AN EXPLORATORY STUDY OF DEGREE COURSE OBJECTIVITY AND GRADUATE PERFORMANCE

A thesis presented in partial fulfilment of the
requirements for the degree of Master of Arts
in Psychology at Massey University.

Sarah Russell

1992

ABSTRACT

This study looks at the issue of degree course objectivity from the perspective of Science and non-Science lecturers. It is an exploratory piece of research, and focuses on a sample of degree courses offered in New Zealand universities. Research was conducted in several steps, and involved the completion of two questionnaires. Participants were also asked to supply a written statement outlining the objectivity of assessment in their own teaching domain. The t-test statistic was used to measure the significance of research findings. In New Zealand, university lecturers recognise that a difference exists in the objectivity of degree course content. Further, they are aware that Science oriented courses lend themselves to greater assessment objectivity than the non-Sciences, despite disagreeing over the exact level of objectivity in the latter field of study. The variance in degree course objectivity has a potential impact on the distribution of 'good' degrees awarded across university departments, yet has not evoked the amount of attention amongst academics that it clearly merits. It is concluded, that in New Zealand, research must continue into the issue of subject matter objectivity as a potential impact on students' degree selection and employee recruitment.

ACKNOWLEDGEMENTS

I would like to thank Mike Smith, my supervisor for providing the idea and the initial direction in conducting a study on degree course objectivity. His practical assistance, commonsense suggestions and insightful comments have helped me refine many of the ideas that appear in this thesis. I am also grateful for his honest review of each successive draft of this text, and for otherwise massaging my efforts into a manuscript which was worthy of print.

I am also deeply indebted to Paul Elliot for the many hours he spent in assisting me format study questionnaires, and for his enduring patience as I came to terms with the WordPerfect programme.

Special thanks must also go to my parents, Heather and Jack for their encouragement and continued support with each successive year of higher education.

TABLE OF CONTENTS

CHAPTER FOUR - Degrees of Success in Different Major Fields

CHAPTER FIVE - The Present Study

CHAPTER SIX - Methodology and Results

CHAPTER SEVEN - Discussion

Appendices

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

Overview

Achieving a high level of academic performance is an important marker of educational success. In turn, the recipients of a 'good' degree may accrue particular benefits within higher education, and the wider field of work. For many years however, the consistency and validity of grading practices has generated a great deal of concern (Wainwright, 1977; Foster, 1985; Kennedy, 1990). Repeatedly, studies reveal that discrepancies in the grades that students receive are not the result of differences in intellectual ability and/or attainment. Instead, several studies suggest that the subject matter of degree courses has a substantial impact on the class of degree awarded (Dale, 1959; Bee & Dolton, 1985; Hindmarch & Bourner, 1986; Johnes & Taylor, 1987; Bolger, 1990). Nevin (1972) is adamant that such a factor should not go unchecked, yet little attention appears to have been paid to his recommendation.

In my own research, one aspect of the subject matter phenomenon will be studied in greater detail. That is, the extent to which course material can be designated as 'subjective' or 'objective' is of concern, and whether this factor impacts on the distribution of grades. Answers to these questions will be of interest not only to the taxpayers who fund tertiary institutions, but also to potential university students and employers. Potential students will be interested knowing the extent to which their chances of obtaining a 'good' degree are influenced by their field of study. Employers may wish to know which student groups graduate with a high proportion of first class degrees when recruiting new staff.

Chapter two is an overview of performance evaluation, and looks at the various measurement tools for assessing work behaviour. Much of the research concerning job performance is conducted in the work place, yet, in many respects, is similar to the evaluation of students' academic performance. The chapter concludes with a discussion on how to interpret evaluation results, and appropriately feed this information back to the recipient.

In Western society, attaining a university qualification confers considerable academic distinction upon the holder. Amongst some employers, certification is also crucially important in obtaining one's first job. In Chapter three, the relationship between society, education and grades is discussed in detail. Desirable psychometric concepts such as reliability, validity and the maintenance of standards is also considered in relation to postgraduate performance.

In Chapter four, several of the studies to address the issue of grade comparability across subject areas are discussed. Also of interest are the factors which may, in part, contribute to the observed variation in degree awards. Much of this research revolves around British data, yet is inconclusive as to why a discrepancy in degree awards might occur.

In Chapter five, an overview of my own research and hypotheses are presented. Of particular concern, is the issue of degree objectivity, and the extent to which this factor impacts on the distribution of grades in Science and non-Science degree courses. A philosophical debate on the meaning of the terms 'objectivity' and 'subjectivity' is presented at the beginning of the chapter.

A description of the subjects and questionnaires which were used in my study are introduced in Chapter six. Research findings are also presented in this section, in both written and diagrammatic form.

Research findings are interpreted in Chapter seven, and some explanations are provided for the observed outcomes. The links between previous studies and my own research are also highlighted and possible reasons for research discrepancies are made. Chapter seven concludes with a reply to the number of concerning issues which were raised by research participants, along with suggestions for future studies.

CHAPTER TWO
Performance Evaluation

2.1 Introduction

Passing judgement is a central part of social behaviour. Continually, we assess those around us, their clothes, accents, beliefs and actions as well as the products and services they offer. Most of the time, such judgements are made at a very personal, informal level. As individuals, we reserve the right to decide for ourselves who we will associate with, what services we will support and what products we will continue to buy.

Of course, in areas of human life in which we lack expertise, very few of us would claim the right to pass any formal value judgement. For this purpose we train and appoint specifically qualified people to judge the value of a particular performance for us. Education is no exception to this general rule. Here, school teachers and university lecturers operate as 'experts', and are entrusted with the task of evaluating students' academic performance.

Since formal evaluation can have a tremendous influence on one's future, such an activity should never be conducted in a casual, haphazard manner. For the benefit of students, academic staff are instead obligated to use systematically planned evaluation procedures. When carefully collected, evaluation data can highlight meaningful distinctions among individuals that correspond to actual behaviourial differences (Wendelken & Inn, 1981). It is important to note however, that evaluation instruments are not intended to replace the thoughtful judgements of teaching staff. Rather, by the removal of extraneous factors from the evaluation process they offer a more dependable basis for making such judgements.

Of course, despite the best intentions, evaluation procedures offer no absolute guarantee that data will be accurate and objective. Issues such as validity, reliability and bias are persistent problems which will often jeopardise the value of many performance evaluations.

2.2 The Meaning of Evaluation

Quite frequently, the terms 'evaluation', 'measurement' and 'testing' are used interchangeably with some confusion. Prior to any discussion of evaluative procedures, the distinction between these three terms should therefore be made clear.

Testing, as seems obvious, means using tests. In general, these consist of a uniform set of tasks to be administered to all group members at a prearranged time and place (Lindeman & Merenda, 1979). As generally used in education, the term testing has a somewhat unfavourable connotation. Among its critics, testing is regarded (whether rightly or wrongly) as a very specific and rather superficial activity (Gronlund, 1985). Specifically, critics suggest that the act of 'testing' simply involves the collation of scores and statistics and that test technicians have little regard for what these results mean in relation to the examinees.

As already mentioned, the overall aim of evaluation is to find out how a person performs when compared to some specific standard or criterion. According to Sax (1989), testing only provides a selective portion of information upon which estimates of success are made.

Measurement is essentially a descriptive process and is generally believed to be more inclusive than testing (Gronlund, 1985). As applied in education, measurement involves the assignment of a percentage score or number to express the extent to which a desirable characteristic is present (i.e., accuracy). To ensure the exactness of performance data, a range of measurement instruments can be used, such as check-lists, rating scales and score cards.

Evaluation is the third and final process by which performance can be estimated.

Here, an emphasis is placed on assessing the value and desirability of results, not just the collation of scores or the assignment of a mark (Lindeman & Merenda, 1979). In the process of making such value judgements, a variety of relevant information is often considered, such as written work, test scores and personal observations.

Throughout educational literature, performance evaluation is conceived as being a more comprehensive and inclusive activity than either testing or measurement (Gronlund, 1985; Carey, 1988). Thus, having selected a legitimate standard for judging performance, evaluation is likely to provide the most relevant information to determine the acceptability of work output.

2.3 A Model of Performance Evaluation

Performance evaluation is not an isolated exercise, nor is it merely a collection of techniques. Instead, evaluation is best thought of as a process which includes many different elements. Figure 2.1 is a graphic representation of how these elements might interact. This model is based on the work of Landy and Farr (1980).

According to this model, performance evaluation can provide a comprehensive description of one's work related strengths and weaknesses. It is apparent from Figure 2.1 however, that the effectiveness of evaluation rests heavily on how well the purpose of evaluation is defined. If the aspects of performance to be evaluated are carefully defined, the most suitable evaluative standard (or criterion) will be selected. Against this standard, performance can then be measured by using some appropriate evaluation technique(s).

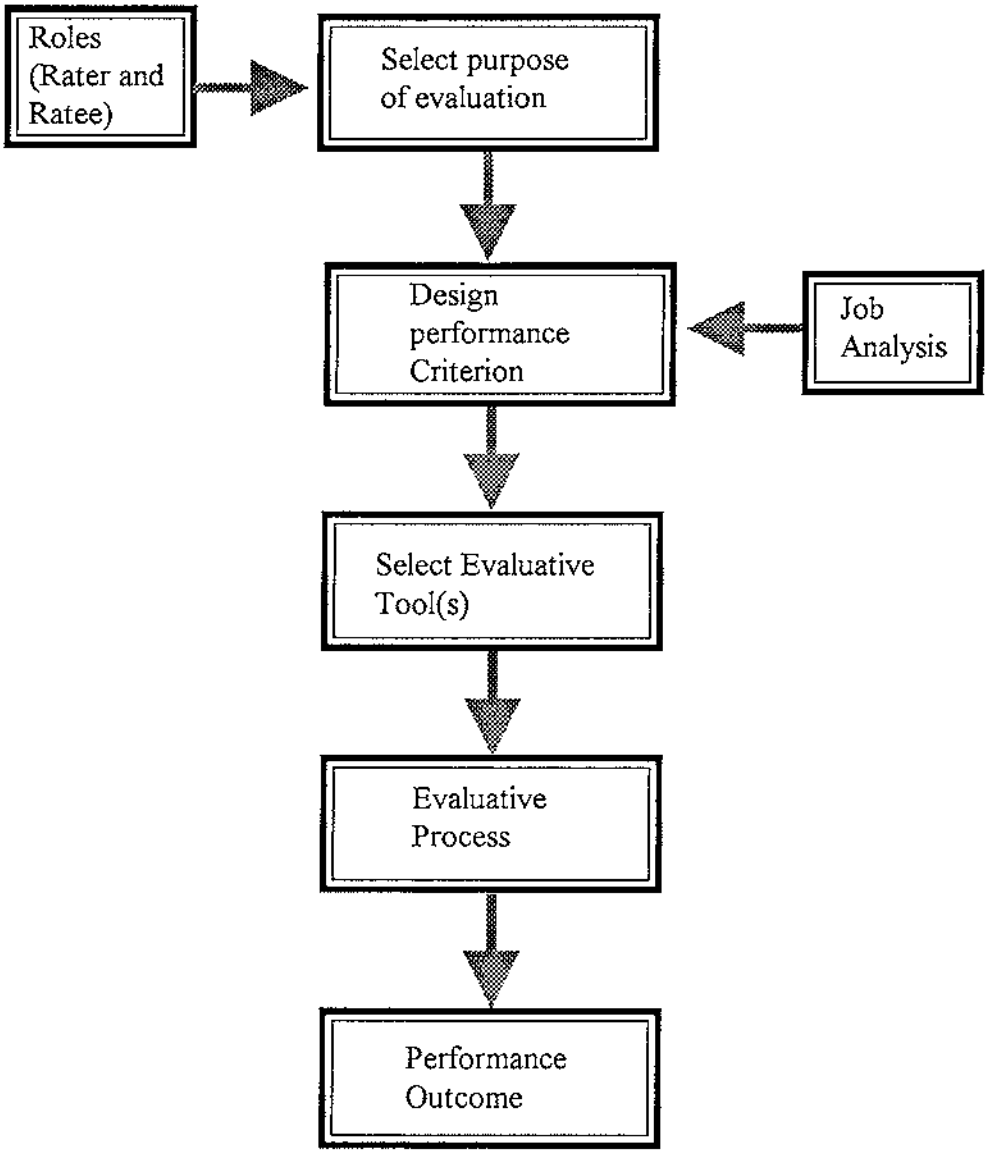

Figure 2.1: A component model of performance evaluation.
Source: Landy & Farr (1980).

2.4 The Function of Evaluation

Broadly conceived, four types of evaluations are made about student performance: placement, formative, diagnostic and summative (Hills, 1981; Gronlund, 1985). Since the purpose of evaluation will dictate which assessment procedures are used, the issue of what to evaluate always has priority in the evaluation process.

As its name implies, placement evaluation is concerned with correct student placement: a) in the most suitable instructional programme, b) at the optimum point of instruction and, c) with an appropriate teacher. Ideally, when making these decisions, various aptitude tests, self-report inventories and readiness tests should be utilised. These will help determine students mastery of prerequisite skills and course objectives.

During instruction, formative evaluation is used to monitor progress towards course mastery, and can reveal work related strengths and weaknesses. Typically, evaluative results are not used for assigning course grades. Instead, they are directed into modifying instruction, prescribing remedial work and reinforcing successful student learning (Sax, 1989). For the purpose of formative evaluation, teacher-made unit or chapter tests are most frequently used (Gronlund, 1985), but customized tests, non-graded 'spot' quizzes, and observational techniques are also applicable (Airasian & Madaus, 1976).

In most instances, learning difficulties can be remediated by the use of alternative methods of instruction (e.g., programmed materials and visual aids). In some cases however, the presence of persistent or recurring difficulties demand a more comprehensive and detailed diagnosis. For this purpose, diagnostic evaluation and the use of specially prepared diagnostic tests are of value. For serious learning problems, Gronlund (1985) also recommends the services of remedial, psychological and medical specialists. To use a medical analogy, formative evaluation provides first aid treatment for simple learning problems, and diagnostic evaluation searches for the underlying cause of problems which don't respond to first aid treatment.

At the end of instruction, summative evaluation is employed to determine programme effectiveness and pupil mastery of intended learning outcomes. To answer these questions a variety of techniques are often used including the teacher-made achievement test, oral reports and research assignments. Typically, the results of a summative evaluation are fed back to pupils in the form of a course grade. In higher education, the final grade obtained in a degree course will often generate both large and small career options (Johnson, 1988). Currently, this use of evaluation is under attack by students and educational critics (Heywood, 1989).

2.5 Developing the Criterion

Before conducting a performance evaluation, the examiner must specify the exact work behaviours to be measured. The process of developing a measure of worker goodness is often referred to as 'developing the criterion' (Blum & Naylor, 1968). In general terms, the criterion is an evaluative standard which can be used to measure a person's performance or to describe success (Landy, 1989). In education, the criterion for measuring a student's success might be the course grade. For a football team, the criterion might be the number of wins versus the number of losses in a season. In practice however, establishing the criterion is not always a simple matter. For Industrial and Organisational psychologists, this has been a problematic area for many years (Landy & Farr, 1983).

For this reason a large amount of research has focused on identifying the necessary requirements for criteria. Repeatedly, this research has highlighted the desirability of such characteristics as freedom from contamination, relevance, predictability and freedom from deficiency. To this list, Blum and Naylor (1968) have added several other items such as cost effectiveness, realism, consistency, and freedom from bias. Although few attempts have been made to operationally define these criterion characteristics, Landy (1989) recommends reducing them into three requirements: reliability, validity, and practicality.

Reliability is the first requirement of a criterion. In general terms, this concept concerns the stability of the criterion measure and can be estimated by a variety of procedures. Three methods frequently used involve: a) the repeated administration of the same criterion measure, b) the administration of a second 'equivalent' measure, and c) tabulating the internal consistency of a single measure (Thorndike, Cunningham, Thorndike & Hagan, 1991).

Validity is the second requirement of a criterion. As the term itself implies, validity refers to the accuracy and relevance of a criterion. That is, the extent to which it accomplishes the purpose for which it was designed. Estimating criterion validity is a highly judgemental procedure. Principally, this is because our actual criterion is only an approximate estimate of the ultimate or ideal criterion of success in any given situation (McCormick & Ilgen, 1981). Since we rarely, if ever, know what constitutes the ultimate criterion, it is impossible to empirically ascertain the relevance and accuracy of the criterion we have selected.

Of course, one can increase the likelihood of using a valid criterion by ensuring that it is neither biased or trivial and is relevant to some important goal of the individual, organisation or society (Smith, 1983). In addition, a criterion (and the measure of it) must be practical and available so that the cost of gathering performance data does not greatly exceed its potential benefit (Landy, 1989). Once the necessary and/or available criterion characteristics are identified, successful performance can be described. Several kinds of data can be used to provide this description.

2.6 Types of Criterion Data

Three viable measures of work behaviour have been identified by Guion (1965): objective data, personnel data, and judgemental data. To grasp the multi-dimensionality of 'job performance', these three categories should be considered simultaneously (Landy, 1989). Unfortunately, this advice is seldom put into practice. Landy and Trumbo (1980) sampled articles from the Journal of Applied Psychology from 1965 to 1975 and found that 72% of published studies only used judgemental indices. Similar findings have been reported by and Blum and Naylor (1968).

In general, criterion measures of work behaviour are used to appraise job performance in the business sector. These measures are a useful administrative tool and are powerful enough to dictate personnel decisions such as: a) staff selection and promotion, b) the selection of training programme objectives, and c) the nature of supervisory feedback and control (Cummings, 1973).

The performance measures identified by Guion (1965) can, of course, be applied to many other facets of life. For example, in education, it is quite legitimate for a teacher to collect a range of objective, personnel and judgemental measures of student work behaviour. In Figure 2.2 a comparable sample of educational and workplace performance measures are shown.

Of the three potential measures of student performance, one might logically expect objective data is the most frequently used. Interestingly enough, this does not appear to be the case. Described by Guion (1965) as "simply a count of the results of work", the objective category of criterion information is quite problematic.

Firstly, there is the simple measurement problem of reliability. Because objective behaviours are typically observed over a short period of time, there is ample opportunity for variance to distort the behaviour of individuals. A second problem concerns the educational significance of objective data. This issue is particularly relevant at the primary school level of schooling, where activities are often interrelated and incorporate multiple skills. In this environment, objective measures tend to poorly represent performance since they must, by nature, focus on rather small and discrete behaviours (Landy, 1989). Finally, for many activities and student projects, it is seldom possible to find the resources or the time necessary to collect objective information from each individual.

