

1. Students were asked to write an essay of 300 to 400 words on one of five selected topics. They had been studying the technique of persuasive writing during the term, hence they were asked to use that particular style on this occasion. Since all the instructors would have been dealing with the same technique, they would look for much the same overall type of performance on the essays. The materials with which they work - texts, calendar of lectures, visual aids etc. - are uniform throughout the department.

2. There was a choice given of one of the following subjects:

   a. Impact of Computers on the Business World.
   b. The Greatest Contribution of Science to Man.
   c. Man's Inhumanity to Man.
   d. Role of Women in Today's World.
   e. The Way in Which Religious Belief Affects Social Progress.

Giving a choice of five topics is in accord with the usual procedure for assignments in the course. While it has obvious drawbacks in adding several more variables, it was felt that to ask all the students to write on one completely arbitrarily selected topic would be unfair. Vernon and Millican concluded in their investigation on reliability of English essays that "the varying performance of candidates when writing essays on different topics is a source of inconsistency in the marking".[121] Hartog had discerned that different markers tend to mark higher or lower depending on the topic and therefore pupils may jeopardize their mark by an unfortunate choice of title.[122] This is no doubt true; on the other hand, as Godshalk et al. point out in their 1966 study, if only one topic is given, it could very likely result

---

[121] Vernon and Millican, p. 73.

[122] Hartog, Rhodes and Burt, p. 147.

in a student not being able to handle it at all. If it happened to be a topic about which the marker had a particularly strong opinion, this would be just as strongly reflected in the assessment. Wiseman, writing on the subject of "Reliability and Validity", says that one of two major problems connected with essay-marking is the question as to whether "the commonly-found variance between titles is due to marker-error or to the abler children tending to choose one title and the less able a different one?" He says that "an investigation by himself and Dr. Jack Wrigley shows the greater part of this variance to be due to children rather than markers - therefore, we can continue to give a choice of title."[123]

Another factor which entered into the decision to give the choice was that to ask markers to evaluate 80 essays all on the same topic would have put an unnecessary burden on them, and would have imposed a marking environment which was both foreign and unrealistic.

The samples of markers and students were selected, the essays were written, collected, put into alphabetical order, then coded so that no marker could identify a paper. The essays were all photographed so that each marker had his own set. A further advantage to this procedure, which was followed by Starch and Elliott in their 1912 study, is that the paper appears exactly as the student wrote it - the handwriting, with all the pitfalls it injects into marking, the mistakes which have been rectified by the student, overall appearance of the manuscript, and so forth. These are important features in subjectivity and they are absent if the students' work is reproduced by typing.

---

[123] S. Wiseman, "Symposium: The use of Essays in Selection at 11+. III. Reliability and Validity." British Journal of Educational Psychology, 26 (1956) 179.

Listed here are the instructions given to the markers; some notes of clarification have been added.

1. They were to give the essays a rapid, impressionistic subjective evaluation, making no attempt to separate the elements that go into effective expression. It was emphasized that they were to mark as rapidly as possible and to resist any tendency to compartmentalize the evaluation. The suggestion was made that approximately 25 papers read in an hour would constitute a reasonably brisk pace.

   (It is recognized that each instructor has her own habits of marking, and no attempt was made to impose a uniform method of handling the scripts. The writer agrees with Morrison and Vernon's comment that "...a general impression always involves a more or less vaguely formulated analysis of the qualities in an essay which are regarded as important.") 124

2. No corrections or comments were to be placed on the papers, although if markers wished to record elsewhere their impressions or momentary reasons for assigning a particular grade, they were urged to do so.

   (This was a justifiable procedure which tended to save time when the pairs came together to re-mark the 80 scripts. Where they had agreed on a mark, there was no need to re-read the paper, but where there were discrepancies, it was helpful if the individuals could recall readily why they had assigned the grade they did. Not all the markers did this, but where it was done it proved to be a useful, time-saving device.)

3. The grades were to be recorded on a separate sheet using the code number assigned, and the papers were to be assessed in accordance with the usual procedure followed in the marking of final examinations in the course.

   (This, traditionally, consists of using a rapid impressionistic method, rather than analytical marking which is mainly confined to routine assignments.)

---

[124] Morrison and Vernon, p. 111.

4. The grades assigned were to follow the same pattern used generally throughout the course.

   (This means using letter-grades ranging from F-failure to A-excellent, these approximating the following numerical values: F (54 and below) with a numerical mark always given to indicate the degree of failure; D- to D+ (55 to 64); C- to C+ (65 to 74); B- to B+ (75 to 84); A (85 and above) designating varying degrees of excellence. It is usual to indicate to the student whether he is on the borderline of a particular zone. For example, an F 54 indicates a borderline failure, whereas a D- 55 to 57 indicates a borderline pass. In the first case the student is alerted that he is progressing towards a passing grade, and in the second case there is an implied warning that although the writing warranted a bare pass the student must continue to work very hard to bring up the level of expression. Similarly in the C range, a C- of 65 to 67 is an acceptable grade but closer to a D, whereas a C+ indicates that the writer is approaching the B category. It is realized that such divisions are quite arbitrary, but they have proven workable under the given set of circumstances.)

5. Instructors were asked to write down any particular feelings they had about the project as it went along.

   (Although very brief, these comments were felt to be sufficiently interesting and insightful to include in the discussion of results in the next chapter.)

With the instructions made clear, the first set of 20 essays was given out. These were completed in a matter of days, and returned with a separate coded list of grades attached. The sets of 20 papers were immediately put back into the remaining sets of 60, making sure that they were distributed in a random order. This would help to ensure that the papers would not be remembered in the next round of marking.

With the results of this first set of marking, it was possible to compute the mean, representing an average grade assigned, and the standard deviation, representing the spread of marks about mean. These two measures were

taken to represent the two main characteristics of the markers - their severity or lenity, and their timidity or recklessness. From the picture thus presented, the pairing of markers was facilitated.

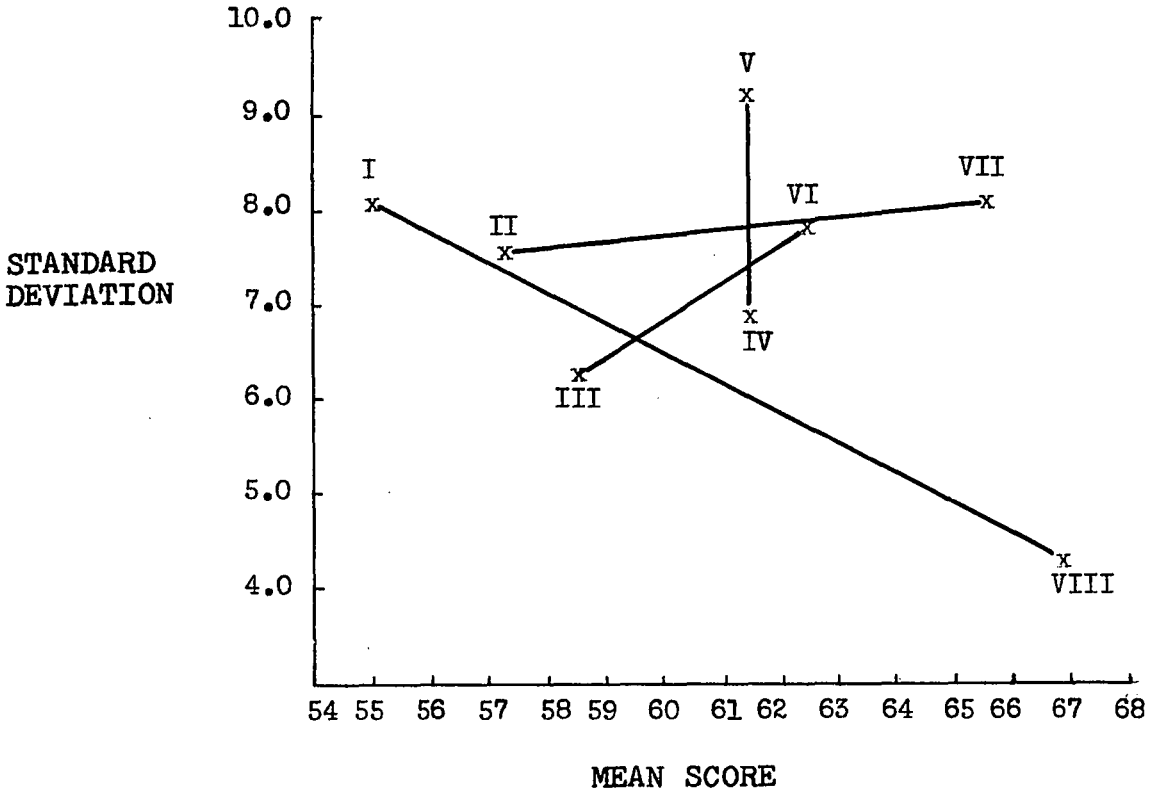
The mean scores and standard deviations of the first marking of 20 essays are shown in Table 1 below:

