A PROCEDURE FOR EQUATING CURRICULUM-BASED PUBLIC EXAMINATIONS USING PROFESSIONAL JUDGMENT INFORMED BY THE PSYCHOMETRIC ANALYSIS OF RESPONSE DATA AND STUDENT SCRIPTS

JOHN LESLIE BENNETT B Math, B Ed Stud, M Ed

This thesis is presented for the degree of
Doctor of Philosophy
at the University of New South Wales

June 1998

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement is made in the text.

John Bennett

CERTIFICATE OF ORIGINALITY

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

.....

ABSTRACT

One of the greatest concerns facing those responsible for conducting large-scale educational programs is whether funds used for such purposes are leading to an increase over time, or at the very least to no decline, in the proportions of students who achieve course outcomes. In order to answer this question in educational systems around the world, students are often given an examination at the conclusion of their course. A number of methods may then be employed to equate an examination to those administered in previous years. Once this is done, it is possible to determine whether more students have reached the desired performance standards than in the past. Equating techniques generally use common items or common persons to establish links between the examinations.

In a number of high-stakes educational programs the examinations are substantial measures of the knowledge and skills students have learnt from studying courses based on traditional subject disciplines. Such curriculum-based examinations commonly employ a variety of different item types appropriate to the curriculum outcomes being assessed. While some of the items may be scored dichotomously, it is not uncommon that the majority of items are scored polytomously using an holistic scoring key. It is also usual in such cases for the examinations to be made available for public consideration after they are administered. Students use past examination papers to practise for their own examination. In such circumstances, traditional equating methods employing common items or common students can not be used.

Following a review of the literature on setting standards and equating it was decided that an Angoff-based approach would be an appropriate way to equate such examinations. It was reasoned that, a team of appropriately qualified judges could develop a set of performance standards based on one examination. These standards could then be described and exemplified using items and student responses. Once this is done it would be possible for a similarly qualified team of judges to internalise those standards and equate examinations administered in different years by determining the scores on a subsequent examination that corresponded to the standards set on the initial examination. The examinations in three courses from the New South Wales Higher School Certificate were used to test the procedure which was developed for this study.

To provide information to the judges to assist them in their task student performance data were analysed using the Extended Logistical Model, a Rasch measurement model. These data were and presented to the judges in a manner most suited to understanding how students of different ability levels had performed in the items in such comprehensive curriculum-based examinations. The feedback provided by this analysis proved to be effective in assisting judges to refine their views. A review of student scripts also assisted in this regard.

This study shows that the procedure developed for equating two curriculum-based examinations is effective. The multi-staged procedure based on the application of informed professional judgment, which utilises the Extended Logistical Model as a way of providing pertinent feedback on student performance, together with consideration of student scripts, delivers promising results when applied to a sample of courses from the NSW Higher School Certificate. The results obtained would indicate that, while certain

refinements may strengthen the process, the procedure is sufficiently flexible that it could be used with virtually any form of examination or test.

ACKNOWLEDGEMENTS

I would like to thank Associate Professor Jim Tognolini for his guidance, encouragement and technical support throughout the development of this study. I would also like to thank Professor Martin Cooper and Emeritus Professor Don Spearitt for their advice and support.

Many other people were involved in this study in a variety of ways. The teachers who were the judges gave generously of their time and expertise. Jung Lay, Maria Gibson and Heather Cooper provided the skills needed in producing the more difficult aspects of the text and layout. Barry Gordon provided editorial support. I am grateful to them all.

I would also like to acknowledge and thank the New South Wales Board of Studies and the Office of the Board of Studies for their support. In particular, I am grateful to Sam Weller (President 1994-97) and John Ward (General Manager) for their encouragement and support. I hope that the procedures developed in this study and the lessons learned will assist the Board of Studies in the task ahead.

Finally, my special thanks to those members of my family who share our house - Glen, Sally and Katie - for their continued support, understanding and tolerance.

DEDICATED TO

Edward (Ted) Bennett (1925 - 1997)
Mary Bennett

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BROAD DESCRIPTION OF THE PROBLEM

In contemporary society, there is strong support for the view that money spent on education should lead to improvements in student learning, or at least, to no declines. Thus, a public examination system that does not enable explicit judgments to be made as to whether students are achieving required standards is of limited value.

To enable such judgments to be made, and thus to exact full value from a curriculum-based examination, student performance needs to be related to some form of pre-defined standards. When these standards are expressed in terms of course outcomes, professional judgment can be used to reference student performance to these standards. That is, student performance will be presented in terms of the things students know and can do. In addition, and more importantly in the context of this thesis, the standards will enable the equating of different examinations. This will allow a direct comparison of the performances of different cohorts across time within the same course, even though they have attempted entirely different examinations.

The desire to make meaningful comparisons between the performances of groups of students who have sat for different examinations in a course on different occasions has been the focus of a number of different approaches. All are similar in that they require the examinations from the different years to be equated, or placed on a common scale. The method of equating, however, differs according to the circumstances under which the examinations are administered. For example, many of the methods commonly used

for equating examinations are not suitable for use with high-stakes, non-secure examinations.

1.2 OVERVIEW OF THE STUDY

This study develops a procedure for equating examinations, including large-scale curriculum-based examinations that contain a variety of different item types. It involves using the professional judgment of teachers, supported by extra statistical and empirical information, to set and describe standards of student performance on one examination in a course and then equate that examination and examinations administered in subsequent years, using these same performance standards.

Many techniques have been advanced for setting standards of performance in examinations. However, most methods are only suitable for use with examinations containing objective-type items that are scored dichotomously - that is, items where there are only two response categories, one of which is correct, the other incorrect. It is only relatively recently that techniques have been proposed for setting standards for examinations consisting of different item types and formats, including items where a student may receive partial credit for a partially correct response. Items that have more than two possible response categories are said to be scored polytomously, or sometimes, polytomously scored.

This study considers a range of standard-setting procedures that have been employed in the past. Building upon techniques which have been shown to be successful, a procedure is developed incorporating aspects of modern measurement theory, specifically those based on the Rasch model (Rasch, 1960/1980). A Rasch model, the

Extended Logistic Model (Andrich, 1978; Tognolini and Andrich, 1995), is used in a manner different to that used in other similar circumstances, specifically because of its potential for directly comparing student performance to curriculum outcomes.

Once a suitable standard-setting procedure has been developed and used to establish a scale of student performance in an examination, the issue then is how to equate other examinations to this scale in order to compare the cross-temporal performances of different cohorts of students who have undertaken a course of study.

One approach is to use the same examination, or at least some common items, year after year. This enables direct comparisons to be made between the performances of students in different years. This method is generally not suitable for large-scale curriculum-based examinations as it is usual for such examinations to be made public after they are administered so that students in later years can use them to prepare for the examinations they will take. In cases where the examination is not released, there can still be a problem with security. Even when examinations are given under strict invigilation, it is difficult, if not impossible, to ensure complete confidentiality. Consequently, the tests become less valid across administrations.

A second approach to equating examinations is to use Modern Test Theory to calibrate a set of items along a scale, commonly referred to as an "achievement scale". These items are usually stored in an item bank and the items withdrawn to produce test forms. The student measures observed from these examinations are represented in the metric of the underlying achievement scale, irrespective of the set of items chosen in any particular examination. Since the measures are on the same scale, the results can be

directly compared. In situations where the examination is high-stakes, consists of a relatively small number of items that are scored polytomously and contains internal choice, this method has limitations.

Another approach to equating two examinations is to have the same students sit either for both examinations, or at least some items from both examinations. Various designs can be employed which use the performances of these common students to place the two examinations on the same scale. Once this is done it is possible to compare the performances of students, irrespective of which examination they have taken.

A different approach is to set and articulate standards of student performance and then to use judges to rate the performances of various cohorts of students against these standards. It is this process of using the professional expertise of judges to compare performance over time that is the focus of this research.

In a number of major curriculum-based public examination programs, professional judgment is employed to establish cut-off scores related to different standards of performance. The General Certificate of Education (GCE) A-level examinations held in England and Wales, the Scottish Certificate of Education (SCE) examinations, the International Baccalaureate (IB) examinations and the Dutch *Voortgezet Wetenschappelijk Onderwijs* (VWO) examinations all use teams of judges to set performance standards and apply them across successive examinations. These examination programs use experienced teachers and examiners, who consider classical statistical measures and student scripts in establishing the cut-off scores.

The approaches used with these major public examinations are outlined in the next chapter, as these examination programs have much in common with the context in which this study is based. In fact, some of the activities employed in setting the cut-off scores for these examinations have been incorporated into the procedure developed for this study.

This study addresses the issue of setting standards and comparing levels of performance over time, by developing and applying a standard-setting procedure which builds upon the strengths of some of the earlier procedures and at the same time taking advantage of developments in modern measurement theory.

One of the main differences in the procedure developed in this study is that Latent Trait models, particularly those developed by the Danish mathematician George Rasch (1960 & 1980) are used to inform the judges' decision-making. Student examination response data are analysed and presented to the judges in a manner that is particularly suitable for a standard-setting exercise. It is considered that this particular application of the Rasch model provides superior advice to judges than that provided by alternate forms of statistical feedback used in other judgmental standard-setting procedures.

The study will show that a team of experienced teachers can use the procedure to set cut-off scores corresponding to standards of performance in a course. It will then show that other teams of judges with similar characteristics can internalise these performance standards from a package of materials which describes and exemplifies the standards. Having done this the judges can then impose these standards on a different examination by the establishment of cut-off scores.

The procedure for equating examinations developed in this study is sufficiently flexible to handle all forms of examination, including large-scale curriculum-based examinations containing items that are scored polytomously. In addition, it is able to set multiple performance standards corresponding to a number of different levels of student performance.