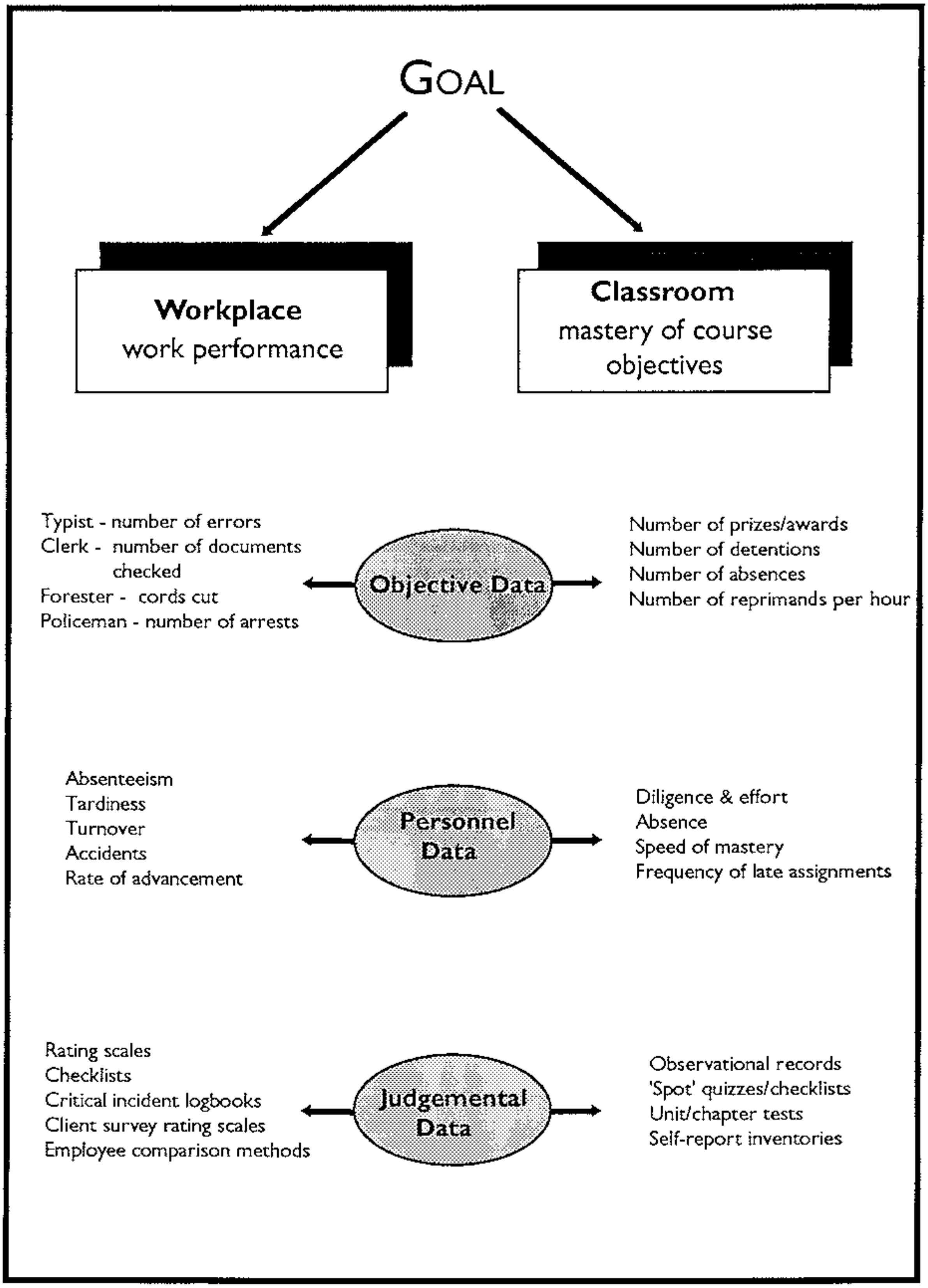

Figure 2.2: A selection of performance measures used in education and the workplace.

Data concerning the number of times a student is absent or late from class, and the number of incomplete and/or overdue assignments they have submitted, are all examples of personnel data. This kind of information is usually available in the permanent record folder of an individual, and can be used to define a 'good' and 'poor' student.

Once again, most of these measures are inappropriate for administrative and research use. This is largely due to the lack of specifity in concepts such as 'absenteeism' and 'tardiness'. Until these concepts are refined, accurate classification and recording of target behaviour will remain unlikely (Landy, 1989). The reliability of personnel data is also generally poor. In the case of absenteeism, reliability estimates seldom fall outside the range of .30 to .50. (Landy, 1989). Of course, this low reliability is hardly surprising, given the ample opportunity for variance to distort personnel measures. Some common types of variance are administrative (i.e., inaccurate recording of attendance reports), and environmental (i.e., ambient flu and atmospheric fluctuations).

The problematic nature of objective and personnel data does not imply that these criteria are inappropriate indexes of performance. Rather, it would suggest that if they are to be useful, a clear relationship must initially be established between the designated classroom activity and the target behaviour.

The third, and most pervasive class of criterion information, is judgemental data. These judgements can take many different forms. For example, they might involve comparing the performance of one student against all others, marking on a continuum scale the level of a student's proficiency, or simply checking the performance of students against a list of statements.

2.7 Rating Scales

In education, the rating scale is the most widely used judgemental measure of performance (Aiken, 1963; Thorndike et al, 1991). Conveniently, these scales can be distinguished from one and other by three major characteristics (Guion, 1965). The first characteristic concerns the manner in which the scale response categories are 'anchored'. Quite simply, scale anchoring involves marking a rating scale into regular and meaningful units. Because scale anchors are designed to guide a rater when measuring performance, attention must be devoted to developing a rating scale which is as efficient and psychometrically sound as possible.

To date, a substantial amount of research has focused on identifying the ideal number of response categories to include on a rating scale. In one study, Jenkins and Taber (1977) revealed that there is little utility in adding scale categories beyond 5; yet reliability will drop with 3 categories or less and with 11 categories or more (Landy & Farr, 1980).

The relative effectiveness of numerical, adjectival or behaviourial anchors has also been the subject of continuing debate (Burnaska & Hollman, 1974; Kingstrom & Bass, 1981; Landy & Farr, 1983). Taken together, this research suggests that there is some advantage to using behaviourial anchors over any other rating format. Specifically, it appears that behaviourial anchors have greater face validity for both the rater and the person being rated, can minimise the use of inaccurate worker stereotypes, and prove useful in providing concrete feedback to appraisees. One must remember however, that the relative accuracy of different scale formats may vary with the nature of the activity under review or the performance function being rated (Siegel & Lane, 1987).

The second descriptive characteristic of rating scales is the degree to which scale ratings can be interpreted. This characteristic refers to 'response clarity', and is largely dependent on the adequacy of scale construction.

The third rating scale characteristic concerns the adequacy with which scale dimensions are defined for the rater. Whenever possible, scale anchors should be limited to work related behaviours that are directly observable (Kane & Lawler, 1979), and should be accompanied by a numerical point system (Jacobs, 1986). If defined precisely, scale anchors are less open to misinterpretation, and will ensure raters direct their attention towards the same aspects of performance in all pupils.

2.8 Rating Scale Errors

Despite their widespread use, judgemental measures of performance are not exempt from error or bias and have been a constant source of dissatisfaction for researchers and practitioners (Landy & Farr, 1983). Three types of rating errors are largely responsible for this dissatisfaction: halo, central tendency and leniency.

Halo error, named by Thorndike (1920), occurs when the rater is unduly influenced by a single favourable or unfavourable trait of the ratee. Once an impression is formed, it then colours the rater's judgement of the individual's other traits. The effect of this psychometric error is most pronounced when multi-factor ratings are required and will yield a higher correlation between factor ratings than would otherwise be the case (McCormick & Ilgen, 1981). To minimise the likelihood of halo error, scale dimensions should be clearly defined and anchored (Landy, 1989) and raters should be trained in overcoming impression formation (Ivancevich, 1979).

The second type of error, central tendency, is characterised by an unwillingness of the rater to assign extremely high or low ratings. As a result, the performance of above average workers is rated about the same as those who do less. This kind of injustice often appears when raters are required to justify extreme ratings, and can easily lead to lower organizational performance (Henderson, 1984).

Leniency-severity errors are most readily committed by raters who are unusually harsh or easy in their ratings. A graphic representation of these errors would show that the former causes ratings to bunch up towards the upper end of the scale, the latter at the lower end. In both cases the rater has applied their own personal standards to the rating scale, and as a result, will reduce the effective width of the scale and make ratings less discriminative (Anastasi, 1988).

One of the most widely used methods of combating the leniency-severity error is the forced distribution technique. This is a procedure which requires the rater to place a certain proportion of his responses into different categories (Blum & Naylor, 1968). Understandably, raters frequently object to being forced into such a strict response pattern which is also time consuming to achieve. Better scale development is a more acceptable procedure to minimise leniency-severity error. Specifically, Landy (1989) recommends reducing the degree of scale ambiguity by improving the definition of dimensions and the nature of scale anchors.

2.9 Interpreting Evaluation Results

In the final phase of a performance evaluation, an appropriate method of interpreting the results must be selected. Interpretation is a vital step in the evaluative process, since it can provide curriculum planners with an accurate description of pupil performance. The more accurately performance results are interpreted, the more effective teachers will be in directing student learning (Gronlund, 1985).

Traditionally, descriptions of pupil performance have fallen into one of two categories. On the one hand, an individual's progress may be compared to the performance of some specified reference group (e.g., other members of the class), or alternatively, against some criteria specified by the teacher (Sax, 1989).

When evaluation instruments are used to compare one pupil with all others, they provide relative, norm-referenced interpretations. When pupil progress is interpreted by indicating what the individual is able to do (i.e., type forty words per minute, or spell ten specified words correctly) the interpretation is referred to as absolute, or criterion-referenced.

Strictly speaking, 'criterion-reference' and 'norm-reference' only refer to the method of interpreting the results. These distinct types of interpretation are likely to be most useful however, when the evaluation instruments are specifically designed for the type of interpretation to be made.

Discussion has already focused on the major descriptive dimensions of one widely popular criterion-referenced technique - the graphic rating scale. A selection of other techniques are also suitable for making absolute, not relative, decisions about student performance. These include the objective or essay test, teacher-made mastery tests, checklists, behaviourial observation scales (Latham, Fay, & Saari, 1979), and the mixed standard rating scale (Blanz & Ghiselli, 1972).

Criterion-referenced techniques are deliberately constructed to reveal student performance in some specific instructional domain. For this reason, task items must be selected on the basis of how well they reflect the instructional objectives being measured (Gronlund, 1985). If the learning tasks of interest are easy, the test items should be easy. Here, no attempt is made to eliminate particular items or to arbitrarily alter item difficulty to obtain a range of scores. Within the context of classroom instruction then, a criterion-referenced test is "deliberately constructed to yield measurements that are directly interpretable in terms of a specified domain of instructionally relevant tasks" (Glaser & Nitko, 1971).

In contrast, norm-referenced evaluation is concerned with the typical performance of typical people. To measure this performance, standardized aptitude and achievement tests are most frequently used. These instruments are designed to rank pupils in order of achievement, and will therefore tend to favour items of a more difficult nature. Gronlund (1985) supports the selective inclusion of items as a way of increasing the range of test scores and maximising performance differences. If a reliable ranking of pupils is achieved, decisions based on relative achievement (e.g., selection, grouping, grading) can be made with greater confidence.

2.10 Providing Feedback

Once a performance evaluation has been conducted, it is appropriate to present the persons who were evaluated with some version of the information. Considered from this perspective, one is faced with the problem of sending a message (the feedback) to a recipient (the worker) from a source (a supervisor, co-worker, or the task itself).

A process model which highlights the manner in which feedback is communicated to recipients is shown in Figure 2.3 (Ilgen, Fisher & Taylor, 1979).

According to this model, performance feedback will pass through four stages of processing before a reply is made. At each stage, there is an interaction between the three basic elements of the model (the feedback source, the message itself and the recipient) which will dictate the way in which information is processed.

Despite its simplicity, Ilgen's model is appealing since it communicates the importance of gathering valid and reliable performance data. If this is not achieved, recipients are unlikely to perceive feedback data as accurate. In turn, this situation may hamper their desire to respond, and set new goals for performance improvement. Numerous studies have attempted to identify critical factors to improving the acceptability of work performance feedback (Kay, Meyer, & French, 1965; Becker, 1978; Latham & Wexley, 1981; Stone, Gueutal, & McIntosh, 1984).

One factor which has received continuing attention is the order in which positive and negative feedback is presented. In one study, Stone et al (1984) revealed that recipients were more receptive towards feedback when positive, rather than negative comments opened the discussion. Interestingly however, the praise first approach was only effective when the recipient had a high level of self esteem and an internal locus of control. Under contrary conditions, the sequence of feedback was immaterial to the recipient's acceptance of comments. This finding is indirectly supported by Kay et al (1965) who argue that criticisms increase defensiveness and block the receipt of positive information.

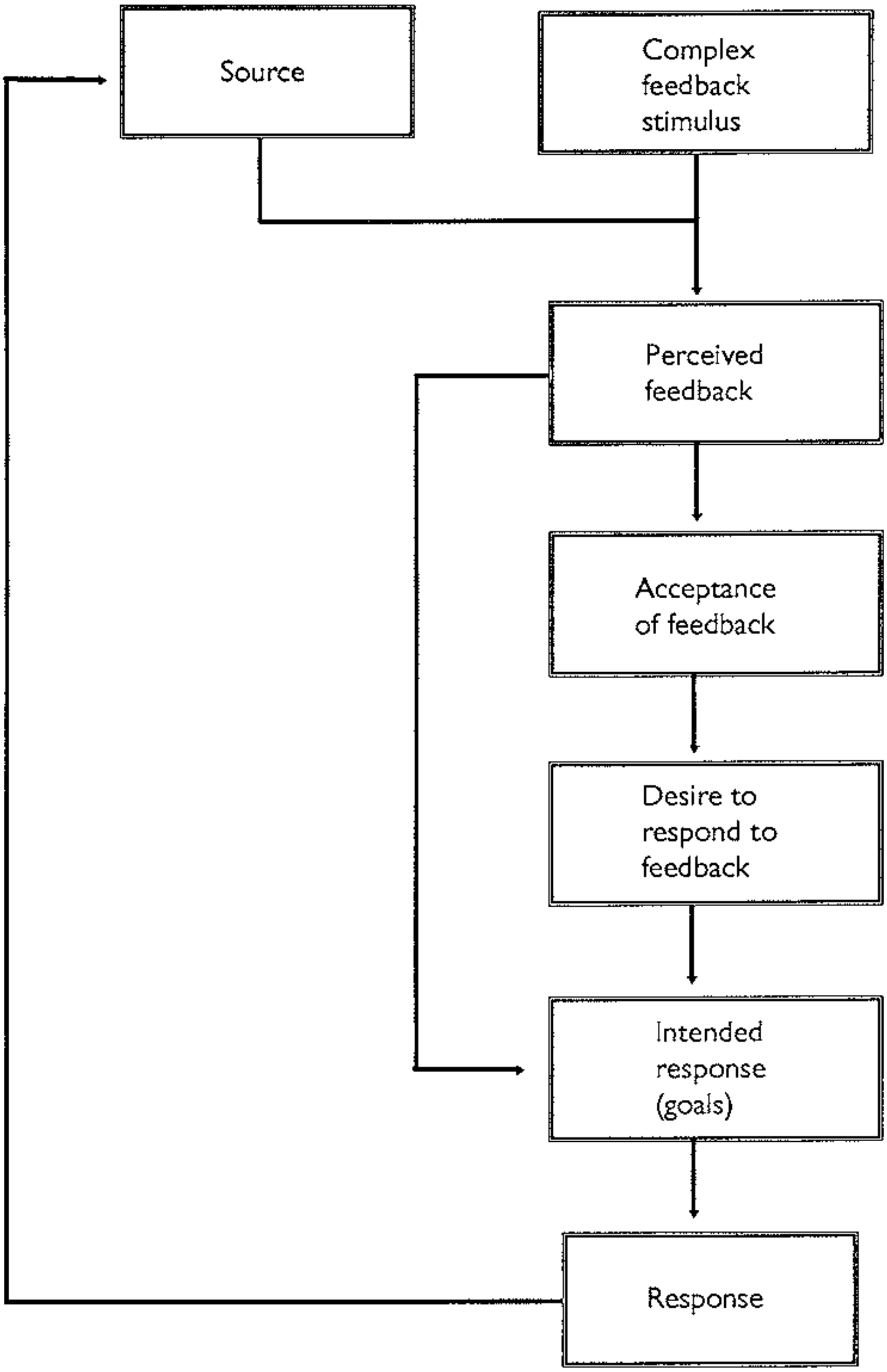
Source
Complex feedback stimulus
Perceived feedback
Acceptance of feedback
Desire to respond to feedback
Intended response (goals)
Response

Figure 2.3 A process model of feedback as a communication process.
Source: Ilgen, Fisher & Taylor (1979).

If negative feedback must be presented, it should be accompanied by specific examples to justify the criticisms, and a specific plan for improving the performance problem (Latham & Wexley, 1981). To enhance the acceptability of feedback, particularly negative feedback, the source must also be perceived as having credibility and job relevant expertise (Stone et al, 1984).

Recipients who are prepared for an appraisal session and take an active part in discussion, also appear to be more receptive towards performance feedback (Latham & Wexley, 1981; Siegel & Lane, 1987). Specifically, evidence suggests that individuals who are encouraged to discuss and respond to appraisal results, perceive the feedback process as both fair and in their own self interest. Under such circumstances, recipients are likely to accept feedback as accurate and experience a higher level of satisfaction with the feedback conference.

Of course, neither acceptance, nor feelings of satisfaction about performance feedback, can guarantee a behaviourial change. To achieve this, Becker (1978) stresses the importance of focusing the feedback discussion on goal attainment. Specifically, he suggests that the provision of precise, challenging, yet attainable goals will motivate the most dramatic change in behaviour.

2.11 Conclusion

Evaluation is a continuous process which underlies all good teaching and learning. It is an integral part of the instructional programme, and provides basic information for a variety of educational decisions. The main emphasis in evaluation, however, is the pupil and his or her learning progress.

In postgraduate education, the learning progress of students is often determined by a multitude of procedures. Some of the most frequently used procedures include, the essay test, laboratory reports, seminar presentations and the final exam. Regardless of which devices are used, performance in higher education is always interpreted in an absolute, not a relative manner. In New Zealand universities, criterion measurement translates into the class of honours awarded with a postgraduate honours or masters degree. Chapter three looks at the relevant issues which surround the selection, sorting and categorization of students at this level of education.

CHAPTER THREE

Postgraduate Performance Evaluation

3.1 The Grading Game

Grading student performance is an important ingredient in contemporary social mobility. Grades 'sift' people into different academic courses, and will present them with both large and small career options (Heywood, 1989). Students who receive low grades may withdraw from a course, or if they persist to graduation, achieve fewer of the benefits that accrue to their more scholarly peers. Of course, this does not imply that persons without a higher education fail to succeed in their chosen field of endeavour. For the majority of people however, some post-secondary school education is essential if they are to become successful.

According to Bourner and Bourner (1985), higher education has a twofold effect on the labour prospects of graduates. Firstly, they suggest that higher education increases the value of students as employees (whether this be in terms of developing specific skills or developing trained minds). Secondly, higher education is a useful device for screening out the most potentially valuable job applicants as employees.

Although higher education almost certainly achieves both these aims, it is reasonable to expect that their value will vary among industries and employers. Research conducted by Heywood (1989) reveals that employers will disregard qualifications if a person can complete a specified job. Similarly, Roizen and Jepson (1985) found that the majority of employers acknowledge the importance of various non-academic qualities, skills and knowledge when recruiting new staff.

In some professions, the 'value' of graduates will also be influenced by the content of their degrees. Specifically, Goldman and Widawski (1976) suggest that the value of people who complete Engineering, Law or Medicine will be tacitly or explicitly determined by the combination of subjects they studied and grade performance. On the other hand, industry and commerce seem to take a naive view of the value of grades and degree content. Typically, Heywood (1989) suggests that employers in both fields will simply take what universities have to offer without too much comment.