TABLE 1 - PAIRING STATISTICS (N=20)

	Mean Score	Standard Deviation
Marker I	54.8	8.07
II	57.1	7.60
III	58.5	6.30
IV	61.3	6.80
V	61.3	9.20
VI	62.3	7.80
VII	65.4	8.00
VIII	66.8	4.30
GROUP AVERAGE	62.2	5.10

It is interesting to note that in a telephone discussion, Markers IV and V, who have identical means, had compared notes. They felt quite elated to find how well their averages agreed until they discovered that where one had given a D, the other had given a B to the same paper, and that in no case had they assigned exactly the same mark to the same paper. It was a rather disturbing, and sobering revelation for each of them, and it underlines the fact that it is not a sound practice to pair markers only according to their mean scores on a set of papers.

CHART I, below, depicts the method of pairing by placing the Means on the abscissa and the Standard Deviations along the ordinate of the co-ordinate system. The Markers are designated by Roman Numerals.



In order to effect a pairing such that markers with as divergent characteristics as possible should become associated with each other, lines are drawn between pairs of points on the Chart in such a manner that they intersect each other as close to a common point as is possible. When selecting pairs in this way, there might easily be two or three feasible combinations; therefore, wherever possible, due care should be taken to match markers who have the best opportunity to carry on a compatible partnership. Such factors as potential personality conflicts, ease with which they may meet for discussion, similar teaching schedules, and so forth, should be taken into consideration.

Table 2 depicts the association of Markers as they have been derived from the Chart.

TABLE 2 - PAIRING OF MARKERS

Markers	Pair
II and VII	A
I and VIII	B
III and VI	C
IV and V	D

The next stage of the experiment was to have the instructors mark as individuals, the total set of 80 papers. The pairing statistics were not discussed with them at all, and no one knew at this point who would be paired with whom. The original 20 papers, as was mentioned, had been placed among the remaining 60, and due to this, and to the fact that at least three weeks had intervened, it was most unlikely that anyone remembered marks originally assigned to any one paper. Thus, the objectivity of the second marking of these papers was not jeopardized.

An interjection should be made here. The selection of markers for this trial was made on the basis of the author's personal knowledge of their professional ability formed over several years of association with them. However, such things as personal standards or consistency in marking patterns at best can only be surmised; even one's own tend to be elusive. Therefore, a way of measuring at least the consistency factor was deemed advisable. Wiseman suggests one such test to be used in selecting markers. He states:

> "The consistency coefficient obtained by a pure mark, re-mark correlation using the same marking method on both occasions, is the one single measure which is quite clearly a true consistency, and one which is closest allied to the normal concept of test reliability. This is the coefficient which

should be used first in selecting markers."[125]

Thus, when the marking of the 80 papers was completed the marks assigned by each marker to the 20 papers were extracted from the sets of 80, and these were used along with the original set of statistics to compute self-consistency co-efficients for each marker. The Pearson Product Moment Coefficient of Correlation* was used and the results are summarized in the following Table 3.

TABLE 3 - COEFFICIENTS OF SELF-CONSISTENCY (N=20)

	Mean Scores		Standard Deviations		Coefficient of Self-Consistency
Markers	First Marking	Second Marking	First Marking	Second Marking	
I	54.8	55.5	8.07	6.02	0.84
II	57.1	61.3	7.60	6.72	0.69
III	58.5	64.5	6.30	4.10	0.63
IV	61.3	63.9	6.80	7.52	0.79
V	61.3	62.1	9.20	6.96	0.55
VI	62.3	63.6	7.80	6.05	0.84
VII	65.4	62.8	8.00	6.99	0.82
VIII	66.8	61.5	4.30	4.85	0.37

The minimum level of self-consistency that can be tolerated in a marking environment must depend to some degree on the particular situation. Stephen Wiseman

---

[125] S. Wiseman. "The Marking of English Composition in Grammar Selection," British Journal of Educational Psychology, 19 (1949) 204.

* As described in J.P. Guilford, Fundamental Statistics in Psychology and Education. 4th Ed. New York: McGraw-Hill Book Company, 1965, p. 95.

stated that in multiple impression evaluation of 11+ essays in county of Devon where considerable experimentation in this field has been done, "A mark re-mark correlation of less than .7 cannot be tolerated... and such an examiner would be replaced."[126] In the present experiment, where the markers were suitably paired for the final assessments, the above correlations, with the obvious exception of Marker VIII, are felt to be acceptable. This will be further explored in the discussion of results.

After another period of three weeks had elapsed, the group met to begin the last stage in the experiment - the marking of the 80 papers by pairs. It was only then that the markers were told with whom they were to be working. After a brief discussion of procedure, the pairs dispersed to go over the individual marks. Where there was agreement on the letter grade, the paper was not discussed. The plus or minus values were not considered important, as long as the letter-grade range was not affected. In the case where one instructor had given an F 54 and the other had given a D 57, the paper was re-read and each marker gave reasons for her decision to either fail or pass the paper. By the mutual sorting out of various factors which had influenced the individuals, it was possible to decide which subjective elements were less desirable and should not have had a bearing on the assessment. For example, one marker might be prepared to admit that she had been unduly influenced by the presence of a number of comma faults, which always tend to unduly prejudice her evaluation of a paper. Having read over the paper again, she now decides that this factor had indeed

---

[126] S. Wiseman. "The Marking of English Composition in Grammar School Selection," p. 206.

over-ridden her better judgment and really the paper did merit a pass. Conversely, in another case, one of the markers may have been influenced by a student's attempt to treat the subject in an unusual way, and had tended to overlook the fact that the overall impression of the paper was so unclear that very little in the way of concrete ideas had been communicated. In this case, she would be prepared to admit that it did not really merit a pass.

After the first 40 papers had been gone over, the markers decided that having established a workable routine, they would profit from a respite before they tackled the remainder. It was decided to call a halt, and the pairs made arrangements to meet later to finish the job. Within a week all the papers had been re-marked and the results handed in.

An analysis and discussion of the results of the experiment are presented in the next chapter.

CHAPTER V

STATISTICAL ANALYSIS AND DISCUSSION OF RESULTS

Statistical Analysis

It will be recalled that the hypothesis to be tested by this experiment is that suitable pairing of markers will minimize effects of individual biasses and idiosyncrasies, and thus produce assessments significantly closer to the true mark than can be derived from individual marking.

In the following paragraphs the statistics used to describe the characteristics of the several markings, and the criteria used to assess the validity of the hypothesis, are summarized.

1. The average of the 8 marks given to each of the 80 papers was used to represent the 'true value' of each paper. The mean of these true marks was taken as the criterion against which all other means were tested for variance.