1.3 THE CONTEXT

The context for this study is the New South Wales (NSW) Higher School Certificate (HSC) examinations, which New South Wales students sit for at the end of their secondary education. These examinations are used to provide norm-referenced, or "cohort-referenced" (McGaw, 1996), measures of student achievement. The courses students take relate to the traditional subject disciplines, such as English, Mathematics, History, Physics, French, and so on. Each year over 60 000 students attempt the examinations in at least one course. The examinations are closely based on course curricula, and employ a variety of different item types, as appropriate. There are presently 148 courses available from 76 subject areas.

Most examinations consist of written response-type items that are scored polytomously. Some examinations also include multiple-choice or short-answer items. Written components generally consist of short or extended responses or the solution to a mathematics problem. However, in some courses there are other, more substantive, manifestations of student work. For example, in Visual Arts, students submit for assessment a piece of artwork they have created. In the examinations for foreign languages, items that assess listening and speaking skills are employed. In Drama and

Music, students are assessed on the quality of their performance of a short play or pieces of music they have prepared. The examinations are usually three hours in duration.

A new examination is produced for each course every year. While the general structure of an examination paper, including the number and type of the items and the maximum possible score for each item, remains similar from year to year, no item is the same as that used in any previous examination.

Courses from the subject areas of English, Mathematics and Biology from the NSW HSC are used to demonstrate the equating procedure. These are intended to be illustrative, however, as the procedure can be applied to all courses. The selected courses are very different in terms of their curricula, the nature of their examinations and the way the examinations are scored.

This research is topical as a New South Wales Government policy document, *Securing Their Future*, released in August 1997, adopted the recommendation made by McGaw (1997) that a "standard-referenced approach to assessment be adopted for the Higher School Certificate by developing achievement scales for each subject" (p. 97). McGaw recommended that examination data be used to clarify performance scales on which student achievement and item difficulties can be represented, to develop descriptors of what the scales measure in broad bands so as to amplify the meaning of the bands. McGaw's proposal approaches the task of establishing performance standards in a different way from that put forward in this study. Nevertheless, the two procedures utilise some common strategies and are aimed at achieving similar outcomes. The

NSW Government has determined that a standards-referenced approach will be first used for the Higher School Certificate examinations held in 2001.

1.4 SUMMARY

Following a search of the literature on equating examinations and setting performance standards, a procedure was devised which incorporates new features - in particular, the application of developments in modern test theory in a way which adds a new dimension to existing standard-setting procedures.

This study examines the application of this procedure and the resulting outcomes, when it is applied to set and link performance standards over time, from the examinations conducted for two consecutive years in three courses which are part of the New South Wales Higher School Certificate. The procedure is also applied to the examinations conducted in a third year as a validation measure.

CHAPTER 2

TEST EQUATING AND

COMPARING STANDARDS ACROSS TIME

2.1 INTRODUCTION

To compare the performances of students who have been examined at different times, some method for creating a link between the examinations needs to be employed. A number of techniques for doing this, and the circumstances under which they can be applied, are addressed in this chapter. Whatever approach is used, it must enable any effects on student scores due to differences in the difficulty of the examinations and differences in the application of the scoring keys to be taken into account. Only then is it possible to determine whether there are any changes in the performance levels of student groups across different years.

In order that the process of comparing the performance standards of different student cohorts is not ad hoc and haphazard, it is essential that the standards be properly set and clearly articulated. The procedure used to establish the standards of student performance to be applied needs to be appropriate for the examination and fully understood by those who will apply it. There are many procedures that can be used, including simply indicating that a nominated proportion of the candidates will be deemed to have reached an acceptable level of performance. On the other hand, some procedures, including the one used in this study, involve the use of trained judges to compare student performance against clearly expressed statements and exemplary

materials which define different levels of student achievement in terms of the key knowledge and skills components of the course.

A review of different standard-setting procedures, involving the use of judges, which have been developed and applied over the past forty years is included in this chapter. By considering approaches used in the past, it will be possible to adapt features which have proved to be successful in other circumstances to develop a standard-setting procedure suitable for a wide range of different examination types, including the type of examinations encountered in this study.

The terms "standards", "performance standards", "standards of performance" and "achievement standards" are used interchangeably in this thesis to refer to what Waltman (1997) calls "performance standards"- namely, "the description of the knowledge, skills and abilities students must have to demonstrate evidence of a specific level of competence" (ibid). The term "cut-off score" is used here, rather than "cutscore" or "cut score", and refers to "points on a score scale that form boundaries between contiguous levels of student performance". The meaning given to the process of "standard-setting" in this thesis is that used by Waltman (1997) - that is, "the method of mapping a set of performance standards onto a particular score scale (ie determining where the cutscores belong)" (p. 102). The term "item" is used to refer to any task in the examination. This includes single multiple choice items and tasks which require students to provide a response to a question or problem in a three or four page essay, or submit a significant project. Some items in the examinations of the type used in the examples of this study are sub-divided into smaller parts.

2.2 COMPARING STUDENT PERFORMANCE

2.2.1 Using the Same Examination

One method of comparing cross-temporal performances of different cohorts of students who have undertaken a course of study is to use an identical examination paper from one year to the next. Provided the scoring key developed for any items which are scored polytomously is applied with equal reliability each year, clear and accurate comparisons can be made between the relative achievements of student groups in the different years.

This approach, however, is generally not suitable for most major testing programs because it is essential that the examination paper be kept secure. It is common in large-scale, high-stakes examinations for papers to be released, after the examinations, for public consideration and for the information of future students. It would be difficult, if not impossible, to ensure the security of an examination paper for more than one or two years because examinees go to elaborate lengths to reproduce and practise on the paper so that they, or their friends, can obtain an advantage in the next year's examination.

2.2.2 Equating

If the approach of using an identical examination paper is not suitable, parallel forms of an examination may be appropriate. Parallel forms of an examination are different examination papers which measure, within acceptable limits, the same psychological functions (Angoff, 1971). In many curriculum-based examination programs, different forms of an examination paper are prepared in accordance with established examination specifications. It is not unreasonable, in such cases, to regard the examinations as being nominally parallel forms.

In spite of considerable effort and care on the part of the examiners, it is not easy to prepare examinations which are precisely equivalent to each other in terms of level and range of difficulty. It therefore becomes important to convert the system of units of one form of the examination to the system of units of the other so that the scores derived from the two forms, after conversion, will be directly equivalent. If two examinations are thus "equated", then it becomes possible to determine whether there is any difference in the standard of the performances of two groups of students, where each group takes a different form of the examination (Angoff, 1971).

Traditionally:

> "two scores, one on Form X and the other on Form Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group are equal" (Lord, 1950).

Two examinations can be equated by an area transformation, or equipercentile mapping, resulting in the distribution of scores on one examination being adjusted to match the distribution of scores of the other. Once different forms of an examination are equated, a student will receive the same converted score irrespective of the form of the examination taken.

In cases where the two examinations yield similar distributions, using a simple linear transformation of the form represented by Equation 2.1 can perform the equating.

Hence, two scores, one on each examination, can then be considered equivalent if they correspond to equal standard-score deviates,

$$\frac{Y - M_y}{S_y} = \frac{X - M_x}{S_x}$$

Equation 2.1

where M_y = mean of form Y

M_x = mean of form X

S_y = standard deviation of form Y

S_x = standard deviation of form X

Angoff (1971) states that, where the distribution shapes of the two examinations are the same, linear equating is a more suitable approach as it does not produce errors due to smoothing. Equipercentile approaches can cause serious distortions in the score range in which data are scant or erratic.

Whatever technique is used to equate two examinations, if two groups of students each undertake one of two examinations which have been equated, it is possible to compare the relative performances of the two groups.

2.2.3 Using Pre-equated Forms of an Examination

In some cases it is possible to equate different forms of an examination prior to their full administration. Such approaches usually make use of samples of students chosen for their similarity to the population to be examined. Once two or more forms of an examination have been equated, the form used can be varied from year to year. Using

parallel forms of an examination in this way can, to a large extent, overcomes concerns regarding security associated with the use of a single form of an examination.

One approach to pre-equating two forms of an examination is to administer a different form to two groups of students selected at random. The students used to equate the examination forms should be as similar as possible in age, ability range and experience to those students whose performances will be measured by the examinations. This technique is usually referred to as Angoff's Design I (Angoff, 1971).

While this method is relatively simple to use, the accuracy of the equating process is dependent on the comparability of the two groups taking each examination. If the groups differ in ability, for example, the equating will not be accurate.

Another approach which seeks to minimise the effects due to differences in the groups is Angoff's Design II (Angoff, 1971). This technique requires that two random groups be selected, but that both groups be given both forms of the examination. To eliminate any bias due to the order in which the students receive the examinations, each of the two groups is divided into two random halves. One half of each group receives Form X first, followed by Form Y. The other half of each group receives Form Y, followed by Form X. While this approach overcomes some of the weaknesses with Design I, it has the disadvantage that all students in the sample are required to take two examinations. Issues of fatigue, while likely to occur no matter how the examinations are presented, may affect the equating process.

Approaches such as these are not feasible in all situations. For many curriculum-based examinations it is not possible to find a group of students who can be used in the equating process. To maintain the security of the items so that they can be used again, it is common to use students from a different education system in the equating process. Some examinations, particularly those where there is a large choice between items, are not particularly suitable for this form of equating, as it can be difficult to get a sufficient number of responses to certain items to make the equating viable.

2.2.4 Using Common Items to Equate Examinations

Equating two examination papers, Form X and Form Y, can also be achieved by setting a number of items which are common to both forms. These common items, referred to by Angoff (1971) as Form U, must represent the same kind of task to both groups of students. The usefulness of Form U for equating depends upon the extent to which it is correlated with Form X and Form Y. Furthermore, Angoff indicates that Form U should consist of at least 20 items or 20% of the number of items in each of Form X and Form Y, whichever is the greater. "Form U should be long enough and reliable enough so that the data obtained can be used for making any fine adjustments for differences between groups that are required" (Angoff, 1971, p. 578). While referred to as Form U, as though they are a single examination or sub-section of a larger examination, the common items can actually be interspersed with the discrete items in Form X and Form Y. Form U can be part of Forms X and Y or it can be separate.