A substantial amount of research has been conducted in the area of graduate recruitment. Taken together, these studies suggest that grades are of moderate to crucial importance in obtaining one's first job. This is somewhat surprising, given the poor relationship that exists between grades and subsequent life achievements (Milton, Pollio & Eison, 1986). One can only speculate that employers rely so heavily on grades since they provide the simplest base line measure for the recruitment process.

Not everyone favours the emphasis that society places on grades or the pervasive influence that examinations hold. In one lively article by Klug (1977), the concerns and sentiments of many people with regard to the grading process have been expressed. Among other things he wrote:

> "not only does the single grade at the end of the course conceal information about a student instead of conveying information, it can also be positively misleading. The process of aggregation of results is such, that a wrong or arbitrary allocation is liable to be made .... statistical difficulties apart, the grading system is misleading not so much because of what it is as because of what it is supposed to be: absolutely objective and timeless .... grading has tenaciously retained its credibility long since empirical evidence has been accumulated which should have undermined confidence in the system" (p.16).

Sixteen years have now passed since Klug (1977) addressed the issue of grading practices in higher education. Given the continuing debate surrounding student assessment and academic study, his criticisms remain pertinent today. Amongst other writers it is also generally acknowledged that the statistical procedures on which grading is based tend to eliminate outstanding performance (Ford, 1977), stifle the most promising and creative student (Heywood, 1989), and too often reward the plodder, the memoriser, and the conformist (Boud, 1990).

However the problem is approached, the practice of academic assessment will continue to pervade the whole domain of university life and influence student approaches to learning. As a result, it is crucially important that exams and the grading process are both reliable and valid. That is, they should be seen to be fair.

3.2 Assessment Reliability

For many years, the issue of assessment reliability has captured the attention of educationalists (Dale, 1959; Cox, 1967; Wainwright, 1977; Foster, 1985; Johnson, 1988). Despite this attention, very little research has ever challenged the conclusion drawn by R. Dale in the University Quarterly (1959). Writing on the issue of university standards, Dale argued that the biggest obstacle to reforming unreliable university exams was the ignorance of staff regarding the pitfalls which surround the examiner. Specifically he stated that:

> "The calm assurance with which lecturers and professors alike believe that they can carry around in their heads an unfailing correct conception of an absolute standard of forty per cent as the pass line is incomprehensible to anyone who has studied the research on the reliability of examinations" (p.186).

Over the years, studies have continued to reiterate Dale's criticisms, and have highlighted which 'pitfalls' exist to obtaining a reliable assessment. Blok (1985) has found for instance, that examiners not only disagree amongst themselves with regards to mark allocation, but also disagree with their own ratings on different occasions. One of the main reasons for such discrepancy is that examiners tend to make random but straightforward mistakes, such as misreading graphs and overlooking points in answers (Cresswell, 1986). Further to this, examiners have rarely been found to share the same implicit ideas about standards, or allocate grades with the same degree of severity or leniency (Johnson, 1988).

In postgraduate education, the marking schedule employed by examiners is not the only area in which reliability is of concern. Reliability is also of the upmost importance when selecting an assessment device. If lacking in this characteristic, a device may be unfairly biased toward some students and against others, provide grades that lack stability over time and place, and only assess a limited subset of educational aims.

For many years, psychometric experts have felt that the unseen essay examination possesses all these disadvantages. When using this device, student grades may be affected by such irrelevant features as the context position of an essay (Hales & Tokar, 1975), the quality of handwriting (Bull & Stevens, 1979), marker expectations (Chase, 1979), and inconsistent marking (Blok, 1985).

Whether these disadvantages are inherent in the essay exam itself, or simply contextual in nature, does not detract from the concern they have generated. In postgraduate education, the essay exam is one of the most frequently used assessment techniques (Henderson, 1980). As such, it should provide grades which are fair and stable over time, and remain consistent when used by different examiners.

Among educationalists, a number of procedures are employed to improve the reliability of both assessment devices and exam marking. With regard to the latter form of unreliability, three corrective procedures have been proposed. Firstly, Heywood (1989) recommends reducing marker errors by simply increasing the number of examiners and taking the mean of their marks. Although effective, this procedure will often disadvantage the student who shows a particular type of brilliance in one area. By averaging the marks, the outstanding candidate becomes just an average student, and can be deprived of remedial attention which might be required in some other area (Cox, 1967).

If exam questions are graded student by student, the shifting of standards from one paper to the next may also contribute to marker unreliability. To minimise this influence, Gronlund (1985) recommends scoring all answers to one question at a time, and shuffling exam scripts between each block of scoring. Apparently, this procedure will assist examiners in maintaining a uniform standard of marking and will minimise the impact of any 'halo error'.

Some evidence suggests that marking schemes are also a viable approach to improving marker reliability (Foster, 1985). At the postgraduate level however, very few departments have a formal statement of the criteria to be used in awarding exam marks. Further, given the nature of the work being assessed, marking schemes in higher education are not always appropriate.

Fortunately, if corrective procedures do not exist, the postgraduate candidate is still partially protected against marker unreliability by the number of papers they take (Dale, 1959). To protect against unreliability in the assessment device itself, a different range of procedures must be adopted. Frequently used procedures include increasing the length of the test, giving preference to specific questions that can be answered briefly (Hills, 1981), including test items that have only one correct or clearly best answer (Gronlund, 1985), and reducing the ease with which students can cheat or guess items correctly (Carey, 1988).

To achieve a reliable, consistent measure of student performance, the design of exam questions and their marking needs to be rigorous. If one is to have any confidence in exam results, educationalists cannot afford to be sanguine about marking; neither can they afford to treat the setting of papers with indifference.

3.3 Assessment Validity

Throughout history, tertiary institutions have been regarded as the official certifiers of an extremely important product: the knowledgeable, educated individual. Accordingly, the measurement of student performance must be consistent, but also legitimate and valid. To accept anything less does a great disservice not only to the taxpayers who fund the institutions, but society as a whole and the individual students.

In psychometric terms, validity refers to the extent to which evaluation results serve the particular use(s) for which they were intended (Anastasi, 1988). If the results are intended to describe pupil achievement, it is imperative that they represent all aspects of achievement one wishes to describe, and nothing else. If the results are used to predict pupil success in some future activity, they should provide as accurate an estimate of future success as possible. Basically then, validity is always concerned with the specific use to be made of exam results and the soundness of one's proposed interpretations (Gronlund, 1985).

In postgraduate education, exam results are typically used to describe the specific knowledge and skills which pupils can demonstrate. To ensure an accurate description of pupil performance, Kennedy (1990) recommends the use of both continuous assessment and the final exam. By using both techniques, the consistently hard working student will not be unduly disadvantaged if they do poorly in the final exam. For some students, the final exam will provide the last opportunity to rectify a poor performance during the year. According to Kennedy (1990) the final exam should be weighed no less than 50 per cent of the overall subject grade. Keeping in mind the importance of continuous assessment, the weight of this component should be no less than 30 per cent.

The implementation of a continuous assessment programme will not, on its own, ensure a valid measure of student performance. To help secure this end, Carey (1988) also highlights the importance of sampling a representative spread of the item universe. In other words, a performance measure should not be overloaded with aspects of the course which readily lend themselves to the preparation of items. Instead, if a test is well constructed, items should cover the objectives of instruction as well as its subject matter. Content must therefore be broadly defined to encourage the demonstration of factual knowledge, the interpretation of data and the application of principles (Anastasi, 1988).

The clarity with which items are written will also affect the validity of an assessment device (Gronlund, 1985). To ensure a valid measure, items must be carefully constructed to avoid ambiguity and unintended complexity. Clearly written items allow students to focus their attention on the actual skill being measured, and will reduce the level of disagreement among examiners on the quality of responses (Thorndike et al, 1991). Moreover, when ambiguities are removed, items become less susceptible to correct guessing (Sax, 1989).

To improve test validity, Carey (1988) also recommends the use of novel items each time an objective is measured. When novelty is introduced into test items, the performance of a skill can be judged with greater accuracy. Examiners are also less likely to make incorrect inferences about one's level of achievement. If novelty is not possible, students may correctly answer exam questions by simply reproducing the desired answer from a previous unit test, practice test, or pre-test.

Thus far, discussion of assessment validity has been limited to test content. One must remember however, that test score validity is also influenced by factors such as test administration and scoring (Gronlund, 1985), inadequate test-taking orientation (Anastasi, 1988), test anxiety or fatigue (Hills, 1981), and the cultural background and experiential history of examinees (Sattler, 1982).

If, for any reason examiners decide to compromise on a valid assessment process, the impact on the tertiary institution in question will always be far reaching and negative. In the long term, Kennedy (1990) believes that assessment invalidity can promote significant staff demoralization, student cynicism regarding the assessment process, and an overall lack of positive spirit among staff and students alike.

3.4 Assessment Methods

Over the years, writers have repeatedly commented on the wide variety of assessment devices which exist at degree level (King, 1976; Thompson, 1979; Johnson, 1988). In Geography departments, an increasing emphasis has been placed on practical work, projects, and dissertations to assess the extent of student knowledge (King, 1976). In the case of Physics, a large portion of assessment is based on laboratory work (Thompson, 1979). While these measurement devices have been used with increasing frequency, the multiple-choice and essay type exam are still the most popular item formats employed (Thorndike et al, 1991).

Although popular, it has already been revealed that the essay exam is too unreliable to form the sole method of assessment. Ager and Weltman (1967) go one step further and suggest that:

> "no single examination technique is completely satisfactory in terms of both reliability and validity .... [and] that a variety of techniques should be used in university examinations, such techniques being chosen according to the functions that the examination performs" (p.272).

The adoption of a wide range of techniques does not, by itself, resolve the problem of reliability and validity. Rather, what is important is that they serve the purpose of testing specific objectives and can integrate learning. This view has been reiterated by numerous psychometric experts (Thorndike et al, 1991; Gronlund, 1985). These experts recommend selecting a particular item format on the basis of such considerations as: the relative ease with which course objectives are measured; the degree of difficulty in constructing or scoring items; the extent to which students can select and integrate their own learning, and the degree of freedom from irrelevant sources of variation in test results.

Of equal importance to this list of considerations, is the need to select assessment procedures which will stimulate more real-life problem solving skills. If such skills as communication, self development and interpersonal management are assessed, Heywood (1989) believe that graduates will be better prepared for subsequent employment.

If such a suggestion was incorporated into the curriculum, important changes in assessment practice must follow. Firstly, to assess interpersonal skills, it is perhaps more appropriate to have the student demonstrate their competence in a group situation, rather than write about the skill. To assess personal and social skills, Swain (1984) favours the keeping of diaries and group projects rather than the traditional unseen essay or multi-choice exam.

Once suitable assessment methods have been selected, one must then identify the most appropriate combination of evaluative tools and when to administer them. According to Ager and Weltman (1967), a strong case exists for using predominantly 'objective-type' tests whenever a critical decision must be made, such as whether a student should be allowed to continue at university or complete a particular course.

In the end of year final exam, Ager and Weltman (1967) favour the use of both essays and various types of 'objective tests' (i.e., multi-choice, true-false and matching). These latter instruments are regarded as being particularly suitable for the conscientious student who cannot write narrative, and can be used as the basis of a pass/fail decision. If carefully constructed, the essay question will also enable 'intellectual high-fliers' to demonstrate their ability to select, integrate and evaluate course material. The adoption of different test measures is also supported by Ager and Weltman (1967) when important decisions such as one's degree class is considered. Under these circumstances, a test battery will ensure that the most reliable measure of student learning is achieved.

Whatever combination of assessment practices are selected, Kennedy (1990) recommends that they be uniformly applied across the whole institution. Where this has not occurred, educationalists have faced problems in judging grade equivalence (Johnson, 1988), and have failed to establish and maintain a fair and common standard of academic excellence (Kennedy, 1990).

3.5 Academic Standards in the University System

In higher education, the maintenance of a fair and equitable standard in degree awards is one of the most important issues faced by academic staff. In Britain, the notion of parity is deeply ingrained in the university system. Indeed, as early as 1843, a passage from the University of Durham Calender asserted that: "The standard of the degree of B.A., as for all other degrees, is the same as that which is required at Oxford" (Piper, 1990, p 1). In 1888, the charter of the Victoria University (Manchester), required that examiners be appointed from other universities for all degrees, to ensure that standards remained "consistent with that of the national university system" (Piper, 1990, p 1). By the beginning of the twentieth century, the external examiner system became crucial in the maintenance of examination standards (Silver & Silver, 1986).

Although seldom appreciated, there is more than one form of parity implied by the ideal of a universal standard. For example, Williams (1979) has identified at least four forms of consistency or equality which need distinguishing. These forms are as follows:

1. The maintenance of standards from year to year in a given course.
2. The monitoring of 'equivalents' between course options.
3. The parity of standards between universities within subjects.
4. Parity between different subjects for nationally recognised levels of accreditation.

Implicit in the ideal of parity between all degrees lies a number of assumptions. In recent years, these assumptions have come under increasing scrutiny (Johnson, 1988).

Firstly, it is assumed that all degree programmes are internally consistent and that the same standards are applied when marking course work. Unfortunately, achieving such consistency within a programme is not always possible. In particular, Piper (1990) highlights the problem of maintaining parity when work is completed under different conditions, - such as that performed during a three hour supervised exam, and that produced as a project over a period of months.

A second problem to maintaining internal consistency arises when some of the examined work is completed before the end of year final exam. This is particularly the case in some Masterate programmes (i.e., in both Arts and Science) where internal work can constitute a large proportion of the final grade. If this work is not marked with the same consistency as the end of year exam, it is likely that degree standards will vary as a consequence. Theoretically however, it is possible to compensate for any variation in marking standards by weighing marks or by taking into account the difficulty of the material (Piper, 1990).

In higher education, the ideal of parity is also premised upon the assumption that examiners can make consistent judgements over an extended period of time. As previously discussed, Dale (1959) questioned the ability of humans to carry within themselves a valid and absolute notion of 'a marking standard'. A perusal of the honours degree results for one British University left Dale in little doubt that this was not the case. Results indicated that almost one student in every four gained a first class honours award in Applied Science, one in fifty in Commerce, yet only one in seventy in Arts. More recent research, indicates that the variation in 'first' degrees awarded between subjects is still surprisingly wide (Goldman & Widawski, 1976; Hindmarch & Bourner, 1980; Bourner & Bourner, 1985).

Implicit in the notion of parity is also the assumption that external examiners are effective in moderating the standard of marking both between and within universities. Since the 1980s, the accuracy of this assumption has increasingly been questioned. In particular, Silver and Silver (1986) challenge the ability of external examiners to monitor the comparability of grading standards and to successfully guard against any arbitrary degree differences. Despite these reservations, the external examination system is still widely regarded as one of the most effective moderating instruments available to academic institutions (Williams, 1979; Johnson, 1988).

3.6 The Role of External Examiners

In both the United States and Commonwealth countries, the external examination system is regarded as one of the major guarantees of quality and equality within higher education courses (Piper, 1985). In the university sector, external examiners uphold this position of responsibility by checking the fairness of marking and marking schemes. They will also act as adjudicator when internal examiners fail to agree on the grade awarded (Williams, 1979), and in some cases, will check examination papers for inaccuracies and ambiguities before they are used (Johnson, 1988).

Although important, Heywood (1989) regards these activities as secondary to monitoring the comparability of grading standards within higher education. This view is reiterated by Williams (1979) who believes "the purpose [of the external examiner] is generally understood to be the maintenance of similar standards between different universities" (p.162).

Given that this is the prime role of the external examiner, one might expect the calibration of marks between institutions to be a universal duty. Instead, available evidence indicates that only two-thirds (47%) of examiners regard this as the case (Piper, 1985). Although specific data is unavailable, the number of mark adjustments to have been made in response to the views of external examiners is also likely to be very few (Williams, 1979).

Quite clearly, there is much confusion about the role of external examiners and little agreement on the importance of their activities. In response to this finding, Bolger (1990) suggests that the academic community is somewhat naive in relying on external examiners to ensure grade comparability, when they themselves, fail to see this as a prominent duty. Yet, given the doubt surrounding the tenability of a common degree standard, external examiners should be commended for disregarding the maintenance of parity as one of their duties. This theme was taken up by Johnson (1988), who, having reviewed some of the evidence on exam practice writes:

> "No single person can possibly have complete insight into the sometimes subtle effects of mark aggregation, nor a 'feel' for absolute grading standards. The likelihood of any individual being able to 'absorb' a feel for the grading standards being applied by some other institution by reviewing examination papers and scripts must be even lower. And the possibility that such an individual, however experienced and long serving, could 'carry' applicable notions about universally appropriate grading standards must be remote" (p. 183).

Are we to conclude from this recital that the external examination system is simply an expensive piece of academic 'window dressing' which achieves very little? According to Johnson (1988) and Piper (1990), the answer is an unequivocal 'no', since they still believe the system has a great deal to recommend it.

Specifically, it is higher education's way of tapping deviant students back towards the centre: it may not position them exactly, but it can effectively turn them in the right direction. Similarly, the system may turn internal examiners in the right direction, so that over time, any anomalies they apply in the grading process can be redressed.

Beyond its immediate efficiency, Piper (1990) regards the external examination system as a manifestation of the great care which goes into assessing degree candidates, and is proof of a commitment to disinterested grading. Above all, the continuing presence of external examiners ensures teaching staff are constantly reminded of such fundamental questions as the original educational intent behind a course.

None of this is to say that we may be complacent for the system has not realised its full potential. Indeed, it is my own belief that many improvements might and should be made. Among other things, the external examination system could: a) increase the input of greater specialist knowledge, b) encourage individuals to keep in closer touch with both their fellow examiners and the students they purport to serve, and c) advocate a little less reliance on one's personal experiences to dictate the course of an examiner's daily work.

3.7 Conclusion

To enrich the learning process, Heywood (1989) recommends improving the methods of testing and learning we use. To better guide and direct pupils in their acquisition of knowledge, performance evaluation must also take place. In modern society, both the private and public sectors rely heavily on student evaluation to dictate the employment package graduates will receive. For this reason, the assessment procedures employed must be seen to be fair.

Unfortunately, as this chapter has shown, a legitimate and valid appraisal of student performance may unwittingly be compromised by the use of less-than-honest or biased assessment procedures. The continued use of such procedures will, in turn, place university standards of academic excellence into serious jeopardy. Unless equivalent academic standards are maintained both within and between universities, considerable diversity may exist in the grades awarded across degree courses. In Chapter four, discussion will focus on several studies to have addressed this issue.

CHAPTER FOUR

Degrees of success in different major fields

4.1 Introduction

The award of a degree with honours, is likely to accord particular benefits to the recipient, both within higher education, and the wider field of work. For this reason, the attainment of an honours degree should depend solely upon the intellectual ability and/or attainment of the student. Any other factors which might improve one's opportunity to obtain a 'good' degree must necessarily be viewed with some concern. Over the years, several studies have considered the impact of the subject studied on student degree performance.

4.2 Subject Studies

One recent study which investigates the comparability of degree awards was conducted by Johnes and Taylor (1987). From an analysis of individual departments within the university, it emerged that Science students were awarded a greater proportion of first class honours awards when compared to both Social Science and Arts students. Specifically, of the students studying Physics, Johnes and Taylor (1987) found that 40% obtained a first class honours degree, compared to 22.2% in Education and 16.1% in Arts. Between the study period of 1976-84, a great deal of stability was found in these degree results. Specifically, this was reflected in the high correlation between degree results obtained in any one year with the results obtained in other years. Of a possible 36 combinations by year, correlations ranged from 0.97 to 0.88.