2. The mean value of any set of marks was used as a measure of severity or lenity of the marker (or pair of markers). The means of the individual marking and of the paired marking were tested for significant variance from the true mean using F ratios. Significance was measured at the 0.05 level.

   A significant reduction in variance from the true mean, when comparing the paired markings with the individual markings, was considered as evidence in favour of accepting the hypothesis.

3. The standard deviation of a set of marks will constitute a measure of recklessness or timidity of the marker (or pair of markers). A better standard of marking by all parties should result in the standard deviations of each set of marks approximating each other.

If the range of standard deviations of the 4 sets of paired markings was less than the range of the 8 sets of individual markings, then this was taken as further evidence to support the hypothesis.

4. Finally, each set of individual marks and each set of paired marks were correlated with the set of 'true' marks. (Pearson Product Moment Coefficients of Correlation were used).

   A correlation of paired marks with the 'true' marks, higher than that of the individual marks with the 'true' marks, was an indication of support for the hypothesis.

True Scores

The mean of the average scores for the 8 sets of 80 papers was 60.9. This represented the mean of the 'true values'.

Mean Scores

Table 4 records (a) mean scores of each of the 8 sets of individual markings with the mean of the 'true values', and (b) the means of the paired markings with the mean of the 'true values'.

Table 4 - Comparison of Means for Individual and Paired Markings with Mean of 'true values'.

	Individual Marker								True Values
	I	II	III	IV	V	VI	VII	VIII	
Mean Score (N=80)	56.1	61.0	62.8	61.5	59.5	64.5	60.6	61.5	60.9

	Paired Markers				
	A(II & VII)	B(I & VIII)	C(III & VI)	D(IV & V)	True Values
Mean Score (N=80)	60.7	59.9	61.5	60.1	60.9

A rapid glance at these figures suggests a significantly lower range of values for the paired means than for the individual means; the paired means range from 59.9 to 61.5, a spread of only 1.6; while the individual means range from 56.1 to 64.5, a spread of 8.4.

Computing F ratios on these statistics to test their significance results in the values shown in the next two tables.

Table 5 - Analysis of Variance (individual markings)

Source of Variance	Sum of Squares	Degrees of Freedom	Estimate of Variance
Between Markers	3390	7	483
Within Markers	27228	553	49.3

$$\text{F Ratio} = \frac{483}{49.3} = 9.8$$

Table 6 - Analysis of Variance (paired markings)

Source of Variance	Sum of Squares	Degrees of Freedom	Estimate of Variance
Between Markers	124.8	3	41.6
Within Markers	11450	316	36.3

$$\text{F Ratio} = \frac{41.6}{36.3} = 1.15$$

Considering first the F ratio produced by the individual markings, with these degrees of freedom, significance at the 0.05 level calls for an F ratio of 2.03. These statistics have produced an F of 9.8 indicating that a highly significant difference exists between the mean of the true values and the individual means. Indeed, this value of F is significant well beyond the 0.01 level (F=2.69).

The paired markings, on the other hand, show an F ratio of 1.15 with a requirement for an F of 2.62 at the 0.05 level of confidence indicating that under the criteria to be used, no statistically significant difference exists between the means of the paired markings and the means of the true values.

Here then is ample evidence of the credibility of the hypothesis.

Standard Deviation

Table 7 sets down the standard deviations computed from the 8 sets of individual markings and from the 4 sets of paired markings.

Table 7 - Standard Deviations

	Individual Marker							
	I	II	III	IV	V	VI	VII	VIII
Standard Deviation (N=80)	6.75	5.05	3.98	6.85	8.28	6.18	7.39	6.68

	Paired Markers			
	A(II & VII)	B(I & VIII)	C(III & VI)	D(IV & V)
Standard Deviation (N=80)	6.04	6.07	5.54	6.26

The range of the standard deviations observed for the individual markers is 4.39 compared to a range of 0.72 for the paired markers. Even when the two extreme values of 3.98 for Marker III and 8.28 for Marker V are excluded, the range of the individual marking reduces only to 2.34, a considerably higher value than the 0.72 observed for the paired markings.

It is interesting to note that the standard error of the lowest standard deviation obtained (5.54) from the paired markings is 0.43. At a confidence level of 0.05, this produces a range from 4.70 to 6.32 within which all observed standard deviations are seen to lie.

This indicates that the variation in observed values is of no statistical significance and that they represent a good approximation to the spread that should exist among these 80 papers.

It seems apparent, then, that the suitable pairing of markers does tend to neutralize the effects of varying degrees of recklessness among individual markers as suggested by the hypothesis.

It is noted that Marker VIII produced a standard deviation of 6.68; considerably higher than the 4.30 obtained when marking the test set of 20 papers - further evidence of this marker's variability.

Correlation

Table 8 tabulates the coefficients of correlation between each set of individual markings and the set of 'true' marks and each set of paired marking and the set of 'true' marks.

Table 8 - Coefficients of Correlation with 'true marks'.

	Individual Marker							
	I	II	III	IV	V	VI	VII	VIII
Pearson r	0.59	0.69	0.62	0.49	0.66	0.35	0.79	0.38

	Paired Markers			
	A(II & VII)	B(I & VIII)	C(III & VI)	D(IV & V)
Pearson r	0.88	0.85	0.89	0.90

Obviously, higher correlations with the set of true marks are obtainable from paired markers than from individual markers. It is also noted that the variability between pairs of markers is considerably less than that between individuals.

In the previous chapter, when discussing the point of tolerance for markers with low consistency correlation, it was suggested that much depends on the kind of marking that is being carried on. Attention was drawn particularly to the statistics of Marker VIII, who was identified as injecting a fair degree of inconsistency into her marking. It is noted here that when marking alone, she correlates only slightly with the true marks ($r=0.38$), but when combined with Marker I, whose correlation with the true marks also is somewhat low ($r=0.59$), their pooled judgment produces marks that show a high degree of correlation with the true ones.

This suggests itself to the author as being strong evidence in favour of the contention that the pairing of markers with opposing characteristics not only tends to neutralize their extremes, but has the very valuable effect of steering the markers toward a more realistic and accurate standard of marking. Furthermore, markers whose grading fluctuates so as to lower their self-consistency unduly, can, by working with a partner, be made more aware of the problem and do something to correct it. That is one justification for accepting the lower correlations recorded in Table 3.

However, with regard to Marker VIII, who is so far below all the others, it is interesting to note that before the experiment was concluded it became evident that, by temperament and personality, she was not a suitable candidate for this kind of marking situation. It casts no reflections on her worth as an individual; rather it makes recognition of the fact that not every person either enjoys or is suited

to group participation. Some find satisfaction and even security in it; others find irritation and feel threatened by it. In such cases, it is best to use another technique for keeping a watch on the subjectivity in their grading.

Summary of Statistical Findings

It has been demonstrated by this experiment that suitable pairing of markers can produce:

1. A considerable reduction in variance of mean values among markers. In other words, a more stable and more reliable assessment of the general ability represented by the papers in question.

2. A reduction in the range of deviation of various markers, i.e. a levelling of the degree of recklessness between markers.

3. Closer correlation with the true scores.

Discussion of Results

While by no means conclusive, the experiment provides strong evidence that this pairing technique does lessen the effect of individual biasses and produce an improvement in the quality of marking.

The variations evidenced by these markings are indicative of the difficulties in obtaining agreement among a group of qualified people as to what constitutes an accurate measure of ability in English composition. Obviously then, it is grossly unfair to permit one marker to pass judgment on a student particularly on borderline cases where passing or failing is at stake.

Furthermore, if opinions of the quality of work vary so much, then it suggests strongly that grading should be done by broad intervals rather than on finely divided ones.

Many researchers in the field have shown that the pooled judgment of a group of markers produces a more reliable measure of the value of a mark and is fairer to the student than any individual assessment. This study suggests, then, that a practical and accurate approximation to this pooled judgment can be obtained by suitably pairing markers, which will produce a significant improvement in the standard of marking.