Angoff (1971) proposes four approaches to common item equating, which he refers to as Designs III, IV, V and VI. The design chosen depends upon the nature of the examinations and the circumstances under which they are administered. Design III, for

example, requires the establishment of two random groups of students. One examination is administered to each group, and a common equating examination is administered to both groups.

If it is not possible to equate different forms of an examination prior to their being administered to the population to be tested, they need to be equated using the results from their administration. In such cases, the common items (Form U) provide this link.

In cases where the two examinations to be equated are administered a year apart, it cannot be assumed that the examinees have been selected from the same population, even if they are similar in many respects. Angoff's Design IV is suitable in such circumstances. It involves administering a common equating examination to both groups in the same manner as Design III. Estimates of the mean and variance on both Form X and Form Y are made for the combined group of students. Substituting these into Equation 2.1 gives a function that relates scores on Form X to scores on Form Y.

In the application of equating techniques using common items, it is possible for Form U to consist of two different forms of an examination, each administered to a different group of students. In such cases, however, these two different forms must be expressed on the same scale.

Whereas the traditional approaches to equating discussed by Angoff (1971) make use of linear and equipercentile techniques, it is now common to use models based on Item Response Theory (or Latent Trait Theory) in the equating process. The measurement models, commonly referred to as Latent Trait models, seek to predict the outcome when

a student attempts an item. The parameters are calculated by analysing the outcomes when a sample of students attempts a set of items.

One of the models often used in this process is based on the work of Rasch (1960/1980). The Rasch model for items that are scored dichotomously uses one parameter, namely difficulty, to describe the location of an item, and one parameter, namely ability, to locate a person on the same achievement continuum. Furthermore, a student's ability is determined solely by the number of items he/she answered correctly. Hence, by analysing the data obtained when a group of students sit for an examination, it is possible to obtain a measure of the difficulty of each item and a measure of the ability of each student. These measures of difficulty and ability are on the same scale. The process of determining item difficulty is referred to as "calibration". The features of the Rasch model are discussed more fully in the next chapter, along with extensions of the model that enable the calibration of items that are scored polytomously.

Wright (1977) describes how if common items (usually referred to as links) are embedded in pairs of otherwise different examinations, each examination can then be taken by different samples of students, with no student taking more than one examination. Using the Rasch model, it is then possible to place all the items in all examinations on a common scale through a network of links.

For each item that is common to the two examinations, the application of the Rasch model produces a pair of separate and independent estimates of difficulty. The estimates for each pair are statistically equivalent except for a single constant of translation that is the same for all items which are included in both examinations. If

examinations X and Y contain a common set of k items and X and Y are given to separate samples of students, then d_{ix} and d_{iy} are the estimated difficulties of item i in each examination. The constant necessary to translate all the item difficulties of the items in examination Y onto the scale defined by examination X is

$$t_{xy} = \sum_{1}^{k} (d_{ix} - d_{iy}) / k \qquad \textbf{Equation 2.2}$$

A number of other researchers (eg Kolen, 1981; Morgan, 1982; Hills, Subhiyah and Hirsch, 1988; Huynh and Ferrara, 1994; and Harris, 1991) have conducted studies comparing the results of equating examinations using traditional processes and Latent Trait models. They found that the latter generally give superior results to those produced by traditional means.

In 1992, Mislevy, Sheehan and Wingersky (1992) reported on how, in an equating study involving Latent Trait Theory, data from other sources were used to support the equating process. They suggested that collateral information - such as content and format specifications, expert opinion, or psychological theories about the skills and strategies needed to solve problems - can be used to support the equating process. It is further suggested that this information be used to enhance, or even replace, examinee responses when linking new examinations to established scales.

Collateral information and response data information differ in one crucial respect. While the linking of two examinations by using student responses can be made arbitrarily accurate by increasing the sample size, the accuracy of the linking by using collateral data is limited by the strength of its relationship to the item operating

characteristics. Mislevy *et al* (1992) conclude that there is no guarantee that collateral information about items in a particular application will be sufficiently rich to eliminate or substantially reduce pre-testing and equating by empirical means (p. 19).

The use of common items can be an effective way of equating two examinations. In some cases it is not even necessary to use the same items in the two examinations. It is possible, in certain circumstances, to use different items selected from a calibrated item bank. Each examination form is equated to the variable defined by the item bank using the pre-calibrated items. Since both forms of the examination are equated to the item bank, they are equated to each other.

There are circumstances, however, when using common items is not suitable. In curriculum-based examinations that consist predominantly or wholly of items that are scored polytomously, it would be unsuitable to have an item in more than one examination. As the number of items students are required to answer may be relatively small, and as each examination paper is usually made public after it has been administered to guide future students, it is not possible to re-use items. Varying an item, even slightly, would mean that the original item and the variant could no longer be considered the same.

In order to equate such examinations, another approach is required. The following section looks at ways in which groups of judges have been used to equate examinations.

2.2.5 Using Judges to Equate Examinations

Teams of judges have been used to equate examinations in situations where empirical methods are not suitable for one reason or another. In such cases, it is common to use as judges those with experience in teaching the course and preparing students for the examination. Where applicable, those who have been responsible for setting the examination and scoring the students' responses are also used. As will be evident from what follows, however, in many situations a more heterogeneous group of judges has been used.

Judges create an achievement scale by defining different standards of performance and ascribe total examination scores that they believe students on the borderline between the different performance levels will achieve. Descriptive statements and other material are prepared which summarise the characteristics of students at each performance level and give meaning to the scale. Once such an achievement scale is created, judges can use the descriptors to equate the scores of subsequent examination papers to the achievement scale, thus ensuring consistent standards are employed from year to year. It is then possible to make comparisons between the performances of the different student groups who have taken the examinations.

The viability of such an approach depends very much on the process used to create the achievement scale. The scale not only needs to be meaningful for the first examination on which it is established, it also must be able to give meaning to the performances of students in subsequent examinations.

While Waltman (1997) refers to the use of judges to link two examinations as "social moderation", the more generic terms "linking" and "equating" are used in this study to establish a correspondence between scores from different examinations, irrespective of the procedure used.

Many methods exist for using the judgments of experienced professionals to establish students' performance levels in an examination. By using one of a number of methods, a team of judges establishes scores that signify the cut-off between one performance level and another following the administration of an examination. Following the administration of a second form of the examination, the same judges determine the cut-off scores for this form. If the same performance standards are applied consistently, the cut-off scores of both examinations are comparable.

It is not necessary that the same judges be used for the second administration. If different judges are used, however, it is essential that they be provided with materials and information that will clearly define and exemplify the standards which were applied on the first occasion. It is also essential that the second group of judges follow the same structured standard-setting procedure employed by the first group.

The use of experienced judges to apply common standards of performance across different years occurs in major curriculum-based examination programs conducted at the end of secondary education in a number of countries. In such programs the challenge is to have judges internalise the standards of student performance which have been established, and then apply them to different forms of the examination administered in different years. Norcini, Shea and Ping (1988), Norcini (1990) and

Norcini and Shea (1992) report on the use of judges to produce cut-off score equivalances across different forms of an examination. These studies show that such procedures can be made sufficiently accurate.

2.2.5.1 The General Certificate of Education (GCE) A-Level Examinations

In the General Certificate of Education (GCE) A-level examinations conducted in England and Wales, the process of determining cut-off scores relating to the various grades awarded involves a team of highly experienced judges who have been involved in the setting and scoring of the examination. Prior to meeting to set the cut-off scores, the judges ensure they are fully conversant with the overall standard of work associated with cut-off scores determined in previous years. As the main objectives are to maintain grade standards over time and across different subjects, question papers, scoring keys and student responses defining grade boundaries for previous examinations are reviewed in the context of relevant statistics. The examining board maintains an archive covering a number of years and containing responses awarded each cut-off score. Evidence from the first year of the examination, when the performance standards were originally set, is also retained to guide the judges in setting their cut-off scores.

The establishment of cut-off scores relating to the different grades awarded requires the judges to work as a group and take account of a variety of factors. These include the examination papers and the scoring keys, samples of student responses to the examination items, technical information relating to the examination and the items (such as facility values for multiple-choice items and mark distributions for papers), statistical information from previous years, grade descriptions, archived examination

scripts, question papers, and details of significant background changes in entry patterns and choice of options (School Curriculum and Assessment Authority, 1996).

2.2.5.2 The Scottish Certificate of Education (SCE) Examinations

In the Scottish Certificate of Education (SCE) examinations, cut-off scores corresponding to the grades awarded are set by subject experts using professional judgment and supported by statistical evidence. The statistical evidence provided includes cut-off scores and distributions of grades awarded in the previous three examinations, and the frequency distribution of students' scores on the current examination.

In order to set the cut-off scores on the examination in each course so that the same standard of performance receives the same grade every year, a meeting is held between senior officers of the Scottish Examinations Board, the Principal Examiner and other subject experts. At this meeting agreement is reached on the cut-off scores to be applied (Scottish Examination Board, 1996).

2.2.5.3 The International Baccalaureate (IB) Examinations

For the International Baccalaureate (IB) examinations, the determination of grade boundaries follows a structured process which entails using the professional judgment of a number of examiners supported by statistical data and the examination papers and samples of student responses from previous years. It is common for different teams of judges (examiners) to consider different components of the examination.

The judges responsible for setting the grade boundaries are required to become familiar with the examination paper and consider feedback provided by those who had scored

the students' work and those who had prepared students to sit the examination. Key points are noted and taken into consideration when samples of student responses are reviewed.

Histograms which show the score distribution for the various components of the examination are also provided. While these are important, the judges are reminded that they should not be used as the sole basis for determining grade boundaries.

Cut-off scores are established by considering a number of scripts produced by students which scored at and around a set of initial cut-off scores suggested by a senior examiner. Once the members of the team have settled on the cut-off scores, they are given the grade distribution percentages from previous examinations. The judges are able to make further adjustments to the cut-off scores, if they feel changes are warranted (International Baccalaureate Organisation, 1996).