Comparisons of this kind have also been made by Klug (1976) and Sear (1983). Again, results reveal large between-subject or between-institution differences in the standard of degree classes. Such reports are interesting, and play an essential role in stimulating the debate about standards, yet according to Johnson (1988), only prove that distributional differences exist between one set of degree results and another. Without taking the general calibre of student intakes into account, Johnson (1988) doubts whether any clear inferences can be drawn.

Interestingly, in the one reported study which did attempt to take initial intake differences into account, strong evidence still pointed to "a wide and systematic difference between the Social Sciences and other disciplines - the Social Sciences offering relatively fewer higher honours degrees" (Nevin, 1972). This study has not been received without criticism. One major shortfall stems from the fact that the analysis is couched in cross-tabular form (Neuman and Ziderman, 1985). As a result, the relative individual impact of particular subjects, faculties and universities on the percentage of honours degrees is not known.

To overcome this limitation, Neuman and Ziderman (1985) conducted an analysis of variance to examine more formally the relative importance of these individual factors on the award of first degrees. Longitudinal data formed the basis of this research, from four Israeli universities for the period 1979-1983. From this analysis, a considerable diversity emerged in the tendency of universities to award first degrees with distinction. Taking the average percentage for the five year period as a whole, the range in first degrees awarded by universities was from 3.4% to 19.3%

Marked differences were also found between faculties within each of the individual universities. Perhaps most striking was the revelation that the natural Science faculties tended to award more degrees with distinction than average (coefficient of +0.21), whilst the Social Sciences awarded less (-0.19). The Arts faculties did not generally differ from the overall average tendency to grant degrees with distinction.

Neuman and Ziderman (1985) selected the Social Science faculty for a more detailed analysis, and in particular, focused on the four subject departments which were common to all universities: Psychology, Education, Geography and Economics. Of these subject majors, Psychology tended to award relatively more degrees with distinction than did the other departments (coefficient of +0.99); Geography and Economics awarded less (-0.57 and -0.42 respectively), whilst Education followed the general average for departments (-0.0008).

In light of these findings, Neuman and Ziderman (1985) concluded that the receipt of an honours degree was not exclusively dependent upon the academic achievements of Israeli students. Apparently, the element of chance relating to the choice of a particular university and field of study also determined the probability of attaining a first degree with distinction. Thus, they argued that there is "a pressing need for universities in Israel, as in England (and possibly other countries too), to set their houses in order through the framing of procedures for the maintenance of common standards in the granting of degrees with distinction, both between as well as within universities" (p. 458-459).

Surprisingly, Neuman and Ziderman's thought-provoking paper has not evoked the amount of attention amongst academics that it clearly merited. Little heed also seems to have been paid to their belief that "a deeper study .... might well reveal systematic factors accounting for the differences in the tendency to award degrees with distinction" (p. 458). This is somewhat disturbing, since it would be naive to expect that class percentages will be equal for different faculties, and that degree results are an accurate reflection of the comparative ability of students.

4.3 Grading Issues

Researchers who have approached the issue of degree equivalency highlight several factors which may contribute to the observed variation in degree awards. In particular, they have attempted to account for the consistent discrepancy in the proportion of first class honours degrees awarded between Science and non-Science faculties.

One persuasive explanation for this variance is that differences exist in the entry standards between the various fields of study. It can be argued that entry qualifications signal student 'quality' and that Science faculties attract students of a higher academic calibre than other departments. For example, in the Science related fields such as Medicine, Engineering and Dentistry, student quality is often calculated by the actual or potential 'A' level performance of degree applicants. Given this as the case, it is often assumed that entry standards across subjects should be strongly and positively correlated with degree results. Evidence elsewhere suggests caution in the acceptance of this easy argument.

In one British study, Clarke (1988) attempted to correlate the 'A' level grades of degree entrants with the proportion of 'good' degrees awarded within the Science department. From his data, student quality was found to explain only a small and non-significant amount of the variation in 'good' degrees awarded. In a related study, Hindmarch and Bourner (1980) did not find Social Science courses recruited poorer quality students than those in the Science faculty, despite awarding a meagre crop of first class degrees.

Departmental size is also proposed as another variable which may impact on the percentage of first class degrees awarded. Specifically, Connolly and Smith (1986) suggest that if the final year numbers in a course are large, students may become more anonymous and receive less contact hours with staff. To test this hypothesis, they correlated the mean size of final year classes, with the proportion of 'good' degrees awarded in a sample of British universities. Although the correlation was small, it was in the expected direction.

Departmental size may explain why Bolger (1990) found the subject area Medicine/Dentistry awarded a high percentage of degrees with first class honours, as departments within this subject area are small in relation to other subjects. On the other hand, several of the departments she included under the created faculty of "Science" had large student numbers (i.e., Chemistry, Microbiology, and Zoology), and still awarded a high proportion of first class honours degrees. In Bolger's study, students were not disadvantaged in the class of degree they received as a result of departmental size. Clearly, more research must be conducted before the impact of these variables is clarified.

In Britain, Nevin (1972) considered the possibility that degree classification was related to the number of students who withdraw from a course without obtaining a degree. Specifically, he suggests that if more 'poor' quality students withdraw from a course, a higher proportion of remaining students are likely to receive good degrees. In this instance, 'poor' quality students were designated as those who withdrew because of academic failure. Once again, an analysis of data failed to reveal any significant relationship between drop-out rate and the likelihood of obtaining a good honours classification. This link can also be disputed since Science faculties tend to admit the highest proportion of non-'A'level, (and hence poorer quality) students, without having any substantial impact on wastage rates (Hindmarch & Bourner, 1980).

One explanation to have found support by several overseas studies was proposed by Bee and Dolton, (1985). They suggest that the award of degrees is largely dependent upon a 'preconceived pattern' which exists in the minds of those responsible for making the awards. Apparently, this pattern develops haphazardly, as individual departments, faculties and institutions develop largely in isolation from each other. This conclusion is also favoured by Connolly and Smith (1986) and Bolger (1990) after failing to find any acceptable explanation for the consistent variation in grades awarded by different departments. Bee and Dolton (1985) regard the impact of any idiosyncratic departmental norms on the awarding of degrees as a 'highly disturbing' possibility (p.49). Bolger (1990) goes one step further and suggests that continued research into their influence is paramount.

Finally, some researchers speculate that the reason for variance in degree standards between faculties lies in the nature of the subject matter (Dale, 1959; Nevin, 1972; Hindmarch & Bourner, 1986). For example, in Science and mathematically based courses, the student is faced with a fairly well-defined area of knowledge, which lends itself to a precise form of assessment. Essentially, the subject matter in these courses can yield questions of either a 'right' or 'wrong' nature, and therefore will generate a greater spread of marks.

In contrast, an agreed upon corpus of knowledge tends not to exist in Social Science and Arts courses. This makes awarding a totally 'right' or 'wrong' mark exceptionally difficult. In subjects such as English and History, students are predominantly required to justify their opinions and judgements in an essay form. Accordingly, there is a tendency among lecturers to compare performance, and simply designate a student's work as better or worse than another student's. Such comparison is likely to 'bunch' the distribution of grades towards the centre of a scale and makes it harder to obtain a first class grade. Using this reasoning, students should also be less likely to receive a third class pass.

The objectivity or subjectivity of subject matter would seem an obvious, yet disconcerting reason as to why differences exist in the class of degrees awarded by departments. Bolger (1990) is slow to accept the possibility of such a phenomenon, despite evidence to suggest that it is not an unlikely explanation for her own results. Specifically, she revealed that students in Science courses were more likely to obtain a first class degree than students in the Arts and Social Science faculties at both the Masterate and Bachelor with honours level. This general finding is in keeping with Dale's (1959) view that subject areas which are inclined towards essay based assessment are less likely to award a 'top' degree.

Interestingly, Bolger (1990) found no evidence to suggest that New Zealand Science students had a greater propensity to obtain a third class honours degree. Confirmation of this was however found amongst English and Welsh students. In this data set, Science subjects awarded three times as many third or fourth class honours degrees when compared to Social Science subjects, and twice as many when compared to Arts courses. Dale (1959) highlights the operation of other forces, such as an unusual selection criteria, or an exceptionally harsh examining board as potential factors which may blur the pattern of results between university departments.

Whatever the reason for Bolger's findings, enough evidence still exists to suggest that the subject matter of degree courses has a substantial impact on departmental grading standards. Having reviewed this evidence, Dale (1959) concluded that "it is time that universities examined this problem and attempted to secure a rough approximation to equity between departments and between faculties" (p. 189). Of course, any attempt to force departments to award the same proportion of 'good' degrees might cause as much inequity as the present system. This is especially so, since students are likely to differ in such characteristics as motivation, work habits and ability. Nevertheless, it is still possible to move closer towards an equitable system if some of the greatest anomalies which surround the grading of students are identified and eliminated.

CHAPTER FIVE
The Present Study

5.1 Introduction

Several British studies and more recent New Zealand research by Bolger (1990) did not discount the possibility that the awarding of student grades is based, to some extent on the nature of the subject matter. Nevin (1972) is adamant that such a factor should not go unchecked, yet little attention appears to have been paid to his recommendation. In my own research, one aspect of this phenomenon will be studied in greater detail. Specifically, the extent to which course material can be designated as 'subjective' or 'objective' is of concern, and whether or not this factor dictates the objectivity of marking course work. It seems intuitively reasonable that more 'objective' courses lend themselves to the development of exam questions with a high factual content. It also infers that the person marking such questions does not need to interpret the extent of accuracy or correctness; a response will be right or wrong regardless of the examiner's individual interests or predilection. Before a more detailed look at these issues, the terms 'objectivity' and 'subjectivity' require further clarification in relation to Science and non-Science degree courses.

5.2 The Objectivity/Subjectivity Debate

In its simplest form, a view or theory is subjective if it is contingent upon the wishes, beliefs, or experiences of the speaker, or people belonging to a particular era or cultural milieu (Gardiner, 1959). For example, the statement that "eating people is wrong" is not an indisputable truth, but simply an expression of the speaker's personal belief. Usually, the dietary preferences and attitudes of people raised in the same society will coincide. This should not however, blind us to the fact that if we encounter someone who does not share our beliefs, then there is no form of proof by which we can demonstrate his or her supposed error.

The word 'objectivity', like the word 'subjectivity', is far from being clear in its definition. Parkinson (1988) suggests that, in general, the term should be restricted to truths or facts that remain independent of the person expressing them, and independent of the time and place of expression. In addition, (Cox, 1971) regards a fact or truth as a statement about a phenomenon which is happening now, or has taken place. This statement must have been produced by following the accepted rules of investigation prescribed by authoritative judges in the domain under investigation. This is distinct from an organised, yet 're-created' narrative of some event, shaped by the subjective interests of the author.

As long as generally accepted mathematical principles are followed, the statement '6 + 3 = 9' remains correct, regardless of any cultural change over time. In the same manner, scientific data such as the molecular makeup of water and the reproductive cycle of animals remain in a relatively pure state of objectivity without requiring any interpretation of accuracy.

Attaining a high level of objectivity is of major importance in both Science and Mathematics. This is especially so when one wants to establish the theoretical significance of a recently discovered phenomenon. Because scientists deal with knowledge and data which can be empirically verified as either 'right' or 'wrong', the opportunity exists to totally reject or accept the significance of any research findings.

Among scientists and non-scientists, it is generally believed that the latter stands in a more intimate relationship with their subject matter. According to Gardiner (1959), this situation precludes them from achieving the kind of objectivity that is possible in scientific work. The subject area of History clearly exemplifies this point, since it is neither the precise, nor complete reproduction of life that people assume it to be. Instead, it is an organised written account of some past events, which are selected on the basis of interest, relevance, suitability and importance.

Since there is a practical impossibility to telling the 'whole truth', Parkinson (1988) regards the selective choosing of facts as a necessity. He argues that provided different selections remain equally in accord with reality, no indictment against objectivity will be made. Parkinson (1988) goes onto suggest that scientists could not derive theoretical 'points of view' without also being selective in their expression of comments. Regardless of the validity of this statement, the situation still remains that scientific view points are prescribed by the state of the subject matter itself. In the non-Sciences, points of view are determined from outside sources, and reflect the personal perception of worth and value of the speaker. Such assumptions of worth and preference are inherently disputable. What is of significance to an historian belonging to one era or milieu may seem unworthy of mention to another, whose background and time will invariably influence the presentation of material.

A conclusion frequently drawn from all this, is that History is infected by some kind of radical and irremediable subjectivity (Gardiner, 1959). Many historians reject this claim on the basis that historical data such as dates, numbers, and persons have nothing essentially to do with the subjective values or attitudes of the historian (Walsh, 1960). Similarly, certain standard documentary records of specific events are empirical evidence, verifiable by investigation only.

Of course, certain raw historical data can exist independently of the interests and personality of the author. When translated into statements concerning human behaviour however, such data is subject to all the vagaries of the english language. Phrases such as "he stated", "he shouted", "he announced", and "he reiterated" may all be attempts to report the same historical event, but each expresses a different interpretation of that event.

Descriptions of people, and what they think and feel, necessarily commit the writer to introducing language that evaluates human life and experience. For this reason, historical accounts, as well as english and philosophical documents are never indisputably 'right' or 'wrong'. Instead, they can be classified as merely better or worse than another narrative of the same phenomenon, based on broadly identical sources.

To gain the greatest accuracy in their assessments, many non-Science lecturers rely heavily on the use of educational tests that provide a broad subject matter coverage and yield both valid and reliable scores (Gronlund, 1985). Among academic staff, two of the most predominately used exam questions are the essay and objective items. Both types serve a useful function for assessment purposes, and should be administered on the basis of their suitability to course material.

According to Anastasi (1988), the inclusion of at least some objective test items is always desirable, to ensure a complete content coverage and reliable evaluation of performance is achieved. When properly written, objective items can be applied to information ranging from simple to complex, allow both rapid and uniform scoring, and reduce the contribution of chance or luck to the total exam score (Carey, 1988).

Amongst Social Science and Arts lecturers, the essay question is also widely used (Heywood, 1989). However, since the subject matter of these courses is largely based on refutable assumptions, the essay will invariably generate subjectivity in the scoring procedure. To improve the adequacy of essay evaluation, Anastasi (1988) recommends the introduction of very systematic scoring techniques. Lecturers are advised to specify in advance the points to be covered by an answer and decide on what weighting to assign particular factors. The preparation of a model answer also helps, as does involving different examiners to mark the same scripts. Ideally, students should also be aware of the criteria for assessment.

By failing to utilise a range of practical scoring devices, the numerous sources of error in the examination system may make all the difference between a pass and a failing grade. Yet, having said this, no amount of 'tinkering' with exam procedure can eliminate the subjective judgements made by teaching staff in the assessment situation. Non-scientific documents are never indisputably 'right' or 'wrong', and may in fact, be graded on the basis of factors apart from the content, coherency and compositional skill of the student (Hales & Tokar, 1975; Bull & Stevens, 1979; Chase, 1979).

Extraneous grading factors would not be tolerated in a scientific context, and imply that a purely objective or 'value-free' marking schedule is largely unattainable in Arts and Social Science subjects (Gardiner, 1959). At the same time, one can not discount the importance of tapping important, albeit subjective, higher order skills (e.g., creative writing, the critical evaluation of arguments, and the application of methodology to new environmental situations).

5.3 The Nature of the Present Research

In particular, my research is concerned with determining whether a difference exists in the perceived objectivity of Science and non-Science course content. Secondly, whether academic performance can be rated more objectively in Science subject groups than Social Science and Arts courses will be addressed.

To determine the extent to which New Zealand lecturers believe courses fall into an 'objective' or 'subjective' category, a statistical analysis of the ratings given to 30 degree courses is conducted. This information is considered in relation to several other factors such as the faculty and course taken by participants, and the perceived objectivity of course assessment.

My study is not a replication of any previous research. As mentioned however, several published studies speculate on the possibility that subject matter could influence the distribution of degrees, both within faculties and between universities (Dale, 1959; Nevin, 1972; Hindmarch & Bourner, 1980; Kornbrot, 1987; Bolger, 1990). These studies are useful since they address the area of interest from a theoretical perspective, yet do not provide any data to support their propositions.

Moreover, unless otherwise stated, all previous investigations into differential grading standards have involved British subjects. Although similar in many respects, it seems reasonable to expect that differences do exist in the structure and content of degree courses in Britain and New Zealand. For this reason, any evidence that grade variance is dependent upon the subject matter of degree courses in Britain, may not accurately reflect the current situation in New Zealand universities.

My research is the first to address the issue of perceived degree objectivity, from the perspective of Science and non-Science lecturers. It is an exploratory piece of work, and deals with degree courses taught by New Zealand academic staff. The variables explored in this study are as follows:

1. Degree course offered
2. Subjective or objective values of the participants
3. Perceived objectivity of course content
4. Perceived objectivity of course assessment

The major objective of my research is to determine the relationship between the Science or non-Science nature of a course and the three other variables. The null hypothesis for this research is that university lecturers do not perceive any difference exists in the objectivity of degree course content and that student assessment is highly objective regardless of course content. The null hypothesis also assumes that Science and non-Science lecturers do not differ in their personal degree of objectivity or subjectivity.

A secondary consideration in this research is whether or not any significant relationships exist between the independent variables. For example, is academic performance more objectively assessed in courses with a highly objective subject matter? Do personality differences in subjectivity and objectivity cause lecturers to perceive course assessment as more or less objective? Do ratings of objectivity differ between participants inside and outside particular fields of study?

CHAPTER SIX
Methodology and Results

6.1 Subjects

Academic staff at Massey University in New Zealand participated in my research. From this population, a random sample of Professors, Senior lecturers and lecturers were invited to complete the three study questionnaires. This sample was targeted given their likely experience in postgraduate assessment, and the awarding of Bachelor with honours or Masters degrees.

Of the 300 academic staff members who were contacted, a total of 131 agreed to participate. Respondents were drawn from twenty-three different departments and six different faculties (i.e., English, Social Sciences, Business Studies, Science, Agriculture/Horticulture and Engineering). Of the participants whose gender was known, a total of 90 (68.7%) subjects were male, and 31 (23.6%) were female. The gender of 10 (7.7%) participants was not known.

6.2 Procedure

The information which formed the basis of this study was extracted from two brief questionnaires, which were estimated to take no longer than half and hour to complete. For the first questionnaire, subjects were asked to complete a 25-item Objectivity/Subjectivity scale which involved a series of hypothetical situations. The subjects' task was to imagine they were involved in each situation and indicate how s/he felt about it. The personal values (VALUES) of participants were coded on a one to five rating scale.

For the second questionnaire, subjects were asked to estimate the objectivity of the subject matter associated with their degree course. Ratings were to be made on a six-point rating scale. The subject matter of 29 other Science and non-Science degree courses was rated in the same manner. Questionnaire ratings were coded by subject headings (see Appendix 1).

In addition, participants were asked to supply a written statement outlining the objectivity of assessment in their own teaching domain. Statements were awarded a mark of one to six according to the degree of objectivity. If a student's performance could be objectively assessed 'almost totally' or 'to a very large extent', a rating of '6' was awarded. On the other hand, if objectivity was only 'moderately' possible, or was dependent upon the presence of certain factors (i.e., the course level or the type of assessment tools being used) a rating of '3' was more appropriate. In situations where objectivity was achievable to 'a limited extent' or was 'currently impossible' ratings of '2' or '1' were given respectively. In each case, ratings were made independently of how respondents marked their course for subject matter objectivity.