It was found that once the underlying principles of the RIMSPA method were grasped and the markers had some practice in effecting them, they were able to identify their own weaknesses, and their strengths - i.e. the points upon which they were not prepared to bend. In short, it tended to make standards less prone to fluctuation. On the other hand, individuals were astonished at times to find their grade so different from their partner's, and when the paper was re-read, the often-heard comment was, "How could I have given that paper such a low (or high) mark?"

An essay which stimulated a great deal of general discussion - some of it quite heated - was one written on the topic, "Man's Inhumanity to Man". This happened to be one of the papers among the batch of 20 which were marked three times - twice by the individual markers and once by the pairs. It is interesting to examine the pattern of grades given to this essay. They ranged from F-failure to B-very good. Table 3 indicates the wide discrepancy in the evaluations of the paper.

TABLE 9 - Range of Assessment for Essay B-35

	Grades Assigned by Individual Markers							
	I	II	III	IV	V	VI	VII	VIII
First Marking	F53	C-	C-	D+	C+	B-	C+	B-
Second Marking	F50	C+	C-	C	D+	C+	C+	C

	Grades Assigned by Paired Markers			
	A(II & VII)	B(I & VIII)	C(III & VI)	D(IV & V)
Third Marking	C+	C-	C-	C-

It is apparent that the range had narrowed appreciably, and the extremes had been smoothed out. The consensus was that it was a good paper, falling into the C range.

In the discussion by the whole group as to why the various grades had been assigned, the marker who failed it said she had taken issue particularly with the treatment of the subject. The student had used Golding's theme in Lord of the Flies to illustrate his own views on "Man's Inhumanity to Man". The marker admitted that the expression did warrant a pass, but that the student had resorted to a rather superficial and cheap "gimmick" and had used psychological jargon in order to make the essay sound more erudite and polished than it really was. She objected to this attempt to throw her evaluation of the writing off-base. The fact that markers who had originally given the essay a B did reconsider and change to a C would indicate that there might have been something to their colleague's complaint that it sounded better than it really was. However the feeling of the majority was that the student had

used his outside knowledge intelligently and had given an excellent example to illustrate the topic. Furthermore, the overall impression was that it was a good essay, which with some polish might have been raised to the B or "very good" category. It would be interesting to know whether the "hard" marker, in this case, would have given such a low grade if the student had been in her own class. But this is another issue.

Two other manifestations of subjectivity are worth mentioning here. The treatment given to the topic, "Impact of Computers on the Business World", gave the writer considerable difficulty in assessment. When going over the results with her partner, she found she had been unduly hard on the students who wrote on that topic. What emerged from the mutual discussion was that low marks had been given to papers containing erroneous information about automation. The writer does not claim to be an authority on the subject, but it is her husband's field and, through discussions with him, her knowledge of it was slightly better than that of her colleagues. Her partner rightly pointed out that the student was not being tested on his knowledge of computers, but rather on his ability to express ideas. Furthermore, in a class essay there was no way in which students could make use of reference books or outside opinion on this subject, and therefore any errors in facts should not have counted heavily against them. The writer agreed and, wherever the expression warranted, the partner's higher grade was awarded.

The other case concerned a student whose handwriting was so poor that the majority of the markers' overall impression of the essay was also very poor - but quite unjustifiably so. Everyone who had given the essay a low mark on individual readings, conceded that when they took more time to decipher the writing they were inclined to raise the mark. No one had actually failed the paper, which

used his outside knowledge intelligently and had given an excellent example to illustrate the topic. Furthermore, the overall impression was that it was a good essay, which with some polish might have been raised to the B or "very good" category. It would be interesting to know whether the "hard" marker, in this case, would have given such a low grade if the student had been in her own class. But this is another issue.

Two other manifestations of subjectivity are worth mentioning here. The treatment given to the topic, "Impact of Computers on the Business World", gave the writer considerable difficulty in assessment. When going over the results with her partner, she found she had been unduly hard on the students who wrote on that topic. What emerged from the mutual discussion was that low marks had been given to papers containing erroneous information about automation. The writer does not claim to be an authority on the subject, but it is her husband's field and, through discussions with him, her knowledge of it was slightly better than that of her colleagues. Her partner rightly pointed out that the student was not being tested on his knowledge of computers, but rather on his ability to express ideas. Furthermore, in a class essay there was no way in which students could make use of reference books or outside opinion on this subject, and therefore any errors in facts should not have counted heavily against them. The writer agreed and, wherever the expression warranted, the partner's higher grade was awarded.

The other case concerned a student whose handwriting was so poor that the majority of the markers' overall impression of the essay was also very poor - but quite unjustifiably so. Everyone who had given the essay a low mark on individual readings, conceded that when they took more time to decipher the writing they were inclined to raise the mark. No one had actually failed the paper, which

possibly indicates that because the writing was difficult to read, the inclination was to give the student the benefit of the doubt. Herein lies one of the dangers of rapid impression marking, but perhaps it can be minimized by the use of the paired assessors.

In another instance, one marker had given a paper a B which had earned a D from the partner. After reading over the paper, the first marker could not understand why she had been so impressed with this essay because it now seemed very weak and inept. Two factors may have been operating here. One, she had been in complete agreement with the student's point of view, and two, she had read the script immediately after a series of very poor ones. This one, by contrast, looked quite polished. She agreed, however, that her initial impression had been very much off-base, and the D grade of the partner stood.

These are but a few examples of the effects of subjectivity in marking composition which arose from this experiment. They seem to the writer, and to her colleagues, to demonstrate the efficacy of using the RIMSPA technique. However, no single experiment is ever conclusive and it is hoped that further verification of the hypothesis will be found. This investigation dealt only with the problem of subjectivity as it affects examiners assessing anonymous scripts. What must be given equal attention is how it operates when the students are known to the markers. In order to look into this, it is proposed that the selected pairs of instructors mark each others students' final examinations. A supplementary investigation of this nature is planned for the future. It is also the intention of the group to continue working together as pairs throughout next year - both for the marking of at least one or two sets of assignments and for the mid-term examinations. The writer intends to observe and record the results with the hope of further exploration of the hypothesis.

It was thought that various comments of the markers might serve as a bridge between the purely mathematical and the more humanistic aspects of this investigation.

Specific Comments of Markers on Subjectivity

> I feel that knowing that these grades will not go against the student in any way has influenced my marking of these essays. I have given the mark that I think each paper is really worth. In an examination situation, or even in term marking, I might be more lenient.
>
> I had a feeling that my marks were higher when I wasn't making corrections.
>
> I find myself giving a low mark to a person who sets out to convince me of something and does not succeed, or to one who tackles a subject from an absurd point of view.
>
> It is possible that a marker might have a special prejudice against a certain type of error, for example, comma splice or poor pronoun sequence. The discovery of one or more errors of this nature might make him lose sight of the over-all value of the whole piece.
>
> A weak beginning might be forgotten in a longer piece, but a weak conclusion would certainly affect the mark.
>
> I find the two worst enemies of my objectivity in marking papers are my personal moods and my pet peeves. In a depressed mood I am likely to be more severe about mistakes than when I am in an expansive frame of mind. When those mistakes are my pet peeves I have to fight prejudice against the writer who made them, no matter what my frame of mind. The hard thing is that the more I mark, the more obsessions I develop. Being aware of these problems helps me combat them, but I don't think I always win.

These honest, extemporaneous remarks may be a reminder that, even in a society overshadowed by computers and other complex machines, whatever are the quantitative elements of any problem, in the final analysis, it is the qualitative ones that govern the results.