2.2.5.4 The *Voortgezet Wetenschappelijk Onderwijs* (VWO) Examinations

To establish the boundary score between a "pass" and "fail" for the Dutch *Voortgezet Wetenschappelijk Onderwijs* (VWO) examinations, meetings of judges with expertise in the field are held. For those examinations that consist of objective items, the judges are consulted about the desired cut-off score before the examination is held. On the basis of a random check of students' results, the judges determine the final cut-off scores after the examination has taken place.

For an examination which consists of a mixture of objective and free-response items, or which only contains free-response items, a different procedure is applied. The scoring

key is determined before the examination is held. This scoring key is sent to schools along with the examination papers. The teachers determine the scores to be awarded to each student immediately after the examination has been held. On the basis of a random check of students' scores and scripts, the cut-off score is finalised by the judges.

In both approaches, the cut-off score may be adjusted on the basis of the actual examination results so that comparable levels of performance receive the same grade every year. The cut-off score is adjusted, for instance, in the case where there is a difference in an examination's level of difficulty compared to the examinations of previous years (CITO, 1990).

In the examination programs referred to above, attempts are made to ensure that the judges fully understand the standards they are to apply in determining the cut-off scores. Providing the judges with copies of examination papers, samples of student scripts, and a variety of statistical data does this. Whatever approach is used, in any standard-setting exercise involving the use of professional judgment to link standards over time, the standards must be articulated in a clear and meaningful manner. Norcini and Shea (1997) emphasise the need for care and rigour in relating standards across different years. "If the pass-fail decisions on the different forms are not equivalent, 'vintages' of licensed or certified professionals are created. This is unfair to examinees who might have been successful if a different form of the test had been used" (p. 48). Equally, it will be unfair in examinations of general education courses if students of comparable achievement in the course, but who sit different examinations are awarded different results.

2.3 JUDGMENTAL STANDARD-SETTING METHODS

The previous sections in this chapter discussed techniques that are used to equate different forms of an examination, including using judges. The following section looks at methods which use judges to establish standards of performance on examinations.

By considering techniques that have been used in the past and adapting them where necessary, a method of setting standards in curriculum-based examinations will be developed. Once a suitable procedure is available it will be possible to use it to set cut-off scores corresponding to standards of performance in an examination. By using basically the same procedure, it will also be possible to set cut-off scores corresponding to those same standards of performance on different forms of the examination.

In undertaking this task, however, it is well to note the challenge identified by Mills (1995):

> "the advent of tests based on complex performance assessment makes it clear that many of the issues researched over the past decade, although providing quite useful results, pale in comparison to the task of determining defensible methods of establishing passing standards on these new types of assessments" (p. 93).

Thus, any procedure for using judges to equate different forms of an examination of the type used in many contemporary assessment programs, will need to be able to handle a range of different item types and examination instruments.

2.3.1 Classifications

Judgmental methods for setting standards may be classified according to various criteria in a number of different ways. Methods which are based on the assumption that competence is a continuously distributed ability are categorised as Judgmental, Judgmental-Empirical and Empirical-Judgmental (Meskauskas, 1976; Berk, 1986). Judgmental methods are based solely on judgment, Judgmental-Empirical methods are based primarily on judgment, whereas those classed as Empirical-Judgmental are based primarily on the performance data of the students.

Kane (1994) classifies standard-setting procedures as either test-centred or examinee-centred. Test-centred procedures are those where the judges set the standards by considering the items in the examination and estimating how marginally competent students (and perhaps those at other points on an achievement scale) will perform on those items. Such procedures include the Angoff (1971), Nedelsky (1954) and Jaeger (1982) methods. Examinee-centred methods are those in which judges set standards by making pass/fail decisions about examinees. The cut-off score is set by choosing the point on the scale where there is inconsistency between the decisions of the judges. The Borderline-group method and the Contrasting-groups method (Livingston and Zieky, 1982) belong to this category.

Plake (1998) refers to Question by Question methods and Holistic approaches. With question by question methods it is common for judges to set a cut-off score on each particular item, and then aggregate these values to obtain the cut-off score for the examination. Holistic approaches "attempt to capture the totality of the candidates' performance by considering the overall examination performance" (p. 72).

In this study the focus will be on approaches which are Judgmental and Judgmental-Empirical according to the Meskauskas/Berk classification, and which are test-centred according to the Kane classification. The procedure applied in this study also fits the Plake classification for a question by question approach. Once the judges have established a cut-off score, however, they are encouraged to look holistically at the outcomes of setting that particular score to ensure that the performances of students awarded that score, match their expectations of the performances of students at that borderline.

Most standard-setting methods evolved from the need to establish a competency or "pass" score in examinations consisting of items that are dichotomously scored. Some of these methods, however, are more flexible than others and can be adapted, not only to handle multiple cut-off scores corresponding to a range of levels, but also to handle items which are polytomously scored. Many of the methods were devised in the 1970s and even earlier. During the 1980s and 1990s, work in this area has mainly focused on improving the traditional approaches and adapting them to suit new types of examination.

Many of the procedures that have been used successfully in recent times are based on a procedure proposed by Angoff in 1971. Adapting the Angoff procedure to handle different item types and building in additional steps has produced standard-setting procedures that are more flexible and yet give more reliable results. The next section outlines the Angoff procedure as originally proposed and considers alternative procedures that have been developed and issues that have arisen in the past. The

standard-setting procedure developed for use in this study is an adaptation of Angoff's procedure.

2.3.2 The Basic Angoff Procedure

Of the early standard-setting methods which use expert opinion to set cut-off scores corresponding to designated levels of performance on examinations, the Angoff (1971) procedure has been the most widely used. It has a number of advantages over other early methods (eg Nedelsky, 1954; Ebel, 1972) in that it is simple to apply, flexible, applicable to examinations containing items which are polytomously scored, and easily adapted to set multiple cut-off scores corresponding to different levels of performance. Understanding the original Angoff procedure, and how it was used, is important in understanding the criticisms of traditional procedures and the developments that have occurred in this area in the 1980s and 1990s.

Angoff's approach relates to the notion of the standard of performance associated with "a minimally acceptable student" or a borderline satisfactory performance on an examination. In its simplest form, it requires a team of judges working independently, with the image of a "minimally acceptable student" in mind, to decide whether this student would obtain the correct answer on each item. In the case of multiple-choice items, which are usually scored as one for a correct answer and zero for an incorrect answer, adding those items which a judge predicts such a student will answer correctly gives the cut-off score proposed by that judge. By averaging the cut-off scores proposed by each judge, the cut-off score to be used is obtained. A similar approach can be used to determine cut-off scores relating to other standards of performance, if required. In most applications of the Angoff procedure, an additional step is added.

Before the judges' decisions are averaged, it is common for the judges to be brought together to discuss their cut-off scores. This can lead to judges refining their values after listening to the views of their colleagues.

In a variation on this procedure for items which are dichotomously scored, the judges record the *probability* that a "minimally acceptable student" will answer the item correctly (Angoff, 1971). Summing the probabilities proposed by a judge across all items gives his/her cut-off score for the examination. Once again, discussions among the judges can be used to refine the initial decisions. Finally, averaging the recommendations from each judge gives the cut-off score for a minimally acceptable student.

Livingston and Zieky (1982) suggest a variation to this method by limiting the particular probabilities that judges can use. For instance, the judges could be asked to chose whether a "minimally acceptable student" would have a probability of 0.1, 0.2, 0.3, 0.4, 0.5, or 0.75 of obtaining the correct answer to an item. Livingston and Zieky do, however, state that this approach may be too limiting and may bias the judges' choices. For example, if the particular probabilities listed above are used, judges would not be able to express the view that students were almost certain to give the correct answer.

Whichever of the above approaches is used, the Angoff method can be applied to determine the cut-off scores relating to different standard levels before students have even sat an examination. It is entirely up to the administrator as to whether, at a later stage, the cut-off scores determined in this way are reviewed and refined in the light of

the performance of students on the examination. Used simply in the manner outlined above, the Angoff procedure is a judgmental one. When Angoff's method is supplemented by steps that build in a review of the cut-off scores in relation to data on student performance on the examination, the approach is test-centred and judgmental-empirical.

2.3.3 An Appraisal of Standard-setting Methods

Glass (1978), reviewing the then current approaches to standard-setting, expresses doubt about the feasibility of setting meaningful standards. He states that in standard-setting exercises, attempts at describing or specifying standards are often poor, the decisions made by the judges sometimes vary considerably, and the whole process is arbitrary. He believes that interpretations and decisions based on absolute levels of performance on examinations are largely meaningless. These absolute levels vary unaccountably with examination content and difficulty, and with the perceived consequences that ensue from the same absolute level of performance. In Glass's opinion, setting performance standards on examinations (by methods in use in 1978) is a waste of time or worse (p. 259).

Scriven, Hambleton, Block, Popham and Linn all disagree with Glass' views about the impossibility of setting accurate and meaningful standards. Scriven (1978) points out that, while some standard-setting approaches may be open to criticism, the problems raised by Glass can be overcome if standard-setting methods are multi-staged and incorporate sound procedures for calibrating and training judges and for synthesising sub-test scores. He believes that consideration should also be given to the purposes for which the assessment results are to be used.

Hambleton (1978) admits the need for further consideration of the issues and methods for determining cut-off scores and the need to develop clear guidelines for implementing the methods. While he expresses some concern with approaches which set a predetermined cut-off standard, he argues strongly that the notion of a cut-off score should not be abandoned simply because cut-off scores are often poorly set. He goes on to claim that, from observations he and others have made, where judges are properly trained and briefed for their task, the cut-off scores proposed by each judge are quite similar. Further, he states that this result holds across tasks in different subject areas and at different grade levels.