Subjects were also asked to supply their name (I.D), their gender (GENDER), and the department (DISCIPLINE) and faculty (FACULTY) in which they taught. (See Appendix 2 for variable codes). To combine this information, the computer programme Word Perfect 5.1 (WordPerfect Corporation, 1989) was used.

6.3 Analysis

The statistical package SPSS-PC version 3.1 (spss Inc., 1988) was used to analyze the data. Prior to generating any results, the data was double checked and corrected for inaccuracies. Once completed, research findings were analysed over several steps.

6.3.1 Step One - Univariate analysis

Univariate information, in the form of variable frequencies were obtained to determine the characteristics of the sample studied. In this phase of analysis, several different types of information were of interest, such as: the number of scientific or non-scientific degree courses included in the study, the ratio of male to female respondents, and the number of individuals who taught within each subject area.

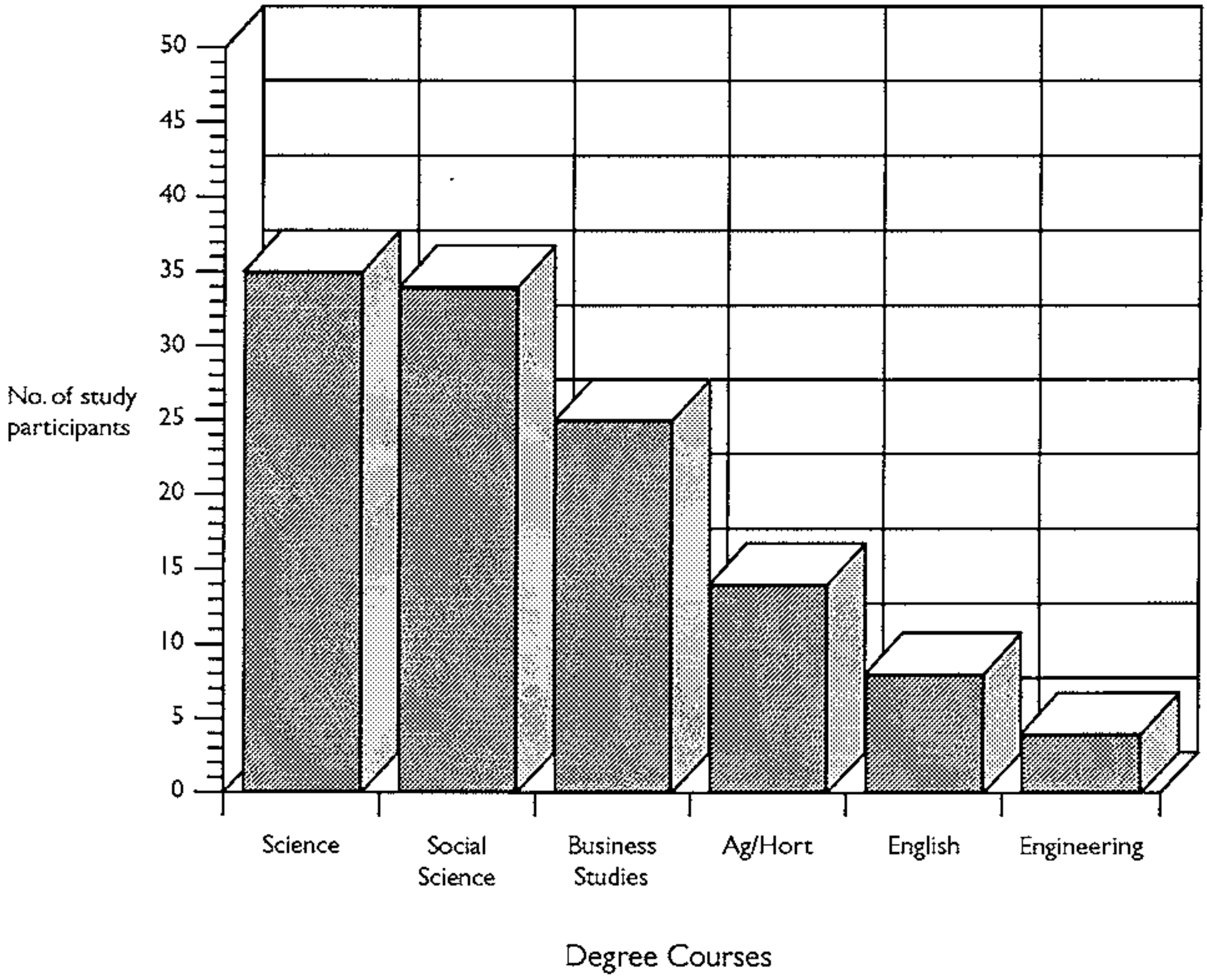

Figure 6.1: The distribution of study participants across Massey University faculties

Figure 6.1 shows the frequency distribution of the faculty in which each participant taught. It is apparent that the number of lecturers who responded to research questionnaires were not equally distributed across the university. Taken together, Science and Social Science courses account for over half of the respondents (53%), whereas Arts and Engineering courses are represented by less than 10% of the total sample.

Table 6.1 shows a crosstabulation of the degree course and gender of each study participant. Of the 121 lecturers whose gender was known, 74.4% were males and 25.6% were females. Slightly more than half the participants (55.4%) were non-Science lecturers, whereas less than half (44.6) took Science courses. Interestingly, males were distributed almost equally amongst the Sciences (36.4%) and the non-Sciences (38.0%). Female lecturers appear to be better represented (17.3%) in the non-Sciences.

Table 6.1: A crosstabulation of the gender and degree course of study participants.

	Science	Non-Science	Total (By gender)
Male	44	46	90
Female	10	21	31
Total (By degree course)	54	67	121

6.3.2 Step Two - A measure of subjectivity and objectivity

In step two of analysis, an estimate of individuals' tolerance for imbalance was sought. This was possible by conducting a frequency check of the objectivity/subjectivity ratings made by each participant. Prior to conducting any analysis, a new variable was created to combine the total scale ratings by academic staff. This variable was labelled TOTOBJ. Responses by Science (SCIENCE) and non-Science (NONSCI) staff were then identified and compared to determine whether a significant difference existed in their tolerance for situational imbalance. Questionnaire ratings were weighed so that a low score would indicate the least tolerant individuals.

Previous research by Blass (1974) found that a great deal of the individual variation in tolerance for imbalance could be accounted for by one's relative degree of subjectivity or objectivity. That is, what distinguishes an objective from a subjective person is the former's ability to tolerate imbalance and the latter's inability to do so. Thus, the situation in which a well-liked television presenter endorses a disliked idealogy is a problem mainly for the subjective individual. For this person, some kind of attitude change must occur, either towards the television presenter or his political views. For the objective individual, however, the imbalance is tolerable or may not even be experienced.

As mentioned, the objectivity\subjectivity distinction is measurable on the VALUES scale. One would expect that objective scorers on the scale would show significantly less attitude change when confronted by an irrelevant imbalanced situation than subjective individuals. This personality distinction is an issue when one asks the question: Do Science and non-Science lecturers differ in their evaluation of certain everyday experiences, and to what extent does this difference influence their perception of objectivity in course assessment? Objectivity is of concern here, since lecturer's evaluative behaviour should be unrelated to their prediction of course objectivity.

An estimate of lecturers' subjective or objective bias is shown in Table 6.2. Of the staff who responded to this questionnaire, 62 were linked with non-Science departments, whilst 60 were from the Science field. For non-Science staff, the mean rating of personality objectivity was 68.7, compared to 72.6 for Science lecturers. As mentioned, a low score on the VALUES scale would reveal the least objective group of respondents. When a t-test was conducted on these two independent groups however, this difference was not found to be significant (t = -1.84, df = 120, p = .069).

Table 6.2: A measure of interpersonal objectivity and subjectivity between Science and non-Science lecturers.

	No. of Cases	Mean	Standard Deviation	Standard Error
Science	60	68.7	10.7	1.36
Non-Science	62	72.6	12.9	1.67

6.3.3 Step Three - Degree course ratings of objectivity

Information was also sought on each degree course variable. Of particular concern, was identifying the number of participants who had rated the content of each course as being highly objective. To measure this effectively, a frequency check was conducted on the ratings assigned to each of the 30 degree variables.

The average mean ratings of objectivity for each degree course studied are shown in Figure 6.2. Courses have been grouped according to faculty. Of the six faculties represented, the Science faculty included degree subjects which had the most objective course content. Courses in the Business Studies and Social Science faculties fell in the mid-range in terms of objectivity, whilst Arts oriented courses were rated as having the least objective subject matter.

Between the Science and Arts degree courses, a mean difference of 2.43 was recorded in the objectivity ratings. A less substantial difference of 1.83 occurred between the Science and Business courses. Overall however, courses in the Science and Engineering faculties were rated as being the least different in terms of objectivity (.04).

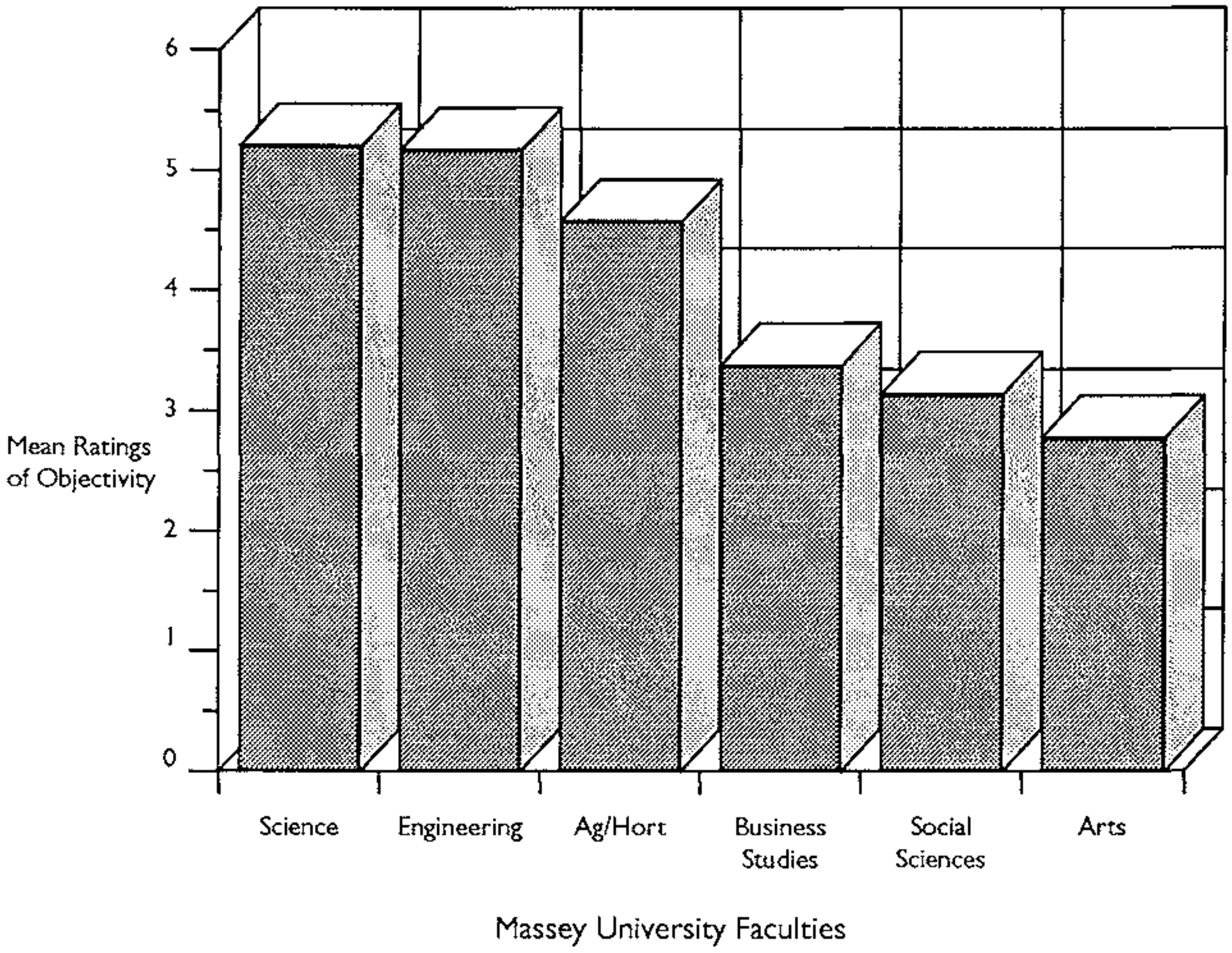

Figure 6.2: The average mean ratings of objectivity for 30 university degree course made by lecturers inside and outside the Business Study faculty.

6.3.4 Step Four - Simple variable correlations

In step four, a series of simple correlations were conducted between pairs of degree course objectivity ratings. By doing so, it was possible to measure the direction and size of association between degree subjects, and identify courses which were rated as having a similar level of objectivity. For example, if two courses are perceived as having highly objective subject matter, one would expect them to have a high, positive correlation. Conversely, a negative correlation might exist between a highly objective degree course and another with very marginally objective course material.

In Table 6.3 a sample of simple correlations are shown between Science and non-Science degree courses. These paired relationships were some of the most significant to occur between the degree courses studied, and were selected from a total pool of 900 correlations. Of all possible relationships, the association between Botany and Microbiology (.79) was the strongest, and highly significant (p = .000). Biochemistry and Botany, Physics and Chemistry were also rated as being very closely related in terms of subject matter objectivity (.77) and had a highly reliable association (p = .000).

Interestingly, the strength of association between non-Science degree courses were not as great as those between the Science courses. Of the subjects studied, Human Resource Management and Management Studies were rated as being the most closely associated in terms of course objectivity (.65). The relationship between Geography and Education (.34) and Sociology and Economics (.23) have less strength, yet still have reliability and statistical significance (p = .000 and .004 respectively).

A number of degree courses are not closely associated as shown in Table 6.4. Of all courses studied, Chemistry and Sociology were regarded as only marginally related (-.004) in terms of course content objectivity. Physics and English (.008), and Veterinary Science and Social Work (.006) also differed widely in subject matter objectivity. As such, these relationships all lacked statistical significance (p = .418, p = .463 and .474 respectively).

Table 6.3: A simple correlation matrix showing highly associated Science and non-Science degree pairs.

	Chemistry	Vet.Sci	Botany	Engineering	Physiology
Physics	.77	.58	.68	.62	.52
Microbiology	.70	.60	.79	.63	.70
Biochemistry	.75	.70	.77	.65	.69
Geology	.56	.67	.76	.65	.70
	Geography	**HR Man.**	**Law**	**Soc.Work**	**Economics**
Education	.34	.51	.21	.53	.32
Psychology	.17	.35	.29	.36	.20
Man. Studies	.17	.65	.32	.41	.40
Sociology	.20	.52	.24	.61	.23

Table 6.4: A simple correlation matrix showing the relationship between pairs of Science and non-Science degree courses.

	Psychology	Soc. Work	Man. Studies	Sociology	English
Physics	.012	-.047	.014	-.037	.008
Mathematics	-.070	-.050	-.012	-.113	.015
Chemistry	.069	-.012	-.041	-.004	-.030
Vet. Science	.118	.006	.003	.091	.008

6.3.5 Step Five - Degree course and assessment objectivity

Due to the exploratory nature of my research, the focus of interest was on global rather than specific variable differences. For this reason, several of the original variables were reduced to a smaller number of categories. Firstly, the variable FACULTY was collapsed into two categories (Science and non-Science), and relabelled SUBDOES. To determine which degree courses should be assigned a Science or non-Science label, the grouping utilised by Bolger (1990) were consulted. By conducting a t-test using the variables SUBDOES and CONTENT it was possible to explore the extent to which assessment objectivity was achieved in Science and non-Science degree courses.

Table 6.5 shows the average ratings of assessment objectivity for Science and non-Science degree courses. Group one denotes the Science subjects, whilst group two includes the non-Sciences. Results indicate that on average, Science subjects were perceived as allowing greater assessment objectivity than non-Science courses. Specifically, on a rating scale of one to six, Science courses were marked objectively to a greater than average extent (4.96), whereas non-Science courses only allowed slightly above average objectivity in assessment (3.2). These results were found to be highly significant (t = 10.05, df = 114, p = .000).

Table 6.5: The average ratings of assessment objectivity for Science and non-Science degree courses.

Subject Group	No. of Cases	Mean Rating	Standard Deviation	Standard Error
Science	53	4.96	.960	.132
Non-Science	63	3.21	.919	.116

6.3.6 Step Six - Departmental differences in course objectivity

In step six of analysis, a series of new degree course variables were created. In the first instance, the subject Accounting (ACCOUNT) was divided into two new categories which were labelled as SELFACC and NONACC. These contained important information about the perceived objectivity of Accounting course subject matter. To identify the objectivity ratings made by Business Study participants, a frequency check was conducted on the SELFACC variable. The same anaylsis was carried out with participants from outside the faculty (NONACC). At this point, it was appropriate to conduct a t-test with the new ACCOUNT variables. This would determine whether or not the Accounting ratings made by Business Study staff were significantly different from other lecturers.

In Figure 6.3 the objectivity ratings for the Accounting degree course have been highlighted. Most notable is the fact that ratings vary widely between lecturers inside and outside the Business Study faculty. Whereas 33% of Business lecturers rated Accounting as having an average level of subject matter objectivity, only 15% of other participants felt this was the case. Unlike their Business colleagues, 39% of lecturers outside the faculty rated Accounting at level '5' in terms of course objectivity. A further 19% believed that total objectivity was possible in this teaching domain. Amongst Business Study participants, no level '6' ratings were awarded. Indeed, only one person in this sample felt that an above average rating of '5' was attainable. As a group, Business Study lecturers rated Accounting as slightly above average (3.17) in terms of subject matter objectivity. Outside the Business faculty, the Accounting course is believed to have a much greater level of objectivity (4.57). The results for these two independent groups were found to be significant at the $p = .001$ level, $t = -4.16$, $df = 13.16$.

In the sixth step of analysis, a t-test was also calculated for the degree course Biochemistry (BIOCHEM). Once again, this variable was divided into two new categories to distinguish between the subject ratings of participants within the faculty (SELFCHM) and those outside it (NONBCHM). Similar analyses were conducted on Education (EDUC) and Psychology (PSYCH) ratings.