CHAPTER VI

SUMMARY AND CONCLUSIONS

This thesis has focussed on the difficulty of accurately assessing examinations in general and composition in particular. The special problem investigated was that of subjectivity in marking English essays and the effects of using a rapid impressionistic evaluation procedure by suitably paired markers. Borrowing a term coined in a pilot investigation carried on under the direction of Dr. Norman France, this method of marking is referred to as the RIMSPA approach. (rapid, impressionistic marking by suitably paired assessors.)

In the pilot study, scripts from several subject areas were evaluated by specialists in those fields. The evidence of that experiment showed that the RIMSPA method had resulted in great improvement in the quality of marking in all subjects (except English literature and composition) and it would lend itself very well to high school leaving assessments. The failure of the paired English markers to reach a better result is attributable to several factors mainly related to the unsuitable physical set-up for marking sessions. These limitations were not operative in the present experiment, the results of which indicated that the RIMSPA method opens another avenue of approach to more efficient and accurate measure of writing ability. They suggest that not only might students benefit from the practice, but the markers may also gain insight into their own strengths and weaknesses in evaluation. These two factors could lead to a general improvement in the

standard of marking - a prospect of considerable interest to educators, whether they are administrators or teachers.

After the writer became aware of some of the ramifications of subjectivity which came to light in the RIMSPA experiment, the possibility arose of doing a similar study among students and lecturers in a Freshman composition course at a local university. The latter is one of the few institutions, if not the only one in Canada, which separates Freshman English into two distinct credits of literature and composition. Only the latter is compulsory for all students, regardless of faculty, and its enrollment usually approximates 2500 students who are taught by one full-time and 45 part-time instructors. It was from this 'universe' that the samples for the experiment were drawn. The writer, who is the only full-time lecturer on staff and acts also as one of the administrators of the program, is responsible for training of new instructors, for helping to set standards and for formulating grading procedures. Needless to say, a very large concern of the department is trying to maintain some kind of consistency in standards of marking in order to ensure the students of fair treatment, especially in their final assessment. This, as has been illustrated, is a complex, even hazardous, venture generally; but the writing and the correction of compositions are such intensely personal activities, that once the lecturer comes to know his class as individuals (and in small groups of 30 or less this point is quickly reached) his chances of being able to remain objective in his judgment of the work are greatly diminished. This student has a likeable personality; that one is unpleasant. This one tries very hard but his results are poor; that one is careless and lazy, but he writes with some flair. This one has some disability or deformity which elicits the instructor's sympathy or pity; that one

always has a bored or smug expression on his face and constantly whispers to his neighbour. The list of personal idiosyncrasies of both students and markers is infinite in variety, and it is these random elements in marking which upset standards and prejudice the students' chances of receiving a reasonably accurate assessment of their ability. When, as in the case of a compulsory composition course taken at university level, the students' progress towards a degree could be unjustly impeded by a low or failing final grade, the situation demands attention, and new approaches must be explored.

The way the problem is presently approached is that instructors who are experienced and who are considered to be dependable, steady markers, grade their own examinations, then hand them in to the office with any doubtful papers placed on top of the pile. These, for some purely subjective reason, the instructor has not felt confident to mark fairly. Hence, two or three other instructors are asked to give a quick reading of the paper, and after general discussion among the markers, a consensus is reached and the resulting mark is assigned. The papers of new instructors, and others who from past marking performances are known to be erratic or unduly prone to personal bias for or against students, are entirely spot checked by a small group of experienced people who, individually, go over each set of these papers as they come in. If grades given do not seem to represent a true evaluation, or if the pattern of marks awarded to assignments seems unusually out of line with the way in which the student has performed on the examination, there is consultation with the instructor and he may be asked to change the grade.

It can be appreciated that with such a high enrollment of students and a fairly large staff of part-time people, this can present an almost overwhelming amount of reading for the small full-time teaching and administrative staff. In fact, as the numbers of Freshman increase each year the load approaches the impossible. Aside from this difficulty, the fallacy of the present system is that there is every possibility that the second reading, whether it is by request or design, may be done by a marker who merely confirms the first one's marking idiosyncrasies, thereby compounding rather than reducing the students' chance of getting an inaccurate assessment.

Hence, the writer saw in the RIMSPA experiment a glimmer of hope - a method whereby instructors could be suitably paired at the onset of term, and could possibly work together both during the year and more especially at exam time. Their combined judgment, arrived at by mutual discussion and not just by averaging of diverse marks, might reduce the students' chances of having grades unduly prejudiced by a single marker.

The idea was conceived and then was translated into action as described in Chapter IV of this thesis.

The results of the experiment clearly supported the hypothesis that the hazards of subjectivity can be reduced by using the technique of suitably paired markers and rapid impression. Whether the judgments rendered by this method are more reliable or valid will always be subject to the criterion against which the measurement is made. As Cast asserted, "All methods of marking English composition contain a large element of unreliability. Yet its amount can evidently be greatly reduced by standardized instructions and by training examiners."[127]

---

[127] B.D.M. Cast, p. 59.

The RIMSPA technique meets both these requirements. It should lead to an increase in efficiency of operation in a course such as the one described in this experiment.

There is no doubt that the present method of dealing with subjectivity, which calls for the reassessment of the majority of examinations by a small committee in order to spot check for erratic evaluation, is tremendously time consuming and costly. Rapid impressionistic multiple marking was shown by Wiseman (1949) to be no less costly in time or money than detailed analytical marking by individual markers. His system calls for a team of four. It is conceivable then that the method of using only two suitably matched markers could represent a real gain in overall efficiency and economy.

The writer does not claim that this experiment conclusively proved the universal applicability of the method; however, the evidence does indicate that it is well worth trying in a large compulsory Freshman English course, and as the RIMSPA pilot investigation revealed it could be useful in external assessment of secondary school leaving examinations or in departmental internal marking situations.

Any experiment is always subject to the limitations imposed by human fallibility, and this one was no exception. It was shown earlier that one marker in the selected group of eight turned out to be unsuited by temperament and personality to working in this way. However, as Mather, France and Sare have underlined in their discussion concerning paired examiners, the trial tests can serve as a training vehicle for all markers and can also be useful in identifying those who are not suited to the technique. They also emphasize the important feature of multiple evaluation that, "Once the differences of view have emerged and the criteria have

been externalized, it is possible to work towards a consensus of opinion and to decide what criteria we are looking for and how we are going to weight them."[128]

One of the most striking results of this experiment was how much the instructors learned about the hazards of subjectivity generally, and about their own attitudes to marking specifically. They considered it a rather sobering but nevertheless profitable experience - from the point of view of what they learned about themselves and what they learned from one another. It is felt that both new and experienced instructors could profit from a similar marking session.

The proviso must always be made, however, that certain people by nature are not 'negotiators', and they react better to having an impartial committee check their results. It is thought quite practical to suggest that all composition markers be given a trial test so that their marking characteristics could be established. Those who wish to work with a partner could do so; those who are not so inclined could resort to other arrangements made for keeping a rein on subjectivity. In this event, as long as patterns of marking can be identified, even roughly, there will be some assurance that when papers are re-read the students will have the benefit of two opposing sets of marking idiosyncrasies which will be carefully weighed. Where there is deadlock as to the final decision, the paper should be given at least a half a dozen further readings, and the average of these marks should dictate the grade.

---

[128] Mather, France and Sare, p. 139.

It must be reiterated that in paired marking there should be no resort to simply averaging differences. This negates the whole point of suitable pairing. The pairs must discuss and mutually agree on what are the most important criteria that should influence the grade. This is not the same thing as merely agreeing that each has a point and therefore they will split the difference. In this experiment during the paired marking the latter kind of compromise was limited to the occasion when, for example, the difference was between a C- and C+. Here, if the pair agreed to give the paper a C, there was no harm done. And in a true situation at examination time, this would also apply. The student still ends up with a C. However, where the difference represented a change of category from F to D, or D to C, or even upon occasion from D to B, then the markers tried hard to be fair and decide which of their criteria should weight more heavily in deciding on the final grade.