Block (1978) states that standard-setting methods are not as arbitrary as Glass claims. Whereas Glass sees the arbitrary nature of standard-setting in a negative way, Block sees the arbitrariness as healthy in that it allows for different professional experiences and opinions to be incorporated. He believes, however, that there is a need to keep researching standard-setting procedures with a view to strengthening them.

Popham (1978) challenges Glass' claim that it will never be possible to establish acceptable standard-setting models. He also attacks the claim of arbitrariness of standard-setting processes put forward by Glass, stating that Glass has misused the term and that "judgmental" would be a more appropriate term to use. Popham claims that judges can demonstrate high levels of inter-rater consistency:

> "If sophisticated standard setters comprehend the nature of their task, and have access to information relevant to that task, then there is no reason to assume that they too

cannot exercise non-arbitrary judgments." (p.298).

Popham also endorses the use of additional data on student performance to strengthen standard-setting procedures, including the use of normative data.

Finally, Linn (1978) argues that standard-setting is fundamentally a judgmental process and that all procedures for setting standards require human judgment at some stage. He also notes that the use of data on student performance is of great value in facilitating judgment. Linn supports the view that the standard-setting process should be an iterative one, permitting adjustments to the judgments based on the accumulation of information over time. He further claims that data on student comparative performance should be provided to the judges - not to be used to set standards, but to inform the judges in their deliberations.

Glass's attack on the standard-setting methods used at that time served an important function. Quite apart from the fact that it encouraged many others to propose ways in which the existing procedures could be applied more effectively, it increased interest in improving these procedures, and in developing new procedures.

One outcome of Glass' criticism was that a number of researchers conducted studies to determine which of the existing standard-setting methods was the best. Mills (1983), Brennan and Lockwood (1980), Harasym (1981), Cross, Impara, Frary and Jaeger (1984), and Smith and Smith (1988) conducted studies which compared the results of using the Angoff procedure with one or more of the other traditional standard-setting procedures. In each case they found that the Angoff procedure generally proved to be the best. It usually gave more satisfactory results than procedures based on the

Nedelsky (1954), Ebel (1972) and Jaeger (1982) procedures, and was more flexible and simpler to use than most other judgmental methods.

Another outcome was the modification of existing procedures, or the development of new ones, which overcame the problems identified by Glass. Researchers examined the standard-setting process and identified a number of important issues to consider in trying to improve the process. The following section examines a number of these key issues

2.3.4 Matters for Consideration in Setting Standards

2.3.4.1 The Use of a Structured Multi-stage or a Single-stage Approach

While early standard-setting procedures (eg Nedelsky, 1954) tended to involve a single process, later methods usually incorporate several stages. In this way, decisions made at one stage can be refined and improved at following stages.

Various procedures which have been proposed (eg Jaeger, 1982; Cross, Impara, Frary and Jaeger, 1984; Cizek, 1996; and Berk, 1996) all advocate the use of a structured, multi-stage approach.

Cizek (1993) expresses the view that standard-setting should be viewed as the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance. He sees standard-setting as a kind of psychometric "due process" that is akin to due process under the law (p. 100).

2.3.4.2 The Selection of Judges

A second issue involves the selection of the judges to be involved in a standard-setting exercise.

Jaeger (1982) used a mixture of registered voters, teachers, school counsellors and principals as judges to set cut-off scores on the North Carolina High School Competency Tests. High school students at the end of their studies take these tests, consisting of 120 multiple-choice items. Busch and Jaeger (1990), in a study designed to determine whether the background of the judges had any effect on the standards they recommended, used panels comprising school teachers and college/university lecturers. Although, during the application of the Angoff procedure, they found some systematic differences in the recommendations made by the judges which were related to their backgrounds, Busch and Jaeger nevertheless recommend that judges from different backgrounds continue to be used.

Jaeger (1991) expresses the view that standard-setting exercises should involve subject specialists, not policy makers. By this he means that decisions should be based on students' performance on the instrument, not simply on an edict that a fixed proportion of students will pass. Jaeger believes that care should be taken in selecting the judges, as a person who may be suitable for one task may not have the necessary understandings and expertise to perform another standard-setting role properly. His view is that, whenever possible, judges should be selected from among those who will have something to do with the students at the next stage, whether it be further education or training.

Norcini, Shea and Grosso (1991) show that where judges have been involved in the development of the examination instrument, as few as five judges and 25 common items can be sufficient to equate examinations by setting cut-off score equivalences. While their findings relate to examinations consisting of items that are dichotomously scored, there is no reason to assume that similar findings will not eventuate when examinations contain items which are polytomously scored. It may not be possible in many cases to have items common to two examinations, nevertheless, it is reasonable to propose that a small team of judges with a strong understanding of student performance standards in a courses of study can set accurate cut-off scores.

Plake, Impara and Potenza (1994) used an Angoff-based approach to determine item ratings on a general education test battery, selecting judges with particular content expertise from a range of content domains. They found that the cut-off scores established by using the recommendations from judges on out-of-content items differed little from the cut-off scores set by using the ratings made by the content specialists. They also found that when performance data were provided to the judges, those rating items outside their content speciality were not more inclined to change their opinions than those working within their content speciality.

For their study, Morrison, Busch and D'Arcy (1994) created three panels. One consisted of primary school teachers, another of secondary school teachers, and the other of grammar school teachers. They found good general agreement between the cut-off scores proposed by the panels when using an Angoff-based procedure to establish "Level 5" cut-off scores on the two tests of the Mathematics Common Assessment Instrument (CAI) of the Northern Ireland curriculum.

In addressing the issue of selection of judges, Berk (1996) states that a broad-based panel of the most qualified and credible judges should be selected.

Norcini and Shea (1997) believe that standard-setters must be recognised as leaders in their field and that it is not appropriate to ask non-experts to make judgments that require knowledge of content. They also claim that reproducible results can be obtained with as few as five to ten judges, but that a larger number will permit the inclusion of judges with different and important competencies. Whatever number of judges is used, Norcini and Shea believe it is necessary that a variety of perspectives are represented.

It can be seen from the above that the number and background of judges used in a standard-setting exercise depends upon the nature of the examination and the purpose of the exercise. In many cases, it may make the process more credible if the cut-off scores have been set by a relatively large team of judges drawn from a cross-section of the population. In other situations, however, it is essential that the judges have a very strong understanding of the subject matter being examined. In these circumstances, a relatively small team of highly qualified judges is more likely to set standards that will be accepted as appropriate by others. Such an approach is used in the setting of cut-off scores in curriculum-based examinations like the English GCE A Level examinations and the Scottish Higher Level examinations. In such cases, teachers with substantial experience in teaching the course and preparing students for the examination are most suitable. University and college lecturers, provided they have a thorough understanding of the range of standards of work produced by students in the course, would also be suitable.

2.3.4.3 The Training of Judges

Many researchers have identified the need to ensure that the judges involved in a standard-setting exercise are properly trained so that they fully understand the process they are to follow and what is required of them. In their response to the concerns expressed by Glass (1978), Scriven (1978) and Hambleton (1978) stress the need to properly train the judges for their task.

Reid (1991) argues that, not only must judges understand and be comfortable with the process to be followed, they also need to be sensitive to the influences of item difficulty on standard-setting. Judges must understand which features of an item may make it more difficult so that they can take account of this when determining how students will respond to it. He argues that the need for training is particularly important if the judges are not generally involved in assessing students. He suggests three criteria which can be built into processes for determining whether a judge is well-trained: standard-setting ratings should be stable over time, standard-setting ratings should be consistent with the relative difficulties of the items, and standard-setting ratings should reflect realistic expectations.

Mills, Melican and Ahluwalia (1991) also support the need to train judges. Their view is that judges must be aware of the process, their role, and how their advice will be used. They point out the importance of taking time to establish a common understanding, among the judges, of minimal competence as it applies to a particular body of knowledge and skills:

> "Without a common understanding of the process and a

> common definition of minimal competence, differences in item ratings may be more related to background variables of judges than to real differences in perceived item difficulty" (p. 7).

Thus, the research is quite explicit in indicating that judges involved in a standard-setting exercise must be thoroughly trained for their task and must have a clear understanding of what they are required to do. Preferably, this would be achieved by bringing the judges together, explaining the steps in the process, and having judges determine cut-off scores on some sample items. The judges should be given the opportunity to ask questions and discuss the process. A set of written instructions should also be provided for judges to use when working individually.

2.3.4.4 The Initial Steps in Establishing Cut-off Scores

In an attempt to find a sound method for setting an initial cut-off score in a multi-stage process, Jaeger (1982) asked a large, diverse group of judges to make a Yes/No decision for each item on an examination as to whether every high school graduate should be able to answer the item correctly. Jaeger reports that there were considerable variations between the opinions of the judges on what the cut-off score should be, even after feedback was provided. These variations may have been because the judges were from such diverse backgrounds. Perhaps a smaller number of judges, who had a more intimate understanding of the knowledge and skills typically possessed by students at the end of secondary education, and with increased opportunity for discussion of their decisions, may have brought greater consistency.

Kane (1986) suggests a modification to the practice of deriving an overall pass score for an examination by simply summing the cut-off scores for the individual items. He suggests choosing a passing score that would cause the probability that students with scores at the passing score answer a given item correctly to be as close as possible to the minimum passing level (MPL) for that item. The MPL for an item is the probability that a minimally competent student could answer the item correctly. The advantage of this method is that it can either be used to set standards or check the validity of cut-off scores set by traditional means.

Kane defined $d_i(x)$ such that

$$d_i(x) = P_i(x) - MPL_i$$

Equation 2.3

where $d_i(x)$ is the difference between the actual passing level for item i, that is, the probability that examinees with observed scores at the passing score answer item i correctly and, the minimum passing level (MPL) generated by the judgmental standard-setting procedure.

The passing score is then set to be the observed score that minimises the sum of the squared discrepancies for all items. This sum is given by

$$D(x) = \sum_i d_i^2(x)$$

Equation 2.4

if the assumption is made that all items have equal weight.