Figure 6.4 presents the objectivity ratings for the Education, Psychology and Biochemistry degree courses. From this, it is apparent that a discrepancy exists in

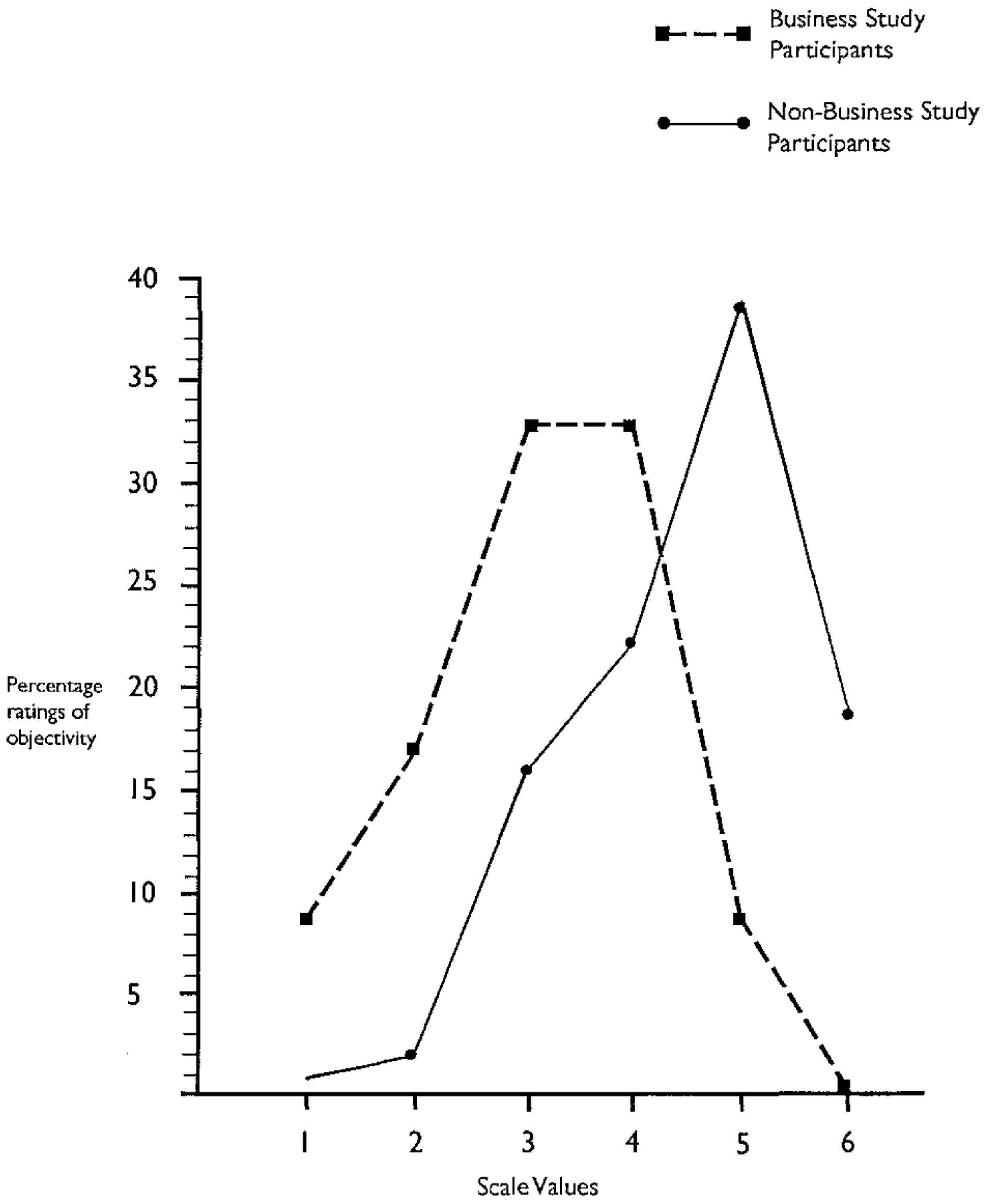

Figure 6.3: The percentage ratings of objectivity for the Accounting degree course made by lecturers inside and outside the Business Study faculty.

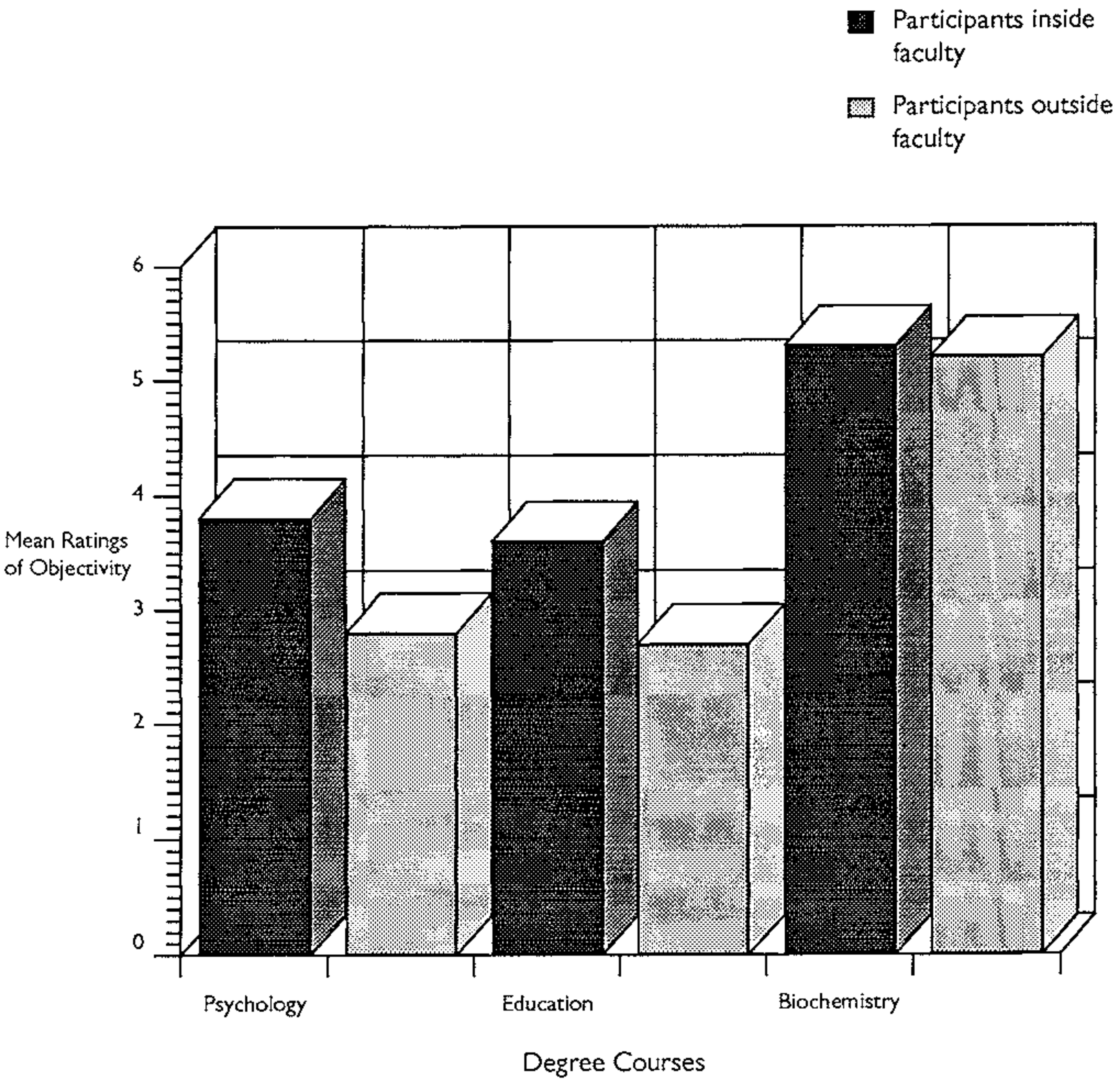

Figure 6.4: The mean ratings of objectivity for selected degree courses made by lecturers inside and outside each faculty.

lecturers' perception of course objectivity for the two non-Science degrees. With regard to Psychology, a mean difference of .96 was recorded in the objectivity ratings of lecturers inside and outside the faculty. Whereas Psychology staff believed course content had an above average level of objectivity (3.79), non-psychologists rated course material as substantially less objective (2.83). The t-test found this difference was significant at the p = .009 level, t = 3.20, df = 10.42.

The objectivity ratings made by educationalists and non-educationalists were less differentiated (.85), but nevertheless were still significant (p = .001, t = 3.79, df = 19). Once again, it was staff inside the faculty who felt their course had more objective content (3.57) than did others outside the field (2.72).

With regard to the degree course Biochemistry, a meagre difference of .13 was recorded between the objectivity ratings of staff inside and outside the Biochemistry faculty. Both groups felt that course content was highly objective, even though non-biochemists were slightly more conservative in their rating (5.20 compared to 5.33 by biochemists.). The t-test for these independent samples did not find this difference to be significant (t = .69, df = 11.73, p = .503).

6.4 Issues

In the process of collecting research data, a small minority of participants expressed their concern over the subject matter of my study or aspects of its methodology. An overwhelming number of complainants were from the Arts faculty, many of whom also choose not to complete study questionnaires. Several issues raised in relation to my research, were commonly held concerns amongst 'literary' staff members.

In the first section of my study, participants were asked to comment on how thoroughly a piece of work could be objectively assessed in their teaching domain. Some lecturers felt that any discussion of this kind was both confusing and inappropriate. Specifically, one History participant felt that:

> ".... If you are seeking information on assessment, I would have to say that any answer couched in terms of objectivity is inappropriate. We encourage students to develop skills of analysis and argument, and assess the quality of that analysis and argument. Students are not assessed on whether their history is 'right' or 'wrong'."

Later, this same participant went onto discuss the importance of collecting data on the specific forms of assessment used by the department. He regarded this information as central to any research into graduate performance. He wrote:

> "To find out how we assess students' work you need to know whether we use some relative as opposed to criterion referenced assessment, and whether there is scaling to guidelines or 'norms'.... I cannot see how you can comment on postgraduate grading and performance unless you find out how and on what basis we assess students' work".

In the second section of my study, participants were required to indicate how pleasurable a series of hypothetical situations would make them feel. Each situation was a separate and independent statement, to which individuals had to respond with a one to five rating.

Of the three questionnaires involved in my study, this was the one which generated the most feedback from participants who opposed my research. Some of the more common concerns are highlighted below. One individual who found the exercise difficult to complete made the following comment:

> ".... I do not have pleasant or unpleasant feelings about any of the situations outlined. I accept life for what it is - if the other person has a different point of view to mine that's fine - they are entitled to it. All the situations outlined are part of life and must be accepted as such".

Another participant was confused about the two-part nature of each statement. He wrote:

> ".... I cannot understand what is needed. Each statement has two pairs, so I wasn't sure which part I was supposed to react to".

In replying to each statement, one participant made a series of comments alongside each response. At the bottom of the questionnaire she wrote:

> "As you can see, the score will always depend on whether or not I'm prepared to accept a conflicting situation, and to what extent I am prepared to change to remove the stress".

For the third questionnaire in my study, participants were required to indicate the subject matter objectivity of a series of degree courses. Ratings were made on a one to six scale. This questionnaire was accurately completed by a large number of study participants. However, there still remained a small minority of staff who were opposed to assigning degree course content a rating of objectivity. Specifically, one participant from the English Department wrote that:

> "I can no longer get myself to believe that anything is either subjective or objective, for in a post-Einsteinian world, such terms are meaningless".

Another participant from the Agricultural/Horticultural Science faculty was also opposed to completing the exercise. He wrote:

> ".... I don't believe that it's possible to gain an accurate perception of the objectivity or subjectivity of degree course content by using a six point rating scale. It seems likely that this device will only serve to hide important differences between and within degree courses".

It is both relevant and appropriate to discuss the basis of these concerns, and clarify the misperceptions which surrounded my study. In Chapter seven, the preceding comments are discussed in relation to each sub-section of my research.

CHAPTER SEVEN
Discussion

7.1 Introduction:

Of the three hypothesis considered in my research, two were supported. University lecturers do recognise that a difference exists in the objectivity of degree course content, and are aware that academic performance is more objectively assessed in some degree courses. In particular, Science subjects were found to lend themselves to greater assessment objectivity than the non-Sciences. Interestingly however, perceptions of objectivity differ significantly between lecturers inside and outside non-Science subjects. Research findings did not support the third hypothesis in my study. That is, a difference does not exist between Science and non-Science lecturers in their personal subjectivity or objectivity. Science lecturers were no more or less objective in their evaluation of persons or objects than non-Science lecturers. As such, it was impossible to determine whether or not this aspect of personality created different perceptions of objectivity in course assessment.

7.2 Degree course and subject matter objectivity

Research finding from my study have shown that a substantial variation exists in the objectivity of degree course subject matter. This is true whether objectivity ratings are measured by estimating the average rating across all study participants (see Figure 6.2, p.53), or by calculating the ratings made by respondents inside and outside particular fields of study (see Figure 6.3 and 6.4, p.56).

The variance in degree course objectivity is a factor which may be expected to affect the distribution of 'good' degrees awarded across university departments. As already mentioned however, very few studies have supported this explanation in the past (Dale, 1959; Nevin, 1972: Ford, 1977). This is somewhat concerning, given the fact that those courses which were rated as having a highly objective subject matter (see

Figure 6.2), also award the greatest proportion of first class degree passes.

In particular, pure Science courses such as Physics and Chemistry were perceived as having the greatest level of content objectivity, closely followed by Maths, Engineering and Agricultural Science. Indeed, on a six point rating scale, the first four named subjects received a rating of '5' for content objectivity, whilst Agricultural Science was rated at level 4.6.

Of the students studying Science between 1960 and 1989, Bolger (1990) revealed that 35% were awarded a first class degree. The degree courses Agriculture/Horticulture and Engineering also favoured awarding a large percentage of 'good' degrees (35.9% and 29.5% respectively).

Figure 6.2 shows that Business Studies courses fall in the mid-range in terms of subject matter objectivity. Of the six courses studied, objectivity ratings ranged from 2.7 (Human Resource Management), to 4.4 (Accounting) on a six point rating scale. These 'average' objectivity ratings also appear to have affected the proportion of Business students who graduate with first class honours. In one study, Bourner and Bourner (1985) compared the distribution of honours degree awards in Accounting with those in other subject areas. Their results revealed that the proportion of first class degrees awarded in the Engineering/Technology and Science faculties exceeded that of Accounting by a factor of seven. Even more dramatically, the probability of a first class award was over four times as likely in these groups than in Accounting.

Research findings reveal that the 'literary' Arts courses are regarded as the least objective when compared to any other subject group (see Figure 6.2). Of the courses studied, Philosophy and English were both awarded objectivity ratings of less than 2.5 on a six point rating scale. Repeatedly, degree courses which possess a low level of content objectivity are recognised as the most reluctant to award a large proportion of first class degrees. In her analysis of New Zealand grading practice, Bolger (1990) revealed that only 22.6% of Arts students graduated with a first class pass. The average for the total sample was 27.2%

7.3 Departmental differences in course objectivity

Even though a substantial difference exists in the objectivity of degree course content, academic staff do not agree as to the 'correct' level of objectivity in several fields of study. As shown in Figure 6.3 and 6.4 (p.56), a difference exists in the objectivity ratings for lecturers inside and outside the Accounting, Education and Psychology fields of study. According to research findings, Business faculty members believe the Accounting degree has less subject matter objectivity than non-Accounting staff believe is the case. On the other hand, both Education and Psychology lecturers regard their own courses as having more objective content than do academic staff outside each faculty. Of the degree courses studied, Biochemistry was the only subject which was regarded as highly objective by both Science and non-Science staff members.

This finding provides further support for Kuhn's (1962) analysis of the paradigm. By "paradigm" Kuhn refers to a body of theory which is subscribed to by all members of a field. By providing a consistent account of most of the phenomena of interest in the area, Kuhn believes that the paradigm serves an important organising function. He specifically designates the 'hard' physical and biological Sciences such as Biochemistry as paradigmatic. Since this degree course is characterised by a greater consensus about content than fields which lack a paradigm, academic staff are understandably more consistent in their ratings of objectivity.

Kuhn (1962) does not regard the Humanities as paradigmatic. Rather, he suggests that the content and method in the 'soft' non-Sciences tends to be idiosyncratic. He goes onto argue that the Social Sciences and Business areas are striving for a paradigm, but have yet to achieve one. It follows, that without a paradigm, the content of non-Science degree courses is not clearly defined. In part, this absence of subject matter clarity may explain why objectivity ratings for Accounting, Psychology and Education are dissimilar for staff inside and outside each faculty. It seems likely, if these degree courses have poorly delineated areas of study, lecturers' perceptions of content objectivity will substantially vary.

An equally plausible explanation for the variance in objectivity ratings is that academic staff differ in their understanding of specific degree course content. According to Biglan (1973a) the best source of information about the characteristics and subject matter of a particular course are the scholars trained within that field. As such, academic staff may hold a simplified view of the course content outside their own field of study. If one follows this line of argument, the only 'correct' ratings of objectivity for the Psychology, Education and Accounting fields were made by the lecturers inside these departments. This argument only holds true for non-Science degree courses, since a difference was not observed between biochemists and non-biochemists in their ratings of this field.

Future research into the issue of departmental differences in perceived course objectivity is desirable. Ideally, these studies should involve more participants than my own research, and sample a broader range of Science and non-Science degree courses. Only then, is it possible to determine whether differences exist in the ratings of objectivity by lecturers inside and outside particular fields of study.

7.4 Degree course and assessment objectivity

The results shown in Table 6.5 (p.57) provide support for the hypothesis that academic performance is more objectively assessed in courses with a highly objective subject matter. Specifically, study participants felt that Science courses could be marked objectively to a greater than average extent, whereas non-Science courses only allowed slightly above average objectivity in assessment.

A large number of lecturers involved in my study also chose to supply a written statement outlining their views on the issue of assessment objectivity. Many of their comments give credence to the finding discussed above, and pinpoint the specific ways in which subject matter can impinge on objective assessment practices. One respondent from the Physics department felt that the presence of ‘hard’ scientific fact was primarily responsible for assessment objectivity in the Science field. He wrote:

> "In Physics, the core subject matter is ‘hard’ scientific fact, which removes the influence of feelings and opinions during student assessment. When marking a piece of work, the distinction between ‘right’ and ‘wrong’ is not a matter of opinion. I think my colleagues would therefore agree that assessment is virtually 100% objective in the Physics field, or is at least possible to a very large extent".

One participant from the Business faculty was also aware that course content impacted on his ability to assess student performance in a highly objective manner. Specifically, he felt that:

> "While you may try to be as objective [in assessment] as you can, your attempts will always be reduced by the nature of the subject matter you are dealing with. In the particular course I teach, the content is not absolute or correspondent with fact, but rather, revolves around refutable judgements and opinion .... As a result, my own presuppositions, values and preferences may impinge on the assessment of a student's work".

7.5 Simple variable correlations

Results from my study also suggest that Science degree courses are similar in terms of subject matter objectivity, and are distinct from the content of non-Science degree courses. Social Science and Arts courses are also related by content, albeit less so than the Sciences (see Table 6.3, p.56). This finding is supported by Biglan's (1973a) analysis of subject matter in different academic areas. In the course of his study, Biglan collected and analyzed the responses of a large group of scholars to questions concerning the similarities among numerous degree courses. From his results, it emerged that academic staff can separate the subject matter of degree courses, and consistently use the 'hard-soft' dimension to do so. That is, they tend to categorize the 'hard', Sciences, Engineering and Agricultural courses as similar, yet distinct from the 'soft', Social Sciences and Humanities. This was found to be the case for scholars in both a university and small college setting.

It seems plausible that lecturers in my own research used Biglan's (1973a) distinction between the 'hard' (factual) Sciences and the 'soft' (judgemental) non-Sciences to rate the objectivity of subject matter. The question still remains however, as to why the Sciences have a closer association in terms of content objectivity than the non-Sciences. In particular, the physical and biological Sciences have the strongest relationship, followed by the relatively pure scientific fields of Engineering, Geology and Veterinary Science (see Table 6.3). Although still substantial, the non-Science areas of Geography, Psychology and Economics have less similar content objectivity than the Sciences. It is clear from Table 6.4 (p.56) however, that both groups are unrelated in terms of subject matter objectivity.