Thus far in this summary attention has been directed to the concept of pairing as it is understood in the thesis; however, equally important to the suggested method of marking is the idea of rapid impression. In the 1966 Godshalk, Swineford, Coffman study, the authors emphasize that the marking which was carried out in the experiment gave as true a measure of the students' performance as any kind of testing procedure will allow. The writer feels that this italicized qualification is important, especially viewed in the light of a rationale for rapid impression marking. When a candidate writes an essay for examination purposes, he is under a certain amount of tension; he has no recourse to a dictionary or source material; he is unable to consult with anyone; and he is under a strict time limit which never allows sufficient time for careful revision. Under these circumstances, the

end-product cannot usually measure up to a very detailed or close scrutiny. Not even the most experienced and competent writer can make his first draft, or even his second, free from weaknesses in syntax, style etc., under similar conditions. Hence, it seems less than fair to subject a student's examination essay to a minutely detailed criticism. Unquestionably it must meet the criterion of having communicated ideas so they can be grasped without strain, but apart from that it should be the overall impression relayed to the reader which governs the assessment. Any detailed or analytical marking should be reserved for assignments, which are testing rather than teaching devices. As for the suggestion made by Professor C.W. Valentine that a composition teacher should look at all essay-type examinations to judge whether or not a student measures up to stringent standards of acceptable writing, one can only pray that the professor be the first and only victim of such a Mephistophelean scheme.

The recommendations for improving standards of marking generally, and in a composition course particularly, then, are:

1. Insist upon clearly set down procedures so that major criteria are agreed upon before marking begins.

2. Use a rapid impressionistic method for mid-term and final examinations. It obviously is not suitable for the marking of assignments which are teaching, not testing, devices.

3. Never rely upon the judgment of one marker. The final assessment should be derived from the combined judgment of the instructor and his matched partner arrived at through mutual discussion and not by simple averaging of differences.

4. Where there is an impasse between the pairs, submit the paper to a committee of five or six experienced colleagues and average these marks in order to arrive at the final evaluation.

POSTSCRIPT

The method of valid and reliable measurement of writing ability as a subject for investigation is a veritable Pandora's box. When the lid is lifted to take out one problem for treatment, a host of others 'fly out' demanding similar attention. Only the surface of a token few have been examined in this thesis. It was established that there is indeed a grave problem as far as accurate evaluation of essays is concerned, and Cyril Burt set the stage for the discussion by laying bare the sinews of subjectivity which form a barrier to any easy solution - those ubiquitous but "personal, limited, and accidental influences" which are at the same time powerful, irrelevant, and irrational.

Long years of patient research have done little more than show that marking practices taken for granted as being 'foolproof' are nothing of the kind - in fact, disastrously the contrary. This experiment represented another attempt to approach the problem at least as it applies to one of the most vulnerable areas - that of composition. But the difficulty in assessing writing ability impinges on a much larger portion of the academic sphere. In fact, the final result of almost every subject that the student undertakes is influenced by it, and therefore it is inextricably tied to the whole question of who should succeed or fail in examinations on all academic levels.

Success or failure cannot be considered without reference to human abilities and exactly what these consist of or how they can be measured are problems which have been eluding clear definition and understanding for a very long time. They may be likened tc threads in a

precisely and intricately designed tapestry whose pattern has been obscured by the haphazard removal of key strands. As one reads progressively through the maze of literature on examinations and marking, one constantly comes upon odd loose strands which belong to this tapestry and must be re-threaded if the original design is to be seen clearly once more. Occasionally, we are led to catch a glimpse of the true pattern, through the genius of men like Galton who saw what now seems so obvious, that human beings are not carbon copies of one another. And then educators expound at length on the importance of taking into account "individual differences" in children, and the teacher training institutes pay lip service to the idea, while the children in many schools continue to be treated as though they were all cut like gingerbread men, from a common mold.

Nowhere has this been more evident than in the attitude towards examinations and subsequent promotion of pupils from one step on the academic ladder to the next. A few of the issues arising from these practices, as they relate to measurement of writing ability, have been aired in this thesis, but a multitude of others are implied. For example, is writing ability an integral part of general intelligence as reflected in academic aptitude; or, as some researchers have declared, does a test of writing ability, i.e. a composition, add something extra to a test of intelligence? Does the ability to write clearly depend upon the ability to think clearly, or vice versa? Or are they even connected? Is the ability to write, a God-given talent or can virtually anyone learn it as a skill? Are some educators inclined to splice two disconnected and only generically related 'strands of the tapestry' - expository and creative expression? The one is reflected in an ability to communicate ideas and facts effectively; the other in the talent for creative expression such as is found in

short stories, novels, poems? Is it a mistake to go on making composition an adjunct of literature, thereby giving it short shrift and demeaning its importance, when in reality it is the basic link among all branches of learning - academic, technical and vocational.

It is well recognized that in the business world, the person who can write a good report stands a better chance of being earmarked for promotion. In the academic sphere, this person has little difficulty with term papers or essay examinations. He may by chance suffer a penalty in subjects which require mathematical aptitude, if he lacks it, but his counterpart, who has considerable numerate ability and none in written expression, is penalized in every essay-type examination he writes, and he has a most difficult time in a compulsory composition course at university.

How do we measure ability in written expression? How do we measure progress in that ability? Is it like physical growth that goes in spurts, or is it more akin to mental or social maturation which seems gradual and continuous? Until some of these issues are clarified, is it fair to go on using an essay as a crucial test of academic worth, especially on school-leaving assessments? If this practice is to be continued, should educators not be applying themselves more assiduously to finding a method of instruction which will not leave so many students (who have all their mental faculties but have neither a special talent for writing nor any appreciation of the structure of English) at a great disadvantage when they are faced with such examinations or when they go into the business world?

The questions that have been raised in this post-script are representative of but a few of the many threads that have become unravelled from the entire tapestry of human abilities. If they present a confused and seemingly

knotty or twisted array of elements that are concerned with the measurement of writing ability, it is because that is exactly what they are. This thesis has done little more than try to re-thread a few small strands into the infinitely large and grand design. Perhaps they have not been placed correctly, or they may even have been joined to the wrong threads. Nevertheless, until all the missing strands are replaced and it is revealed what constitutes human achievement, and how it can be measured, little can be done but take the loose ends as they occur and try to find how they are linked to the others, and where they fit into the total design.

When and if they are all in place and secure, perhaps researchers can cease trying to decide whether general intelligence has anything to do with the way a student writes an essay, and there will be no need to seek further for ways of assessing writing ability accurately and fairly at the crossroads of matriculation and future opportunity. That particular Pandora's box will be closed for all time.

BIBLIOGRAPHY

Anderson, C.C. "The New Step Essay Test as a Measure of Composition Ability," Educational and Psychological Measurement, Spring, 1960, pp.95-102.

Ballard, P.B. The New Examiner. London: University of London Press, 1923.

Boyd, W. Measuring Devices in Composition, Spelling and Arithmetic. London: Harrap, 1924.

Burt, C.L. Mental and Scholastic Tests. London: King, 1921.

Cast, B.D.M. "The Efficiency of Different Methods of Marking English Composition," (Part I), British Journal of Educational Psychology, 9(1939) 257-269.

_____. "Efficiency of Methods of Marking Composition," (Part II), British Journal of Educational Psychology, 10(1940) 49-60.

Coward, Ann F. "The Method of Reading the Foreign Service Examination in Composition," Research Bulletin RB-50-57. Princeton, N.J.: Educational Testing Service, 1950.

Diederich, Paul B. "Reading and Grading" in Improving English Composition. (Arno Jewett and Charles E. Bish, eds.). Washington, D.C.: National Education Association, 1965, Chapter 11.