Norcini, Shea and Ping (1988) demonstrate that it is not always necessary for all judges to consider every item when determining the cut-off score in an examination. They

created five teams of judges and allocated the items from an examination to them at random, with no duplication. The judges set cut-off scores for their sub-tests utilising careful briefings, group discussions and consideration of student performance data. Norcini *et al* were able to demonstrate that there was very little difference between the cut-off scores predicted for the total examination by each team based on the cut-off scores calculated on their sub-tests. Their study demonstrated a high level of internal consistency between judges.

Fehrmann, Woehr and Arthur (1991) hypothesise that the reliability of the judgments made during the use of the Angoff procedure can be improved by incorporating "frame of reference" training for the judges prior to their use of the procedure. Their study consisted of setting up a No Frame of Reference group (NFR), a Standard Frame-of-Reference group (SFR) and a Consensus Frame of Reference group (CFR). In all cases, an explanation of the Angoff procedure was provided.

Fehrmann *et al* found that when judges are given a detailed description of the knowledge and skills typically possessed by a marginally competent student (SFR), or where through a process of discussion judges are able to develop their own detailed description (CFR), they are able to set accurate, reliable and consistent cut-off scores. On the other hand, when judges are simply given a general description of a marginally competent student (NFR), they have difficulty in reaching agreement on the cut-off scores.

The results of this study clearly show that providing judges with a frame of reference (in the form of exemplary materials and feedback and the opportunity to discuss student

performance data) leads to higher levels of inter-judge reliability, consistency and accuracy in setting standards.

Faggen, Melican and Powers (1995) demonstrated that the form of presentation of items (ie using a computer screen or paper) was irrelevant to the results achieved during a standard-setting exercise.

Cross, Frary, Kelly, Small and Impara (1985), in a procedure designed for setting standards on essay items, decided that this should initially be done with judges being unaware of the scores awarded to the essays they were given. This information was later provided when they had made their initial decisions, and they were given the opportunity to discuss and change these initial decisions.

The studies reported above show different ways in which attempts have been made to provide support and guidance to judges in the initial stages of a standard-setting procedure. If judges can be assisted to develop an accurate understanding of the standards they are to apply, the initial decisions they make will be relatively accurate, and merely require review and refinement at later stages in the procedure.

2.3.4.5 Discussion and Refinement of the Initial Cut-off Scores

Early standard-setting procedures (eg Nedelsky, 1954; Angoff, 1971) simply involved collecting the decisions of the individual judges and then averaging them to determine the cut-off score. The judges were not given the opportunity to refine their initial opinions as a result of discussion with their fellow judges.

In many of the standard-setting procedures used today, judges are provided with the decisions made by other judges, these decisions are discussed, and the judges are then given the opportunity to vary their own decisions if they wish. After these discussions, there may still be differences between judges. It is usual then for an average of their decisions to be recorded as the cut-off score. In some cases, rather than calculate the average, judges continue to discuss their decisions until consensus is reached.

Norcini, Lipner, Langdon and Strecker (1987) selected a group of judges and had them apply the Angoff technique under three different conditions. Initially, the judges were given an explanation of the Angoff procedure and the opportunity in a group meeting to apply it to a short test consisting of items similar to those in the examination that was to be used in the study. For the first condition, the judges were sent instructions, a booklet containing half the items of an examination and normative data. They were asked to determine, and record for each item, the proportion of borderline competent/not competent students who would answer the item correctly. For the second condition, the judges were brought together and given the opportunity to discuss their earlier decisions with their colleagues. They were able to change their decisions where they felt it was warranted. For the third condition, a month after the group meeting the judges were mailed a booklet containing the second half of the items from the examination and asked to make judgments without consulting each other.

Norcini *et al* (1987) found that, while there was a considerable range in the cut-off scores set when the judges worked on their own (the first condition), there was close agreement following the discussions (the second condition). They also found that there was close agreement between the judges on the second set of items (the third condition),

even though the judges did not have the opportunity to discuss their decisions. This led them to conclude that the group discussion process is an important step in establishing standards, and that once the judges have established standards at the group meeting, these standards tend to stay with them.

A number of other researchers report that giving the judges the opportunity to discuss their decisions, and to refine these decisions on the basis of the discussion, is a very important step in attaining consistency and accuracy in setting standards. Among those who support this approach are Jaeger (1982), Morrison, Busch and D'Arcy (1994) and Berk (1996).

2.3.4.6 Feedback to Judges and Refinement of the Initial Cut-off Scores

In addition to information on the decisions of the other judges and the opportunity to discuss those decisions, giving judges statistical feedback on the performance of students in the examination is seen as a means of improving the quality of the decisions they make. The type of information provided varies. In some studies, item analysis data are provided. In other cases, the data consist of frequency distributions of the scores gained by the students. Samples of student scripts is another form of feedback which can be provided.

Geisinger (1991) identifies various forms and sources of data that can be used in the standard-setting process. He identifies three forms of primary data which are commonly used: the cut-off scores themselves, determined by a process of expert judgment; acceptable passing and failing rates; and the relative costs of miscalculation errors. He also suggests a number of different supplementary types of information that

can be useful, including errors of measurement, errors of rating, anomalies in the rating process, and the results from different standard-setting sessions or techniques. Geisinger holds that one can expect the process of setting standards to be more reliable and accurate when a variety of data are used in combination. This is particularly the case should it be necessary in a high-stakes situation to use a compromise model, such as those proposed by Beuk (1984), De Gruijter (1985) and Hofstee (1983). A compromise approach is often used in situations where it is necessary to reach a balance so that absolute standards are maintained and yet a politically acceptable proportion of students is deemed to have reached the required standard.

Reid (1991) cautions that the use of normative data as a form of feedback needs to be handled with care. He claims that, while it has an important role to play and can be particularly helpful, care must be taken to ensure that judges do not simply change their initial determinations to fall in line with such data. Discrepancies between a judge's decisions and the performance data may be caused either by inaccurate expectations on the part of the judge or by variations in the performances of the students. In most cases, it is not possible to determine which factor has caused the discrepancy. Indeed, both may have contributed. Reid believes that judges need to be aware of the limitations of normative performance data so that cut-off scores are not simply set to match the status quo.

Popham (1978), Linn (1978), Jaeger (1982), Cross *et al* (1984), and Norcini, Shea and Kanya (1988) all support the use of student performance data to assist judges in refining their initial decisions. Their research suggests that providing the judges, with either statistical data on student performance or with samples of student examination scripts

improves the quality of the decisions made. Norcini and Shea (1997) indicate that the credibility of the standard can be enhanced by including data from external sources in the process. They claim that performance data provide judges with an anchor in reality, but that empirical data should only be used "through the filter of their judgment" (p. 44).

Wiliam (1996) indicates that a danger with test-centred standard-setting procedures is that they can generate standards which appear quite reasonable, but which can be difficult for students to achieve. Judges, asked to set cut-off scores with little or no guidance, may set cut-off scores that are too high. It is for this reason that either explicit use is made of normative data in the original standard-setting process, or empirical data are used to assist judges in the finalisation of the cut-off scores.

By examining a sample of student scripts which have been awarded scores at or around their proposed cut-off score, the judges can note whether students who gain the actual cut-off score demonstrate skills and knowledge commensurate with their image of that standard. This improves the validity of the decisions. The research evidence is clear, however, that statistical data on student performance and student scripts should be used to help judges review and refine decisions they have made. They should be used to inform professional judgment, not replace it.

2.3.4.7 Articulating the Standards

The value of describing standards of student performance in terms of the knowledge and skills typically displayed by students who reach each standard is recognised by a

number of researchers. Such descriptions are particularly helpful in the standard-setting process, as well as in reporting student achievement to various audiences.

A clear and comprehensive description of standards enables judges to understand and internalise the standards to be applied when setting the cut-off scores. As Fehrmann *et al* (1991) showed, once they have developed a good understanding of the standards, judges are able to apply them with considerable consistency in setting cut-off scores for examinations.

Kane (1986) shows that it is possible to develop a performance-based interpretation of passing scores. His approach, for tests consisting of items that are dichotomously scored, is to identify those items which passing students are more likely to answer correctly than failing students. By considering the course content covered by such items, it is possible to make interpretations about the nature of the achievement of a passing student.

Mills, Melican and Ahluwalia (1991) indicate that, in cases where the assessment is being conducted for the purpose of certification, it should be possible to bring together judges with a thorough understanding of the domain. Mills *et al* note that, along with this understanding, the judges will bring with them different perceptions of student achievement and minimal competence. These differences are due to such factors as their familiarity with the curriculum, the range of abilities and achievements of the students with whom they have been involved, and their own experience in assessing students. In spite of these differences the judges can determine and describe, through a process of negotiation, those skills and knowledge required for minimal competence.

If a process is put in place where the judges work to build up an agreed description of the knowledge and skills typically displayed by students who reach a particular standard, this description should improve the quality of the decisions made by the judges. Once such knowledge and skills are clearly articulated, judges can use these descriptions, and other support materials such as student responses, to set cut-off scores on other forms of the examination.

2.3.5 Recent Developments and Current Attitudes to Setting Standards

In recent times there has been a resurgence of interest in standard-setting procedures. This has come as a result of a need to find ways of setting standards in complex, multi-dimensional performance assessments. While these new methods often need to cope with collating judgments taken across a number of quite diverse tasks, they incorporate many of the features and steps inherent in the earlier procedures. In this, and the next, section - which examines the use of Latent Trait Theory in the setting of standards - new methods and developments in judgmental standard-setting are discussed.

Webb and Miller (1995) propose two methods for setting standards on examinations involving items that are polytomously scored. The Paper Selection Method requires judges to first conceptualise students who are at the borderlines of the various categories of performance. They then read a large number of student scripts and choose three, one at each of the borderlines (Basic, Proficient and Advanced). Judges are not told what scores, on a four-point scale, have been awarded to the scripts. The procedure involves performing this process several times, with judges being given feedback in the form of intra-judge and inter-judge consistency at the end of each cycle, and the opportunity to

discuss their decisions. Scores for the borderline scripts selected after the final repetition are then combined with cut-off scores determined on other sections of the examination to produce the final cut-off scores.