These results provide further empirical support for Kuhn's (1962) notion of the paradigm. It seems reasonable to expect, that if lecturers agree on the content of Science degree courses, then ratings of objectivity will tend to be highly consistent. Once again, further research into this question would be of value.

7.6 Selection of a course of study

Thus far, discussion has focused on the issue of subject matter objectivity from the perspective of university lecturers. Of course, the presence or absence of clearly defined subject matter will also create unique experiences for the student. In particular, Nevin (1972) highlights the difficulty faced by the non-Science student of entering a course in which there are not only '[intellectual] mountains to climb, but guides in manifest disagreement to cope with as well' (p. 672). It is not uncommon, (certainly in the final stages of a degree) for the non-Science student to face a deluge of lectures which contradict information presented in their first year. Often, they are also compelled to consult learned journal articles in which controversy is the norm.

In contrast, the Science student who is faced with a more precise corpus of knowledge, can be assured that course content will be reasonably similar when displayed by different lecturers. As such, they may have an intellectual mountain to climb, yet according to Nevin (1972), they will not be additionally hampered by signposts pointing in opposite directions.

These sentiments were reiterated by one participant in my research who is a lecturer in the Social Science faculty. From her comments below, she is clearly aware that Science and non-Science courses call for distinct abilities to be utilised. Unlike Nevin (1972) however, she does not imply that Social Science courses are more 'difficult' given the often inconsistent nature of course material. Specifically, she felt that:

> "In the particular course I teach, there is rarely any perfect information about the characteristics of the environment on people's activities. Students are therefore required to consult articles which may be at odds with a view presented to them in class. This is particularly so at postgraduate level, where there is not a broad measure of agreement on very much of what I teach. This of course is not a bad thing, for it forces students to establish their own views and critically weigh up available information in order to justify their position. Because many Science subjects deal with quantitative material, I'm not sure these skills are utilised to such an extent. On the other hand, students in my course are not required to demonstrate their numerate ability which is demanded in the Sciences".

At the undergraduate level, this participant also mentioned the importance of providing students with a solid base of supported material. She wrote:

> ".... At the undergraduate level, (especially in the first year) it is easier to restrict course content to fairly objective material and 'hurdles'. This is important in the sense that it provides the student with a solid base,

which they can later amend and modify".

Where there is an absence of a commonly defined set of 'rules', the undergraduate 'novice' may be thrust into a state of some confusion and frustration (Nevin, 1972). Despite this, the withdrawal rate from non-Science courses remain substantially less than in the Sciences (Heywood, 1989). Nevertheless, some researchers remain concerned that the difficulty of achieving high grades in the Humanities may subtly influence students' choice of a degree major (Goldman, Schmidt, Hewitt & Fisher, 1974; Bee & Dolton, 1985). Underlying this assumption is that students are aware of the different grading practices in various subject areas. Hewitt and Jacobs (1978) reveal that students can accurately detect subject grading standards, and will choose a degree major with standards which are comparable with their ability.

Having selected a suitable course, students also believe that degree success is an achievable aim. In one study, Bosworth and Ford (1985) asked 261 university entrants to rate their perception of degree success on a seven point rating scale. Apparently, all student groups felt that Science (5.3), Business (5.0), and Arts (5.4) courses could be successfully completed at a comparable level.

Despite the ability to accurately select a suitable degree course, students appear to be unaware of the ways in which course objectivity may impinge on their success. This is apparent in the ratings above which found that degree success was equally achievable in Science and non-Science courses. Clearly, further research must be conducted in this area. Ideally, this should focus on students' perceptions of course objectivity and whether their views impact on course selection. In addition, students should also be questioned on whether they perceive a relationship exists between subject matter objectivity and the award of postgraduate scholarships.

In one study which implies that a connection does exist, Accounting graduates were found to be less successful in obtaining postgraduate scholarships than their Science counterparts (Bourner & Bourner, 1985). Further, over the period 1979-81, less than

1.5% of Accounting graduates proceeded with academic study. As already mentioned, Business degrees have less objective subject matter than Science courses. As such, only a small proportion of students may receive a first or upper second class degree. Without a 'good' degree, the award of a postgraduate scholarship becomes increasingly unlikely (Bourner & Bourner, 1985).

Issues surrounding the award of postgraduate scholarships and continued study by graduates were not addressed in my study. Nevertheless, one female participant was also aware of the relative lack of scholarships obtained by non-Science graduates in New Zealand. She wrote:

> "I don't know whether it is a tradition or not, but there appears to be a reluctance on behalf of Social Scientists to award their students with 'high' grades. As a result, students in my faculty (i.e., Business Studies) are less likely to get awards or scholarships when grade point average is the criteria".

Since an increasing number of students are pursuing postgraduate study, it is imperative that they are informed of the possible relationship between subject matter objectivity and the award of scholarships. Consultation with students on this issue may also reveal that subtle factors influence career choice in ways academic staff have not considered.

7.7 The Objectivity/Subjectivity Debate

In Chapter five of my study, some of the central issues surrounding the objectivity/subjectivity debate were discussed in detail. There is general agreement amongst philosophical theorists that the term 'objectivity' should be restricted to facts that are correspondent with reality (Gardiner, 1959; Walsh, 1960; Flew, 1985; Parkinson, 1988). In contrast, a view or theory is labelled as 'subjective' if it is contingent upon the attitudes, beliefs or background experiences of the speaker (Lacey, 1976).

Many of the comments made by academic staff on the issue of assessment objectivity are of particular relevance in this section of my thesis. With regard to the Sciences, several staff held the view that objective assessment was an achievable goal, since the opportunity existed to totally accept or reject the accuracy of students' work. One participant who was a lecturer of Chemistry and Biochemistry felt that:

> "A high level of objectivity is possible in assessment, as it generally involves the comparison of answers with universally accepted facts. When assessing new data and [the] conclusions drawn from it, there are standard criteria to judge the validity and logic by which conclusions are reached. The range and variation of acceptable answers is usually very narrow or non-existent. It can be reasonably expected that colleagues will arrive at assessments that are similar to one's own - this is not a bit worrying, it's comforting".

A similar view was held by a participant in the school of Mathematical and Information Systems. Specifically, he believed that:

> "In Mathematics, it is possible to be quite objective as the material which is dealt with has a high factual content. I only deal with facts, models and algorithms .... There is no input required at the subjective level".

In the university system however, students are often required to do more than simply regurgitate masses of accumulated facts and figures on demand. In many instances, they must present this information in a report form and justify its application to solving current environmental or social concerns. One research participant from the Biotechnology Department also shares these views. He wrote:

> "A high level of assessment objectivity is possible [in Biotechnology], as most courses deal with scientific facts. The main areas which are less

> objective relate to the skill in report writing (i.e., putting together and expressing information), and moral issues, such as the interaction of technology with the environment".

In the Agriculture/Horticulture faculty, some research subjects also recognised the interaction between 'hard' factual course material and the assessment of subjective course requirements. One lecturer in the Soil Science Department wrote:

> "My teaching responsibilities are primarily remote sensing, geographic information systems and soil conservation. The science and technology of these subjects can be assessed with a high degree of objectivity. Some of the design considerations and applications of this information are slightly more subjective".

As argued by Kuhn (1962), it seems likely that the paradigmatic physical and biological Sciences have the most clearly delineated methods of research and validation procedures. Further, it appears from the highlighted comments above, that the fields of Biotechnology and Soil Science lack some of the objectivity of their natural Science counterparts. This is despite a grounding in pure fields of scientific study.

Even so, it is generally believed that the non-scientist stands in a more intimate relationship with their subject matter (Gardiner, 1959; Walsh, 1960). Further, the widespread suspicion is that, in producing a selective narrative, the historian and philosopher will be indulging their personal prejudices and moral, political and religious attitudes (Parkinson, 1988). For this reason, I previously argued the view that non-scientific documents are never indisputably 'right' or 'wrong', but merely better or worse than another narrative of the same phenomenon. These sentiments were reiterated by one study participant from the Education faculty. Specifically, she felt that:

> "By the use of pre-set criteria and detailed marking schemes one can try to be objective in their assessment. Yet, many of the issues in the course I teach do not lend themselves to a single 'right' answer, therefore to say that a student was 100% correct in their essay response is an impossible situation".

Many non-Science lecturers involved in my research make reference to using the essay question as an assessment device. However, as previously discussed, criticism of the essay tool is prevalent throughout academic articles and texts (Cresswell, 1986; Johnson, 1988; Heywood, 1989). To obtain the most valid and reliable measure of performance, one English lecturer highlighted the importance of clarifying the criteria for essay assessment with his students. Criteria were set in an attempt to explain the points by which assessment would proceed, and to indicate their priority of importance in the assessment process. He wrote:

> "In the courses for which I am responsible, students are made aware of the criteria they are required to meet .... Assessment then involves attempting to deal with these criteria as given, and focuses on the extent to which, and the ways in which criteria are fulfilled .... Of course, judgement about the manner and degree of fulfilment (and particularly the former), certainly involves a large element of subjectivity. There is, for instance, certainly an element of subjectivity in evaluating the finer points of 'quality' in a student's assignment. This can be kept in check by exercises in marking standards conducted with other course tutors, but it would be impossible to eliminate it entirely".

Several Education lecturers also highlighted the importance of developing and using procedures which minimise scorer subjectivity. One female staff member believed that:

> "Assessment objectivity can be achieved to a large extent in the

> education field. If the essay is used, objectivity is achieved by having a detailed marking schedule which corresponds to a comprehensive essay question (and student guidelines). Experience in assessment is also important, as is having more than one judgement of the same performance. Hence, we endeavour to have all students' work assessed by more than one judge".

By the introduction of a range of practical procedures, this participant felt that assessment objectivity was achievable when using the essay question. Other psychometric experts argue that the inherent drawbacks of the essay exam make assessment objectivity an unrealistic goal (Foster, 1985). As suggested, the disputable nature of non-scientific subject matter may be the largest single factor contributing to this point of view. Coupled with this, is the fact that essay exams may be graded on the basis of factors which are extraneous to the assessment process. Some of the irrelevant factors which can impinge on essay marking were highlighted by one English lecturer involved in my research. She wrote:

> "Frankly, assessment objectivity [in English] is very difficult - impossible if by 'objectivity' one takes the paradigm of Mathematics. Naturally, one looks for certain characteristics in written essays and exams which remain constant, but marking literary essays remains a judgement call, highly dependent on one's current expectations, the standard of performance in previously graded essays (i.e., scripts already marked in the stack), and one's own frame of mind at the time. Subsequently, I always remain open to a re-reading if a student is troubled by a mark".

At the undergraduate level, my own personal view is that assessment objectivity can be improved in the non-Sciences. Repeatedly, participants in my research have emphasised a range of procedures to achieve this goal, and have been supported by psychometric experts (Hills, 1981; Gronlund, 1985; Foster, 1985). If these recommendations are implemented, it seems likely that academic staff will be less influenced by extraneous factors in their grading of students. The situation remains however, that indisputably right or wrong assessments will not be made.

7.8 Issues

In the preceding chapter, a series of issues were raised in relation to my research. In general, these questions reflect a concern over the choice of my topic, or the particular content of the study questionnaires. In the following sub-sections, each issue is discussed in relation to the questionnaire of concern.

7.8.1 Section One - assessment objectivity

In the first section of my study, participants were asked to comment on how thoroughly a piece of work could be objectively assessed in their teaching domain. This single question was sufficiently open-ended to give lecturers the freedom to explain 'assessment objectivity' within the context of their own department. A suitable definition of 'objectivity' was given, yet beyond this, participants were not provided with any other indication as to what constituted an 'appropriate response'. From the statements provided by academic staff, it appears that this question was both relevant and sufficiently well worded for the majority of staff to interpret correctly. However, for a small minority of study participants, the issue of assessment objectivity was both confusing and inappropriate.

As noted, one participant felt that the assessment of work 'quality' was more suitable than assessing whether it was 'right' or 'wrong'. Ideally, he also felt that assessment should be based on the relevance and suitability of their arguments. By its very nature, judgements of this kind lack a high level of objectivity, since they require the marker to interpret historical discussion against one's own social and cultural assumptions of the era. As such, the passing of a 'right' or 'wrong' objective judgement is not so much inappropriate as impossible in the area of History.

Later, the same participant went onto discuss the importance of collecting data on the specific forms of assessment (ie., criterion-referenced or norm-referenced) which are used by the department. He regarded this information as central to any research into graduate performance.

In considering whether some departments unfairly award higher grades, one must indeed question the extent to which different departments assess students in the same way. Repeatedly, researchers reveal a wide variety of assessment practices at degree level, within the same subjects, in the same university (King, 1976; Thompson, 1979; Griffiths & McLone, 1984a). Such revelations have generated pleas for greater consensus on course objectives and for these to be more explicitly described (Johnson, 1988). There is a problem however, in equating the performance of students, who differ in ability and motivation across departments which use widely varying assessment procedures. Despite this difficulty, future research must necessarily look within, as well as between departments to enhance our understanding of why grading practises differ across subject areas.

This however, was not the focus of my own research. As already mentioned, this was an exploratory investigation which considered grade variation as a result of the different subject matter of degree courses. Despite support for this argument, there are still numerous variables which must be explored in relation to grading standards.

7.8.2 Section Two - A measure of subjectivity and objectivity

As already mentioned, the second study questionnaire was designed to see whether a subjective/objective distinction could be drawn between Science and non-Science lecturers. To test the validity of this distinction, each statement in the questionnaire was couched in a triad form. As noted previously, this caused some confusion for at least one participant who found it difficult to respond to a two-part statement.

The basic assumption underlying the triad statement is that a situation is less stress provoking when all three units in that scenario are in balance (Heider, 1958). The theory predicts that a situation is more preferable when, for example, person P likes both person A and his attitudes. The situation is presumably more stressful when person P likes person A, but is opposed to his views. When confronted with an imbalanced situation, Blass (1974) suggests that the individual will tend to restructure it into a balanced one. Thus, balance theory predicts that person P, would experience a state of imbalance, upon hearing that his friend argued against a philosophy he deeply believed. Person P can reinstate this balance by either adjusting his philosophy to be in line with his friend's view, or changing his attitude towards his friend.

Balance theory has received empirical support from many studies (Zajonc, 1968; Blass, 1974; Sherman & Gorkin, 1980). There is however, a great deal of variation in peoples' tolerance for imbalance. For example, in everyday life, some people can value another person's friendship, while disagreeing with his or her political views. Some people however, cannot separate the two.

In my own research, the imbalanced triadic situation was of value in assessing the extent of peoples' tolerance for imbalance. In particular, it was used to reveal whether objective scorers changed their attitude significantly less when confronted by an imbalanced situation than subjective individuals. As noted, at least one participant did not find any of the questionnaire items stress provoking. This is not uncommon, and simply highlights the person's ability to assess people and objects they are in unit with, each on their intrinsic merits. Thus, a co-worker may still have your respect, despite advocating a personal lifestyle which is different from your own.

7.8.3 Section Three - A measure of degree objectivity

As noted, the third questionnaire caused confusion for at least one participant who felt that labels such as 'subjectivity' and 'objectivity' were meaningless in relation to course content. This conclusion was based on the belief that all academic staff had access to techniques which allowed students' performance to be objectively assessed. Such a comment is somewhat surprising given the results of my own study, yet appears to have some support in selective undergraduate courses. In particular, one

Geography participant felt that at least 50% of course material could be marked with a substantial degree of objectivity by the use of objective test items. She wrote:

> "In planning, approximately 50% of the course content could be graded with almost total objectivity by using short answer, multi-choice, or true-false items. These would be suited to material which is based on historical facts about the development of theory, the basic planning paradigms developed by theorists, and the accepted 'pros and cons' of each theoretical approach".

In the Sociology Department, one staff member also felt it was possible to achieve a high level of agreement on the correctness of academic performance. Specifically, he suggested that:

> " .... If I wish, I can set a multi-choice test that is highly objective, in the sense of being able to point specifically to the 'correct' answer in the text. This is perhaps easier to achieve at entry level where basic concepts, methods and information are involved .... At the postgraduate level, sociological theses and research projects do not lend themselves to assessment of this kind".

At the postgraduate level, a large majority of assessment is essay based, and will take the form of course work, a thesis, or more commonly, a combination of the two. At this stage of university education, the ease of maintaining an objective standard in marking is questionable. Foster (1985) suggests that, in general, this is because very few departments actually have any formal marking scheme for awarding exam marks.

The close working relationship between a thesis supervisor and the students they oversee may also make the awarding of a completely impartial grade particularly difficult. This situation was found to exist for many of my own research participants when they had knowledge of such factors as student motivation, commitment and effort. At the postgraduate level, students also tend to chose their own research topics, which vary in content difficulty, and require different amounts of assistance to complete. It seems reasonable to assume, that these factors will also impinge on the maintenance of objectivity in postgraduate assessment. In future studies, consideration needs to be given to whether these factors have a differential impact on the grading practices in Science and non-Science fields.

The third questionnaire was also contested by one participant, on the grounds that it simplified the 'true' nature of degree course content. Further, he felt that a one to six rating scale would not capture subtle differences in content objectivity either between or within degree subjects. It is true that very definite limits are imposed on describing an activity or event when using the rating scale, yet it is often the best available measurement device (Henderson, 1984, Siegel & Lane, 1987). Moreover, given the exploratory nature of my research, the rating scale was a useful tool for generating global rather than specific degree course differences.

As discussed, my research has revealed some very substantial global differences in the content objectivity of degree courses. Only now, is it appropriate to look more closely at the nature of these differences and the potential impact they may have.

7.9 Recommendations for future study

As shown in Table 6.2 (see p.53), the results from my research did not provide further support for the construct of interpersonal objectivity/subjectivity. Previously, Blass (1974) found this aspect of personality was responsible for differences in interpersonal evaluative behaviour. In my own study, Science lecturers were no more or less objective in their evaluation of persons and objects than non-Science lecturers. As such, it was impossible to determine whether personality differences in objectivity is a salient issue in the assessment of student performance.

In future studies, it would be of interest to continue the exploration of the objectivity/subjectivity construct in relation to the issue of assessment practice. In these investigations, it may be appropriate to refine the Objectivity questionnaire used in my own study, and increase its face validity for university staff members. Future research should also consider the extent to which personality differences exist between departmental staff.

As previously discussed, theorists are in disagreement as to why lecturers inside and outside non-Science fields of study differ in their perception of course objectivity. To what extent this difference is due to a lack of course understanding is another interesting area for future research. As an alternative explanation, one should also consider the extent of support for Kuhn's (1962) paradigm theory. Specifically, any future research should look at whether academic staff regard the subject matter of the 'soft' non-Sciences as less differentiated than content in the 'hard' physical Sciences. Once this issue is resolved, the relationship between course clarity and subject matter objectivity can be more fully explored.

Ideally, any future research into degree course objectivity should consider the issue from the perspective of university students. In particular, research must focus on whether or not students are aware of the possible link between course content objectivity and degree success. The extent to which this relationship influences one's selection of a major field also requires greater attention. Findings by Bosworth and Ford (1985) suggest that students are uninformed as to how assessment objectivity impacts on graduate performance, and believe that success is equally achievable in Science and non-Science degree courses.