_____. "The Problem of Grading Essays," Princeton, N.J.: Educational Testing Service, 1957.

Diederich, Paul B., French J.W., Carlton, S.T. "Factors in Judgment of Writing Ability," Research Bulletin Series R.B.-61-15, Princeton, N.J.: Educational Testing Service, 1961.

"Examinations," Encyclopaedia Britannica, Vol. 8, Chicago: Encyclopaedia Britannica Inc., 1963.

Examinations Bulletin No. 3, "The Certificate of Secondary Education: An introduction to some techniques of examining," Secondary School Examinations Council, London: Her Majesty's Stationery Office, 1964.

Examinations Bulletin No. 5, "The Certificate of Secondary Education: School-based examinations," The Schools Council, London: Her Majesty's Stationery Office, 1965.

Examinations Bulletin No. 16. "The Certificate of Secondary Education Trial Examinations: Written English," The Schools Council, London: Her Majesty's Stationery Office, 1967.

Finlayson, Douglas S. "The Reliability of The Marking of Essays," British Journal of Educational Psychology, 21(1951) 125-34.

France, Norman. "Examinations: The Contribution of the School," McGill Journal of Education, Vol. 1, No. 1, 'Spring, 1966), 60-64.

France, Norman and Associates. "A Technique for the Assessment of High School Examinations: Rapid Impressionistic Marking by Suitably Paired Assessors (RIMSPA)," McGill University, 1967, (to be published in the Canadian Education Digest, September, 1967).

Gage, N.L. ed., Handbook of Research on Teaching. Chicago: Rand, McNally and Company, 1963.

Godshalk, F.L., Swineford, F., and Coffman, W.E. The Measure of Writing Ability. New York: College Entrance Examinations Board, 1966.

Grinnell, J.E. "What Makes Ability in English?" School Review, 45(1937) 602-604.

Guilford, J.P. Fundamental Statistics in Psychology and Education, 4th ed. New York: McGraw-Hill Book Company, 1965.

Hamilton, G.A. "Teacher Opinion on Acceptable Grade Eleven Writing," Unpublished Master's Thesis, McGill University, 1966.

Hartog, Sir Philip. The Marking of English Essays. London: MacMillan and Co. Ltd., 1941.

Hartog, Sir Philip, Rhodes, E.C. An Examination of Examinations, 2nd. ed. London: MacMillan and Co. Ltd., 1936.

Hartog, Sir Philip, Rhodes, E.C., Burt, C.L. The Marks of Examiners. London: MacMillan and Co. Ltd., 1936.

Huddelston, Edith. "Measurement of Writing Ability at the College-Entrance Level: Objective vs Subjective Techniques. Journal of Experimental Education, 22(March, 1954) 165-213.

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

Kitzhaber, Albert R. Themes, Theories, and Therapy: The Teaching of Writing in College. New York: McGraw-Hill Book Company Inc., 1963.

Lamb, H. "The English Essay in Secondary Examinations: A Comparison of Two Systems of Marking," British Journal of Educational Psychology, 23(1953) 131-133.

Lang, Albert R. Modern Methods In Written Examinations. Boston: Houghton Mifflin Company, 1930.

Mather, D.R. France, N. & Sare, G.T. The Certificate of Secondary Education: A Handbook for Moderators. London: Collins, 1965.

Meckel, Henry. "Research on Teaching Composition and Literature," Handbook of Research on Teaching, N.L. Gage, ed. Chicago: Rnad, McNally and Company, 1963.

Monroe, W.S. "The Unreliability of tne Measurement of Ability in Written Composition," Yearbook of the National Society for the Study of Education. 22 (1923) 169-171.

Morrison, R.L. and Vernon, P.E. "A New Method of Marking Composition," British Journal of Educational Psychology, 11(1941) 109-119.

Myers, A.E., Coffman, W.E. and McConville, C.B. "Simplex Structure in the Grading of Essay Tests," Educational and Psychological Measurement, 26 (Spring, 1966) 41-54.

News from The Schools Council. London: March 10, 1967.

Nisbet, J.D. "English Composition in Secondary School Selection," British Journal of Educational Psychology, 25(1955) 51-54.

Noyes, E.S., Sale, W.M. and Stalnaker, J.M. Report on the First Six Tests in English Composition. New York: College Entrance Examination Board, 1945, 72 pp.

Peel, E.A. and Armstrong, H.G. "Symposium: The Use of Essays in Selection at 11+, II.-The Predictive Power of the English Composition in the 11+ Examination," British Journal of Educational Psychology, 26(1956) 163-171.

Penfold, D.M. Edwards. "Symposium: The Use of Essays in Selection at 11+, I.-Essay Marking Experiments: Shorter and Longer Essays," British Journal of Educational Psychology, 26(1956) 128-136.

Shirts, Morris A. "When College Students 'Contract' for Their Grades," College Board Review, No. 63, Spring, 1967.

Smith, C.E. ed. The Marking of English Essays. Toronto: Macmillan and Company, 1941.

Stalnaker, John M. & Stalnaker, Ruth C. "Reliable Reading of Essay Tests," School Review, 42(1934) 599-615.

Starch, Daniel and Elliott, Edward C. "Reliability of the Grading of High School Work in English," School Review, 20(1912) 442-457.

_____. "Reliability of the Grading of High School Work in Mathematics, School Review, 21 (1913) 254-259.

Steele, J.H. and Talman, J. The Marking of English Composition. London: James Nisbet, 1936.

Thorndike, R.L. and Thorndike, Elizabeth. Measurement and Evaluation in Psychology and Education, 2nd ed. New York: John Wiley & Sons, Inc., 1964.

Valentine, C.W. The Reliability of Examinations. London: University of London Press, 1932.

Vernon, Philip, ed. The Measurement of Abilities. London: University of London Press, 1940. (Revised edition, 1956).

_____. Secondary School Selection, A British Psychological Society Inquiry. London: Methuen & Co. Ltd., 1957.

Vernon, P.E. and Millican, G.D. "A Further Study of the Reliability of English Essays," Part II, British Journal of Educational Psychology, 7(November, 1954) 65-74.

Wiseman, Stephen, ed. Examinations and English Education. Manchester: Manchester University Press, 1961.

Wiseman, Stephen. "The Marking of English Compositions for Grammar School Selection," British Journal of Educational Psychology, 19(1949) 200-9.

_____. "Symposium: The Use of Essays in Selection at 11+, III.-Reliability and Validity," British Journal of Educational Psychology, 26(1956) 172-9.

APPENDIX I

Pairing Statistics From Which Tables 1 & 3 Were Derived

Markers I-IV

* I Markings		II Markings		III Markings		IV Markings	
First:	Second	First:	Second	First:	Second	First:	Second
50	54	50	55	55	60	50	55
50	55	55	64	60	64	55	60
45	50	64	60	60	64	64	64
45	50	45	55	50	60	60	55
74	74	74	74	65	75	75	75
50	50	55	55	50	64	70	64
45	50	50	55	50	64	60	65
60	55	60	65	55	65	65	74
60	60	65	70	65	70	65	74
55	50	60	64	60	64	65	74
55	55	64	55	65	65	60	60
60	55	53	55	60	65	55	60
55	55	50	55	55	65	60	65
60	60	64	60	60	64	65	65
53	50	65	74	65	65	64	70
40	53	50	64	50	55	50	50
50	50	45	55	50	60	50	55
65	60	64	60	64	70	74	65
60	60	64	65	70	65	55	55
64	64	55	65	60	65	64	74
$\bar{X}$	54.8	55.5	57.1	61.3	58.5	64.5	61.3	63.9
SD	8.07	6.02	7.60	6.72	6.30	4.10	6.80	7.52
r	0.84		0.69		0.63		0.79	

* Candidates' papers were labelled A 1-40 and B 1-40. These papers were written by Candidates B21-40.