The Contrasting Groups Method, applied by Webb and Miller (1995) to the 1993 New Jersey Early Warning Test, requires judges to sort the scripts of students responding to constructed-response items into three categories (viz. does not need instructional intervention, may or may not need instructional intervention, does need instructional intervention). This process is repeated across three rounds, with feedback on intra-judge and inter-judge consistency provided as for the Paper Selection Method. Cut-off scores are computed by using a formula which averages the scores awarded to the student scripts allocated to the first two categories and the average score of those scripts which are allocated to the third, or "needs instructional intervention", category.

Both methods use large panels of judges, with the Paper Selection Method, in particular, requiring a considerable amount of judges' time. Webb and Miller (1995) are of the view that both methods tend to overestimate the performances of students. However, their results fit criteria for judging the effectiveness of standard-setting procedures proposed by Plake (1995). These include the accuracy of the decisions resulting from the application of the standard, the ease of administration, the judges' comfort with the final decision rule, the judges' confidence in the results, and the potential replicability of the decision rule resulting from the standard-setting procedure (p. 89).

Poggio and Glasnapp (1994) propose a new judgmental method for setting standards which they feel overcomes some of the shortcomings of the Angoff (1971) and Ebel

(1972) methods, including suitability for use both with items which are dichotomously scored and which are polytomously scored. After the items in the examination are analysed for cognitive demand and importance, each judge specifies a distribution of examination scores which he/she believes to be the minimum acceptable score distribution for students assigned to each performance level. The mean (or median) of the distribution of scores proposed by a judge for each standard level gives the nominated cut-off scores for that judge. The cut-off scores proposed by the individual judges can then be averaged to produce the final cut-off scores. One disadvantage of this approach, however, is that a relatively large number of judges is needed.

In seeking a standard-setting procedure which is suitable for use with the type of complex performance assessments used in certification programs, comparisons were made of the results of applying the Judgmental Policy Capturing process (Jaeger, 1995), an extended Angoff procedure (Hambleton and Plake, 1995) and the Multi-stage Dominant Profile method (Putham, Pence and Jaeger, 1995). The methods were applied to the National Board for Professional Teaching Standards (NBPTS) assessment program for teacher competence.

The Two-Stage Judgmental Policy Capturing process (Jaeger, 1995) involves judges responding to a large number of profiles relating to student performance in a complex task. If student competence (and/or merit and/or excellence) is to be determined on the basis of performance across a number of tasks where each task contains several different skills or content areas, it is possible to identify a number of profiles corresponding to the different strengths and weaknesses of students. In the first stage of the Judgmental Policy Capturing method, the judges classify each profile on a five-

point scale. Their job is to assign a score of 1 (poor), 2 (mediocre), 3 (satisfactory), 4 (noteworthy) or 5 (excellent) to each of the profiles which theoretically evolve from each task. In the second stage the judges classify a further large set of profiles consisting of the performance of hypothetical students across all the tasks. The judges classify them as belonging to a student who is a Novice, or Competent, or Accomplished, or Highly Accomplished. After each stage, mathematical models are used to enable a single decision to be made about a student's performance given his/her profile of performance across all tasks.

The method is reasonably straightforward but is quite lengthy and involves some relatively complex statistical procedures. Berk (1995) points out that it was developed from a non-educational measurement foundation and so has limitations as a result. In particular, the lack of feedback to the judges - and of the opportunity for them to discuss and vary their original decisions - probably results in far more variability between the decisions than is necessary.

Hambleton and Plake (1995) propose a model based on Angoff's procedure for setting standards on complex tasks. Under this model, the judges independently estimate the expected score that a marginally competent student will receive on each discrete skill or content area within a task. In addition, they suggest what weighting they think each skill or content area should have in the task. Once they have done this for every task, they are provided with feedback on the estimates proposed by all judges and have the opportunity to change their values. The average of the judges' values is calculated and the cut-off score for a marginally competent student on each task determined. The judges are then given the job of determining what weighting each task should have

when all tasks are combined. Again, they are given feedback showing the values proposed by all judges and can vary their recommendations. The various weightings proposed are then averaged and applied to the cut-off scores determined for each task. This gives the cut-off score for a marginally competent student across the total assessment.

This model has come in for some criticism for employing what is termed a "compensatory approach". Under such a model, students can be weak at one or more components, or perform poorly on one or more tasks, but can still obtain a score which sees them classified as competent. Additional conditions can be set in order to make the model more like a "conjunctive" approach - that is, one which requires students to obtain at least a certain score on particular key components or tasks. For example, students may be examined for a driver's licence in a manner that involves both a written and a practical component. Under a conjunctive model, they would need to demonstrate a satisfactory level of knowledge and skills in both the written and practical components. Using a conjunctive approach would alleviate some of the criticisms related to the use of the Hambleton and Plake procedure in such applications as licensing or certifying competence for professional standing. In other cases, however, particularly where an examination covers a wide domain, it is usually considered that a compensatory approach is more suitable.

Berk (1995) identifies another issue with this procedure. In assigning weights to the tasks or components, the judges tend to propose weightings that are all very nearly equal. This situation tends to become even more evident after the averages of the different judges' values are calculated. The model is, nevertheless, simple to use and

flexible enough to be adjusted to take account of these concerns. Plake (1995) notes that the judges were comfortable with this method due to the use of group discussion and refinement that is typical of modern Angoff-type approaches. This method, however, does not readily support the match of the performance standards established to the underlying decision rule policies of the judges.

Putham, Pence and Jaeger (1995) suggest that their Multi-Stage Dominant Profile method overcomes some of the shortcomings of the Judgmental Policy Capturing and the Extended Angoff procedures. The initial step in the Multi-Stage Dominant Profile method requires judges to create decision rules which they can apply to assess competence across a series of multi-dimensional assessment tasks. These rules generally consist of statements about the level of performance they would expect of a student in the various components of the tasks (from 1 to 4 as in the Judgmental Policy Capturing method). The rules proposed by each judge are then analysed by a coordinator and synthesised to provide a set of profiles. The judges consider these profiles independently, in order to decide which of the profiles they feel would enable pass/fail decisions to be made. The results of the judges' decisions are analysed, using both compensatory and conjunctive models to ascertain which model provides the best fit to the decision rules used by the judges.

As a way of providing more structure and order to the Judgmental Policy Capturing method, the Multi-Stage Dominant Profile method has some potential. Berk (1995) indicates that it enables judges explicitly to create and state their decision policies holistically across the entire assessment package, rather than make decisions exercise by exercise and then amalgamate them into a consolidated policy. This could help to

overcome one of the concerns expressed about the use of the Extended Angoff method. In order for this method to work effectively, however, considerable care is needed in training the judges and preparing them for their task. This improves the likelihood of devising clearer and more consistent decision rules so that concerns about replicability can be minimised. In applying this method, it is advisable to allow the judges some discretion in the application of the rules. Cases may arise where a student's profile does not satisfy the criteria for a "pass" result according to the strict application of the rule, but on examination of the data it could be argued that the student is competent.

Plake, Hambleton and Jaeger (1997) investigated the Dominant Profile Judgment method in the certification of teachers using the Early Adolescence Generalist assessment. They identify the use of a conjunctive approach as a cause for concern in relation to measurement reliability. It is possible that under a conjunctive approach the decision to fail a student on the whole assessment may rest on the student's performance on a single item or task. This may be appropriate in cases where the assessment is conducted for the purpose of certifying competence in performing a process where issues of safety or security are paramount. It is doubtful, however, whether such an approach is appropriate in an assessment in a general education course. Plake *et al* also identified possible problems in gaining agreement amongst the judges on the final standard-setting policy. They acknowledge that the goal is to obtain consensus, but indicate it is unclear what action should be taken when the judges disagree.

Impara and Plake (1996) suggest that estimating item difficulty is a difficult task for judges and that if they cannot perform this task, the validity of the standards based on

their estimates is in question. This being the case, it further emphasises the need for the judges to be experienced, to be adequately trained for their task, to be provided with feedback on their decisions and to be given the opportunity to discuss those decisions.

Berk (1996) notes that two major changes in testing practices necessitate the development of new standard-setting procedures. These changes are the increased use of items that are scored polytomously, and the establishment of multiple standards, rather than a simple "pass/fail" arrangement. Berk reports that various groups have experimented with the use of descriptor statements where certain points on a scale, referred to as "anchor points", are described in terms of the knowledge, skills and abilities exhibited by students at or near those points.

Berk proposes his Generic Eclectic Method as a means of setting standards on complex tasks. He emphasises, however, that before commencing the process it is essential that a broad-based panel of the most qualified and credible judges be selected, and that they be carefully and thoroughly trained so as to minimise the effects due to the instrument and to maximise intra-judge consistency.

Where an examination-centred standard-setting approach is being used, the Generic Eclectic Method involves judges meeting to define achievement levels and preparing explicit behavioural descriptions, based on consensus. A sample of items (anchor items), similar to those in the examination, is presented to the judges, who collectively select items which are at the upper and lower ends of each of the achievement categories. Using the behavioural descriptions and the anchor items, the judges independently match the examination items to the achievement level categories. Judges

are given feedback on the decisions made by their colleagues, as well as meaningful performance data, and are given the opportunity to revise their initial decisions. Discussion is held amongst the judges concerning their revised decisions, without any requirement to reach consensus. Following this stage the judges submit their classifications, with the cut-off scores for each achievement level being the mean or median of the judges' individual scores.

Mills and Melican (1988) claim that

> "Achievement is generally considered to be a continuous variable. Therefore, dichotomising that variable into two mutually exclusive categories, such as mastery status and non-mastery status, is difficult and some classification errors are inevitable. As a result, no standard-setting method will be ideal. Nonetheless, decisions do have to be made about individuals, and test scores can be an important piece of information in the decision-making process" (p. 273).