Despite empirical support, the objectivity/subjectivity debate does not fully explain the variance in graduate performance. As previously discussed, several factors have been found to contribute to the discrepancy in degree awards. To date however, the majority of these factors have surfaced from studies which only look at degree variance across university departments (Nevin, 1972; Connolly & Smith, 1986; Bolger, 1990). In future, research must look within subject areas, and focus on the impact of various measurement practices on departmental grading systems. Some areas of interest include, the impact of norm-referenced or criterion-referenced assessment in specific subject areas, the variable emphasis placed on exam tools (i.e., the unseen essay, dissertations, and practical exercises), and whether departments use scaling to guidelines or 'norms'. Several empirical investigations by Goldmand and associates (Goldmand & Hewitt, 1975; Goldman & Widawski, 1976) have shown evidence of differential grading practices in assorted major fields within the same university. Some of these differences are quite substantial. As yet, this research has not been supported by any New Zealand analysis.

From a student's point of view, it matters little how differences in the award of 'good' degrees arise. What does matter is that large differences do exist which can be consistent over time. In my own research, the impact of course subject matter on graduate performance was explored. Prospective students would be wise to carefully study this research before selecting a major field of study. For those employers who interpret degree performance as an indicator of achievement, caution is also urged. Primarily, this is because the award of a degree with distinction may be unduly influenced by the subject matter objectivity of one's degree field.

APPENDIX 1

Questionnaires Codes by Subject Headings

Subject:	**Code:**
Accounting	ACCOUNT
Agricultural Science	AGSCI
Anthropology	ANTHRO
Architecture	ARCHIT
Biochemistry	BIOCHEM
Botany	BOTANY
Business Administration	BUSADM
Chemistry	CHEM
Computer Science	COMPSCI
Economics	ECON
Education	EDUC
Engineering	ENG
English	ENGLISH
Geography	GEOG
Geology	GEOLOGY
History	HISTORY
Human Resource Management	HRMAN
Law	LAW
Management Studies	MANSTUD
Mathematics	MATHS
Microbiology	MICRO
Philosophy	PHILOSP
Physics	PHYSICS
Physiology	PHYSIOL
Political Studies	POLITST
Psychology	PSYCH
Regional Planning	REGPLAN
Social Work	SOCWRK
Sociology	SOCIOL
Veterinary Science	VET

APPENDIX 2

Study Categories and Codes

Gender	**Code**
Male	1
Female	2
Unknown	-

University Departments in Alphabetical Order

Accounting	15
Agricultural and Horticultural Systems Management	06
Biotechnology	07
Chemistry and Biochemistry	04
Computer Science	16
Ecology	12
Economics	21
Education	14
English	01
Finance	23
Geography	13
Human Resource Management	19
Marketing	20
Mathematics	02
Microbiology and Genetics	08
Philosophy	22
Plant Biology	10
Plant Science	11
Physics and Biophysics	03
Psychology	18

Sociology	17
Soil Science	09
Veterinary Clinical Sciences	05

Faculties

Humanities	01
Science	02
Agriculture and Horticulture	03
Engineering	04
Social Science	05
Business Studies	06

APPENDIX 3

'A'

MEASUREMENT OF OBJECTIVITY-SUBJECTIVITY

Directions: The following is a list of hypothetical situations, no two of which are alike. After reading each situation, try and estimate how it makes you feel; that is, how pleasant or unpleasant you find each of them. In the space beside each situation, please indicate your feelings, by assigning to each item a score of 1 to 5, the score of 1 being "unpleasant" and 5 being "pleasant". Important: The score you assign should represent your feelings about the situation as a whole. Make each situation a separate and independent judgement.

	Scale Items	Rating
1	You like person A who likes tramping. You don't like tramping.	
2	A and B are enemies. You are friends with both.	
3	You read a poem which you enjoyed. You then find out that person P, whom you dislike wrote it.	
4	You have always enjoyed listening to the works of a certain composer. Now you find out that he is a political conservative. you are a liberal.	
5	You dislike person C who likes painting. You enjoy painting.	
6	You are listening to a lecture by a person you admire, only to hear him/her attack a position you have always advocated.	
7	X and Y are brothers/sisters. You are married to X, but detest Y.	
8	Your close friend, a sculptor, has worked for five years on a statue. You see it and think it is a monstrosity.	
9	You are listening to a lecture by a person you admire, only to find yourself in opposition to a position he/she now advocates.	
10	You are in love with a man/woman whose style of dressing you find distasteful.	
11	You have always disliked person A. You now find out that A's philosophy on life is almost identical to yours.	
12	Eliot White is your favourite novelist. He has recently married a woman who is of a different religion than he is. You are against intermarriage.	
13	You listen to a speech by the head of a political party, whom you abhor, yet find yourself agreeing with a lot of what he says.	
14	You dislike a particular brand of breakfast cereal. You now read that $1 from every box of cereal sold, will be donated to a cause you deeply believe in.	
15	'You like person C who dislikes sailing. You enjoy sailing.	
16	You dislike your co-worker because he is a bigot. You boss, whom you respect, has just praised your co-worker on his job performance.	
17	You go to a cocktail party with your wife/husband. You drink, but your partner does not.	
18	You read a poem to which you react negatively. You then find out that person A, whom you like, wrote it.	
19	You enjoy eating a particular brand of Tuna fish. Now you read that it was caught by driftnet fishermen.	
20	You read a poem which you enjoyed. You then find out that that your closest friend, who also read the poem, disliked it immensely.	
21	You are in love with a man/woman who is an atheist and you are religious.	
22	You are a sculptor, who dislikes Ryan Jones, another sculptor. Jones has now produced a piece of work which you think is magnificent.	
23	You are listening to the radio and hear a new record release which you instantly like, yet the announcer doesn't say who the recording artist was.	
24	You have written a book which you know is trite and superficial.	
25	You are a political Conservative, and your wife/husband is a Liberal.	

APPENDIX 4

'B'

DEGREE OBJECTIVITY

Directions: The following is a list of university degree courses. I am interested in knowing how objective the subject matter is of each course; that is, the extent to which it deals with external 'things' or facts that can be verified. On each rating scale please indicate the degree of objectivity <u>you</u> feel each course has, by assigning to it a score of 1 to 6, the score of 1 being "low objectivity" and 6 being "high objectivity". Please circle your response.

1) **Mathematics**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

2) **English**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

3) **Chemistry**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

4) **Education**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

5) **Geography**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

6) **Psychology**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

7) History

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

8) Veterinary Science

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

9) Biochemistry

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

10) Economics

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

11) Botany

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

12) Human Resource Management

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

13) Geology

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

14) Physics

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

15) **Agricultural Science**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

16) **Anthropology**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

17) **Law**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

18) **Computer Science**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

19) **Accounting**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

20) **Microbiology**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

21) **Management Studies**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

22) **Philosophy**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

23) **Architecture**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

24) **Social Work**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

25) **Business Administration**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

26) **Sociology**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

27) **Political Studies**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

28) **Regional Planning**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

29) **Engineering**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

30) **Physiology**

1	2	3	4	5	6
low objectivity stresses internal feelings and opinions					high objectivity stresses external facts and events

REFERENCES

Ager, M., & Weltman, J. (1967). The present structure of examinations. Universities Quarterly, 21 (3), 272.

Aiken, L.R. (1963). The grading behaviour of a college faculty. Educational and Psychological Measurement, 23, 319-322.

Airasian, P.W., & Madaus, G.F. (1976). Functional types of student evaluation. In W.A Mehrens Readings in Measurement and Evaluation in Education and Psychology, (p.9-25). New York: Holt, Rinehart and Winston.

Anastasi, A. (1988). Psychological Testing (6th ed.). New York: Macmillan.

Becker, L.J. (1978). Joint effect of feedback and goal setting on performance: A field study of residential energy conservation. Journal of Applied Psychology, 63(4), 428-433.

Bee, M., & Dolton, P. (1985). Degree class and pass rate: An inter-university comparison. Higher Education Review, 17, 45-52.

Biglan, A. (1973a). The characteristics of subject matter in different academic areas. Journal of Applied Psychology, 57, 195-203.

Blanz, F., & Ghiselli, E.E. (1972). The mixed standard scale: A new rating system. Personnel Psychology, 25, 185-200.

Blass T. (1974). Measurement of objectivity-subjectivity effects on tolerance for imbalance and grades on evaluations of teachers. Psychological Reports, 34, 1199-1213.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. Journal of Educational Measurement, 22(1), 41-52.

Blum, M., & Naylor, J. (1968). Industrial Psychology. New York: Harper and Row.

Bolger, P. (1990). An exploratory study of final grades awarded to bachelor with honours and masters students, Unpublished Thesis.

Bosworth, D. & Ford, J. (1985). Perceptions of higher education by university entrants: An exploratory study. Studies in Higher Education, 10(3), 256-267.

Boud, D. (1990). Assessment and the promotion of academic values. Studies in Higher Education, 15(1), 101-110.

Bourner, J., & Bourner, T. (1985). Degrees of success in accounting. Studies in Higher Education, 10, 55-68.

Bull, R., & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. Journal of Occupational Psychology, 52, 53-59.

Burnaska, R.F., & Hollmann, T.D. (1974). An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 59, 307-312.

Carey, L.M. (1988). Measuring and Evaluating School Learning. Boston: Allyn and Bacon.

Chase, C.I. (1979). Impact of achievement expectations and handwriting quality on scoring essay tests. Journal of Educational Measurement, 16, 39-42.

Clarke, S. (1988). Another look at the degree results of men and women. Studies in Higher Education, 13(3), 315-331.

Connolly, K.J., & Smith, P.K. (1986). What makes a "good" degree: Variations between different departments. Bulletin of the British Psychological Society, 39, 48-51.

Cox, C.B. (1971). Histography. In L.C. Deighton (Ed.), The Encyclopedia of Education (Vol 4). London: Macmillan Press.

Cox, R. (1967). Examinations and higher education: A survey of the literature. Universities Quarterly, 21, 292-340.

Cresswell, M.J. (1986a). Examination grades: How many should there be? British Educational Research Journal, 12(1), 37-54.

Cummings, L.L. (1973). A field study and experimental study of the effects of two performance appraisal systems. Personnel Psychology, 26, 489-502.

Dale, R.R. (1959). University Standards. Universities Quarterly, 13, 186-195.

Flew, A. (1985). A Dictionary of Philosophy: Updated and Revised Edition. London: Macmillan.

Foster, J.J. (1985). Assessing student learning: Psychologists in blinkers? Bulletin of the British Psychological Society, 38, 370-374.

Ford, B. (1977). The dubious meaning of a first. New University Quarterly, 31(4), 396-412.

Gardiner, P. (1959). Theories of History. Glencoe, Illinois: Free Press.

Glaser, R., & Nitko, A.J. (1971). Measurement in learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement, (p. 625-670). Washington, D.C: American Council on Education.

Goldman, R.D., & Hewitt, B.N. (1975). Adaption-level as an explanation for differential standards in college grading. Journal of Educational Measurement, 12(3), 149-161.

Goldman, R.D., Schmidt, D.E., Hewitt, B.N., & Fisher, R. (1974). Grading practices in different major fields. American Education Research Journal, 11(4), 343-357.

Goldman, R.D., & Widawski, M.H. (1976). A within-subject technique for comparing college grading standards: Implications in the validity of the evaluation of college achievement. Educational and Psychological Measurement, 36, 381-390.

Griffiths, H.B., & McLone, R.R. (1984a). A critical analysis of university examinations in mathematics. Part 1: a problem of design. Educational Studies in Mathematics, 15, 291-311.

Gronlund, N.E. (1985). Measurement and Evaluation in Teaching (5th ed.). New York: Macmillan.

Guion, R.M. (1965). Personnel Testing. New York: McGraw-Hill.

Hales, L.W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 12, 115-117.

Heider, F. (1958). The Psychology of Interpersonal Relations. New York: Wiley.

Henderson, E.S. (1980). The essay in continuous assessment. Studies in Higher Education, 5(2), 197-203.

Henderson, R.I. (1984). Performance Appraisal (2nd ed.). Reston, Virginia: Prentice-Hall.

Hewitt, B.N. & Jacobs, R. (1978). Student perceptions of grading practices in different major fields. Journal of Educational Measurement, 15(3), 213-217.

Heywood, J. (1989). Assessment in Higher Education (2nd ed.). Chichester: John Wiley and Sons.

Hills, J.R. (1981). Measurement and Evaluation in the Classroom (2nd ed.). Columbus, Ohio: Charles E Merrill.

Hindmarch, A., & Bourner, T. (1980). C.N.A.A. Degrees in the social sciences: A comparative analysis. Studies in Higher Education, 5(1), 17-31.

Ilgen, D.R., Fisher, C.D., & Taylor, S.M. (1979). Consequences of individual feedback on behaviour in organisations. Journal of Applied Psychology, 64(4), 349-371.

Ivancevich, J.M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64, 502-508.

Jacobs, R.R. (1986). Numerical rating scales. In R.A. Berk (Ed.), Performance Assessment: Methods and Applications, (p.82-99). Baltimore: John Hopkins University Press.

Jenkins, G.D., & Taber, T. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. Journal of Applied Psychology, 62, 392-398.

Johnes, J., & Taylor, J. (1987). Degree quality: An investigation into differences between U.K. universities. Higher Education, 16, 581-602.

Johnson, S. (1988). Comparability in degree awards: Implications of two decades of secondary level examination research. Studies in Higher Education, 13(2), 177-187.

Kane, J.S., & Lawler, E.E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B.M. Staw (Ed.), Research in Organisational Behaviour, (p. 425-478). Greenwich, Connecticut: JAI Press.

Kay, E., Meyer, H.H., & French, J.R.P. (1965). Effects of threat in a performance appraisal interview. Journal of Applied Psychology, 49, 311-317.

Kennedy, C.N. (1990). An unbiased, standardized method for quality in student assessment at the post-secondary level. In C. Bell, & D. Harris (Eds.), Assessment and Evaluation, (p. 214-229). London: Kogan Page.

King, R. (1976). Assessment in geography: Approaches to the formulation of objectives. Studies in Higher Education, 1, 223-232.

Kingstrom, P.O., & Bass, A.R. (1981). A critical analysis of studies comparing behaviourally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.

Klug, B. (1976). To grade or not to grade. Studies in Higher Education, 1(2), 197-207.

Klug, B. (1977). The Grading Game. NUS, 1-17.

Kornbrot, D.E. (1987). Degree performance as a function of discipline studied, parental occupation and gender. Higher Education, 16, 513-534.

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Lacey, A.R. (1976). A Dictionary of Philosophy. London: Routledge and Kegan Paul.

Landy, F.J. (1989). Psychology of Work Behaviour (4th ed.). Illinois: Dorsey Press.

Landy, F.J., & Farr, J.L. (1980). Performance Rating. Psychological Bulletin, 87(1),72-107.

Landy, F.J., & Farr, J.L. (1983). The Measurement of Work Performance: Methods, Theory, and Application. New York: Academic Press.

Landy, F.J., & Trumbo, D.A. (1980). Psychology of Work Behaviour (rev. ed.). Homewood, Illinois: Dorsey Press.

Latham, G.P., Fay, C., & Saari, L. (1979). The development of behavioural observation scales for appraising the performance of foremen. Personnel Psychology, 32, 299-311.

Latham, G.P., & Wexley, K.N. (1981). Increasing Productivity through Performance Appraisal. Reading, Massachusetts: Addison-Wesley.

Lindeman, R.H., & Merenda, P.F. (1979). Educational Measurement (2nd ed.). Glenview, Illinois: Scott, Foreman and Co.

McCormick, E.J., & Ilgen, D. (1981). Industrial Psychology (7th ed.). London: George Allen and Unwin.

Milton, O., Pollio, H.R., & Eison, J.A. (1986). Making Sense of College Grades. San Francisco: Jossey-Bass.

Neuman, S., & Ziderman, A. (1985). Do universities maintain common standards in awarding 1st degrees with distinction? The case of Israel. Higher Education, 14, 447-459.

Nevin, E. (1972). How not to get a first. The Economic Journal, 82, 658-673.

Parkinson, G.H.R. (1988). An Encyclopedia of Philosophy (Ed.). London: Routledge.

Piper, D.W. (1985). Enquiry into the role of external examiners. Studies in Higher Education, 10, 331-342.

Piper, D.W. (1990). Quality control in british higher education. In C. Bell & D. Harris (Eds.), Assessment and Evaluation, (p. 1-21). London: Kogan Page.

Roizen, J., & Jepson, M. (1985). Degrees to Jobs: Employer Expectations of Higher Education. Guildford: Society for Research into Higher Education.

Sattler, J.M. (1982). Assessment of Children's Intelligence and Special Abilities (2nd ed.). Boston: Allyn and Bacon.

Sax, G. (1989). Principles of Educational and Psychological Measurement and Evaluation (3rd ed.). Belmont, California: Wadsworth.

Sear, K. (1983). The correlation between 'A' level grades and degree results in England and Wales. Higher Education, 12, 609-619.

Sherman, S.J., & Gorkin, L. (1980). Attitude bolstering when behaviour is inconsistent with central attitudes. Journal of Experimental Social Psychology, 16, 388-403.

Siegel, L., & Lane, I.M. (1987). Personnel and Organizational Psychology (2nd ed.). Homewood, Illinois: Irwin.

Silver, H., & Silver, P. (1986). The escaping answer. In Moodie, G.C. (Ed.), Standards and Criteria in Higher Education, (p.9-30). Surrey: SRHE & NFER-Nelson.

Smith, P.C. (1983). Behaviours, results, and organisational effectiveness: The problem of the criteria. In M.D. Dunette (Ed.), Handbook of Industrial and Organisational Psychology, (p.745-775). New York: John Wiley and Sons.

Stone, D.L., Gueutal, H.G., & McIntosh, B. (1984). The effects of feedback sequence and expertise of the rater on performance feedback accuracy. Personnel Psychology, 37(3), 487-506.

SPSS Inc. (1988) SPSS-PC Users Guide (2nd ed.). Chicago: McGraw-Hill.

Swain, R. (1984). The teaching of personal and social skills (PASS): A group centred perspective. In D. Rose & J. Radford, Teaching Psychology: Information and Resources. Leicester: The British Psychological Society.

Thompson, N. (1979). The assessment of candidates for degrees in Physics. Studies in Higher Education, 4(2), 169-179.

Thorndike, E.L. (1920). A constant error in psychological ratings. Journal of Applied Psychology, 4, 25-29.

Thorndike, R,M., Cunningham, G.K., Thorndike, R.L., & Hagan, E.P. (1991). Measurement and Evaluation in Psychology and Education (5th ed.). New York: Macmillan.

Wainwright, C. (1977). The Degree Merchants. Albany, Auckland: Stockton House.

Walsh, W.H. (1960). Philosophy of History: An Introduction. New York: Harper.

Wendelken, D.J., & Inn, A. (1981). Non performance influences on performance evaluation: A laboratory phenomenon? Journal of Applied Psychology, 66, 149-158.

Williams, W.F. (1979). The role of the external examiner in first degrees. Studies in Higher Education, 4, 161-168.

WordPerfect Corporation. (1989). WordPerfect Version 5.1. Utah: WordPerfect Corporation.

Zajonc, R.B. (1968). Cognitive theories of social behaviour. In G. Lindzey & E. Aronson (Eds.), Handbook of Social Psychology, (p. 320-411). Reading, Mass.: Addison-Wesley.