APPENDIX I (continued)

Pairing Statistics From Which Tables 1 & 3 Were Derived

V Markings		VI Markings		VII Markings		VIII Markings	
First:	Second	First:	Second	First:	Second	First:	Second
55	55	64	60	64	60	64	64
64	54	60	64	64	55	64	64
60	56	55	60	55	54	64	55
64	55	55	55	60	54	64	60
75	75	74	74	74	74	64	64
55	54	55	55	75	70	60	55
55	54	60	60	60	60	65	55
65	74	65	64	70	65	70	60
65	64	65	70	65	65	60	60
50	55	60	64	70	70	70	60
40	60	64	60	65	64	70	60
64	60	70	70	64	60	64	64
80	74	74	65	70	60	70	64
60	65	65	65	64	64	74	64
74	64	75	74	74	74	75	70
55	60	52	64	40	50	65	64
54	60	50	54	60	54	70	50
55	70	48	55	70	70	64	64
65	70	70	70	74	70	65	64
70	64	65	70	70	64	74	70
$\bar{X}$	61.3 62.1		62.3 63.6		65.4 62.8		66.8 61.5
SD	9.2 6.96		7.80 6.05		8.00 6.99		4.30 4.85
r	0.55		0.84		0.82		0.37

APPENDIX II

Statistics for Individual Markings of 80 Papers

I	II	III	IV	V	VI	VII	VIII	Group Average
55	70	64	64	60	70	64	65	64.0
65	64	64	60	55	60	64	60	61.4
65	64	70	55	70	74	65	65	66.0
70	70	70	74	65	74	80	74	72.2
60	64	60	64	55	65	55	64	60.8
65	64	65	55	75	75	70	50	64.4
65	60	65	64	70	75	65	60	65,6
65	70	60	64	60	74	70	65	66.0
64	64	55	64	55	65	65	65	62.2
45	54	60	50	40	55	50	50	50.5
65	65	64	70	60	74	64	80	67.8
60	64	65	64	64	65	65	65	64.0
60	65	70	65	55	70	65	64	64.3
74	60	60	64	64	64	65	55	63.2
60	64	64	65	60	70	65	64	64.0
60	60	64	64	65	65	55	64	62.2
55	55	60	54	55	60	54	54	57.2
55	65	64	65	60	64	54	54	60.2
50	55	60	50	40	55	50	60	52.5
50	55	60	50	55	60	54	40	53.0
50	60	60	50	50	55	52	53	53.7
50	55	60	55	40	60	50	53	53.3
55	60	64	55	50	64	50	55	56.7
60	60	60	64	64	60	70	60	62.3
60	65	65	65	60	65	54	60	61.8
55	60	65	55	64	55	50	50	56.7
64	60	65	65	65	65	64	65	64.1
60	64	64	70	55	65	65	70	64.1
64	64	65	74	65	70	70	60	66.5
53	60	70	64	54	70	54	54	59.8
64	60	65	65	64	75	55	55	62.9
58	60	60	64	60	70	54	64	61.3
50	65	65	60	54	64	50	54	57.8
55	60	64	55	70	70	54	60	61.0
60	64	70	60	60	54	55	60	60.4
64	70	70	55	74	70	55	64	65.3
64	60	64	64	60	55	54	64	60.7
60	64	70	60	65	65	60	64	63.5
55	64	65	60	70	64	50	60	61.1
55	65	65	64	64	74	54	60	62.7

APPENDIX II (Continued)

Statistics for Individual Markings of 80 Papers

I	II	III	IV	V	VI	VII	VIII	Group Average
55	55	60	60	50	64	60	70	59.3
50	54	60	55	54	55	54	65	55.8
50	54	65	64	64	65	64	74	62.5
64	64	70	65	70	70	70	74	67.2
40	54	54	50	40	54	54	50	49.5
40	55	55	50	55	54	50	53	51.5
45	54	60	60	40	54	54	60	53.3
65	64	60	70	60	70	70	75	66.7
60	65	60	64	66	64	65	64	63.5
60	54	60	50	50	60	54	64	56.5
65	64	65	74	60	60	70	74	66.5
55	54	60	54	50	65	60	55	56.5
55	55	60	64	66	64	65	64	61.6
55	55	60	64	66	64	60	60	60.5
60	60	65	65	64	65	74	64	64,6
60	55	64	55	64	70	60	65	61.6
54	60	60	55	60	60	74	55	59.7
55	64	65	64	40	60	55	64	58.3
60	64	64	50	60	65	70	64	62.1
64	60	64	60	50	70	70	65	62.9
54	55	60	55	55	60	60	64	57.9
55	64	64	60	54	64	55	64	60.0
50	60	64	64	56	60	54	55	58.0
50	55	60	55	55	55	54	60	55.5
74	74	75	75	75	74	74	64	73.1
50	55	64	64	54	55	70	55	58.3
50	55	64	65	54	60	60	55	58.0
55	65	65	74	74	64	65	60	65.3
60	70	70	74	64	70	65	60	66.6
50	64	64	74	55	64	70	60	62.6
55	55	65	60	60	60	64	60	59.9
55	55	65	60	60	70	60	64	61.1
55	55	65	65	74	65	60	64	63.0
60	60	64	65	65	65	64	64	63.3
50	74	65	70	64	74	74	70	67.6
53	64	55	50	60	64	50	64	57.5
50	55	60	55	60	54	54	50	54.7
60	60	70	65	70	55	70	64	64.2
60	65	65	55	70	70	70	64	64.8
64	65	65	74	64	70	64	70	67.0
$\bar{X}$ 56.1	61.0	62.8	61.5	59.5	64.5	60.6	61.5	60.9 (true mean)
SD 6.75	5.05	3.98	6.85	8.28	6.18	7.39	6.68	
r 0.59	0.69	0.62	0.49	0.66	0.35	0.79	0.38	

(with true mean).

APPENDIX III

Statistics for Paired Markings of 80 Papers

A(II & VII)	B(I & VIII)	C(III & VI)	D(IV & V)
66	64	66	60
64	60	64	60
66	65	65	64
76	72	74	70
60	63	64	60
66	64	65	66
64	60	65	66
70	65	70	65
64	64	60	60
50	45	50	45
66	70	70	66
64	62	66	64
65	63	64	60
63	62	64	64
64	63	64	60
60	63	64	64
55	58	58	55
58	55	60	60
50	55	55	45
50	45	50	50
52	50	54	50
50	50	54	50
54	54	54	50
64	60	60	64
60	60	60	60
60	55	60	56
64	64	64	64
65	65	65	65
65	65	70	70
55	53	55	54
55	55	60	60
60	58	64	64
50	50	54	54
54	58	60	60
60	60	60	60
60	64	64	60
55	60	60	60
64	65	64	64
58	58	60	60
65	64	65	64
55	60	60	55

APPENDIX III (continued)

Statistics for Paired Markings of 80 Papers

A(II & VII)	B(I & VIII)	C(III & VI)	D(IV & V)
54	54	54	54
58	60	64	64
65	64	65	65
50	45	50	45
50	45	50	50
54	54	54	50
70	70	68	68
68	65	65	66
54	54	50	50
68	70	68	65
55	55	55	55
60	60	60	60
60	60	64	64
70	65	65	68
60	60	63	58
64	60	60	60
60	60	64	58
65	64	65	60
60	64	65	58
58	60	60	58
55	60	60	58
58	58	60	60
55	55	55	55
74	74	74	76
60	55	55	56
55	50	58	56
65	64	65	70
65	65	68	66
64	60	64	60
60	58	64	58
60	60	64	64
65	65	65	68
64	60	64	64
74	65	65	65
55	60	55	56
50	50	54	54
65	64	64	65
65	65	65	66
65	65	70	70
$\bar{X}$ 60.7	59.9	61.5	60.1
SD 6.04	6.07	5.54	6.26
r 0.88 (with true mean)	0.85	0.89	0.90