Zieky (1996) describes the history of standard-setting by referring to the Ages of Innocence, Awakening, Disillusionment, and Realistic Acceptance. He claims that the move from the Age of Innocence, where little thought was given to the process or the implications of the consequences, to the Age of Awakening coincided with the introduction of the criterion-referenced testing movement. It became imperative that standards be set according to formal procedures. The Angoff procedure (Angoff, 1971) was one of the procedures developed in this period. The Age of Disillusionment arose when the critics of standard-setting, notably Glass (1978), claimed that all procedures were subjective and gave different results from each other. From this period, Zieky believes that, after much work in the area, we have entered the Age of Realistic

Acceptance. It is generally accepted that there is no objective way of setting standards, and that it is quite right and proper to set standards by using judgments. Zieky does indicate, however, that "a standard is based on the values of some group of people and as long as different people hold different values, standard-setting will remain controversial" (Personal Correspondence).

Wiliam (1996) supports this notion and contends that "no standard-setting method is ideal, but one may support the most important inferences which may be drawn from the results better than others". He holds that "the meaning of standards can only be defined relative to a community of interpreters. Those who do not share the assumptions of the community will not agree about the meaning of the standards" (p. 303).

2.3.6 The Use of Latent Trait Theory in the Standard-setting Process

Latent Trait Theory has been used in judgmental standard-setting procedures in a number of different ways. Early work in this area by Van der Linden (1982) focused on the use of Latent Trait Theory to validate intra-judge reliability. In recent times, Latent Trait Theory has been used not only as a means of monitoring reliability, but also as a key step in establishing the cut-off scores.

While latent trait measurement models have their detractors (eg Goldstein, 1979; Goldstein, 1980; McLean and Ragsdale, 1983; Divgi, 1986), it is evident from the many studies undertaken that these models can be used most effectively in measuring and reporting student achievement.

Kane (1987) uses Latent Trait Theory to analyse the results of a judgmental standard-setting exercise. He evaluates the decisions of judges to determine whether they fit a latent trait model. He argues that even if examinee performance data have been shown to fit the model, this provides no assurance that the ratings fit the same (or any different) model (p. 334).

His paper questions the assumption that combining the average of the judges' decisions on each item to obtain a passing score for the test provides the best estimate of the passing score. Kane believes that if the ratings do not fit the model, the use of the model as a basis for combining the minimum pass levels over judges and items to obtain a passing score for the test, or for estimating the expected error in passing scores, should be considered suspect (p. 336).

The issues raised by Kane in relation to the fit to latent trait models, both in terms of examinee response and judges' decisions, need to be considered. The relevance of these issues, however, will depend on the role played by, and emphasis placed on, the latent trait model used in any particular standard-setting exercise.

Plake and Kane (1991) further investigated the methods proposed by Kane (1987) for establishing a passing score on a test based on the item-by-item minimum passing level approach. Using simulated data from hypothetical judges, they applied a three-parameter latent trait model to generate data for minimum passing levels on a test. One method used the common approach, adopted in most standard-setting situations, of simply summing the estimates of the minimum passing levels. The other two methods adopted procedures for differentially weighting the minimum passing levels of the

items. Very little difference was found between the results of the three methods. They therefore recommended that the simple approach of applying equal weightings be used.

McKinley, Newman and Wiser (1996) used teams of judges to establish cut-off scores for satisfactory performance using the Angoff (1971) procedure in examinations consisting of 50 four-option multiple-choice items. They also used a second procedure, based on a Rasch model, which they term "item mapping". The Rasch-based process leads to the calibration of items. Different forms of the examination are then placed on the same metric, thus creating an item bank. Items that do not fit the Rasch model are excluded from use in the item-mapping procedure. The items are next placed on a scale based on their difficulty, with items of similar difficulty grouped together. Judges are asked to indicate which group contains items which borderline students have a 50:50 chance of answering correctly.

Once the judges agree on an item difficulty at which the borderline examinee has a 50% chance of a correct response, probabilities for harder and easier items are re-examined with respect to that point. The cut-off scores for the various forms of the examination can then be determined.

When comparing the results from the Angoff procedure and a Rasch model, McKinley *et al* found some minor differences in the cut-off scores for certain forms of the examination. They concluded that the Rasch model has much to offer in that it facilitates the presentation of information about both item and student performance. In providing this information, the procedure addresses the issue of judges' precision in predicting borderline examinee performance, reducing the likelihood that additional

adjustments have to be made in order to determine the passing score. Further, they conclude that the procedure is relatively time-efficient, as only a representative sample of items needs to be rated by the judges, and the item-banking procedure allows more time to be spent discussing those items where consensus is not easily reached. The applicability of such an approach will be limited, however, when the examinations consist predominantly of items that are scored polytomously and it is not possible to have common items in two examinations.

Kahl, Crockett, DePascale and Rindfleisch (1994) report on the use of the Student-Based Constructed Response method and the Item-based Constructed Response method in the Maine Educational Assessment (MEA). Both methods are designed to enable standards to be set on tests containing items that are scored polytomously. Descriptions of student performance corresponding to Distinguished, Proficient, Apprentice and Novice have been developed to be used both in the establishment of the cut-off scores and the reporting of student achievement. In the Student-based Constructed Response method, students are placed on a Rasch ability scale based on their scores on common items. Judges then review the complete set of responses for a sample of students. The Item-based Constructed Response method places the score points for the individual items on the Rasch ability scale. The judges then review the student responses sorted by awarded score by item.

Kahl *et al* (1994) note that the judges involved in the standard-setting exercises report that they are generally able to relate student responses to definitions of proficiency levels. They are, however, more comfortable with relating complete sets of the responses of the students to the performance descriptors associated with the Student-

based Constructed Response method than they are with making judgments relating to individual items/score points using the Item-based Constructed Response method.

Latent Trait Theory has been proposed as a way of analysing and reporting student performance in curriculum-based examinations containing items that are scored polytomously. In McGaw (1997), Stephanou developed an achievement scale using data from the 1995 Higher School Certificate (HSC) Physics examination given to New South Wales students. This analysis was performed using the Partial Credit Model (Masters, 1982; Wright and Masters, 1982), which enabled the difficulty level associated with obtaining each possible score on each item to be placed on the same scale. Stephanou then created a number of bands of achievement corresponding to adjacent score ranges. By analysing the knowledge and skills that Physics students needed in order to obtain the scores on each item which fell within the band, he prepared statements which described what students in each band generally know and can do. This approach has similarities to the scales created as part of the study reported in this thesis using the Extended Logistic Model (Andrich, 1978; Tognolini and Andrich, 1995). The way the scales are used, the establishment of the standards of achievement, and the construction of the descriptor statements, however, are different.

Engelhard and Cramer (1997) used the Binomial Trials Model (BTM), a Rasch measurement model, as a means of analysing data relating to judges' decisions in a standard-setting exercise. This technique compares the observed and expected values of each judge's decisions as a means of determining the validity of the process, through the identification of inconsistent judges and those whose severity is significantly different from that of their colleagues. They found the correlations between judges' predictions

of the difficulty of dichotomously scored items and the empirical item difficulties presented by the BTM to be high.

Engelhard and Anderson (1996) use the BTM in evaluating the quality of judgments obtained from judges. A modified Angoff procedure, with three rounds of judgments, is used as the standard-setting procedure, with feedback being provided to judges between rounds. They conclude that the BTM provides useful information regarding the judgments which is not available from other procedures.

Engelhard and Gordon (1997) used experienced judges to make pass/fail decisions on the Georgia High School Writing Test (GHSWT). This test requires students to write an essay of two pages in length in a 90-minute period. These essays are analytically scored on four domains of effective writing (Content and Organisation, Style, Conventions, and Sentence Formation), using a rating scale with four categories.

They created batches of student scripts with the same scores and asked judges to determine whether the scripts in each batch were worthy of a "pass". They then used a Rasch measurement model to map the ratings from judges to a judgmental scale of writing competence to identify judges whose views differed significantly from those of the others.

Using a three-stage process, with judges presented with feedback and the opportunity to discuss their ratings at the end of each stage, Engelhard and Gordon found that the agreement between the judges regarding the quality of writing represented in the batches of student scripts increased over the rounds. They also found that this

procedure provided an effective technique for establishing the cut-off score for this type of test.

In addition to the studies outlined above, Hambleton and Cook (1977), Lord (1980), Smith (1986) and Becker and Forsyth (1992) advocate the application of latent trait models to the investigation and analysis of student performance data.

2.3.7 Common Directions in Judgmental Standard-setting Procedures

Developments in assessment practices, such as the greater use of items that are scored polytomously to examine complex tasks, require changes to traditional standard-setting procedures. It can be seen from the studies reported above that most contemporary procedures improve upon earlier procedures by ensuring that the judges are carefully briefed and trained for their task and by providing opportunities for judges to discuss and change their initial decisions. They also provide feedback to judges in the form of statistical data on student performance and they examine samples of student responses to the items. In particular, a number of recent studies have been conducted where Latent Trait Theory has been used, either in setting the cut-off scores or in providing limited statistical feedback.

2.4 ISSUES RELATING TO RELIABILITY AND VALIDITY

While a wide variety of standard-setting methods have been developed and applied over the last 40 years, questions of reliability and validity still remain. Indeed, the motivation for developing many of the methods was to improve the reliability and validity of existing procedures, particularly following the criticism by Glass (1978). Studies have also attempted to develop measures to ascertain whether particular

standard-setting procedures are giving reliable and valid results (eg DeMauro and Powers, 1993; Livingstone and Lewis, 1995).

2.4.1 Reliability

Jaeger (1990) claims that the

> "reliability of a standard-setting procedure is the extent to which it produces consistent classifications of an examinee as 'competent' or 'incompetent' when it is applied to different samples of items from the domain of generalisation, by different samples of judges, on different occasions of judgment" (p. 16).

It is generally accepted that there are two aspects of reliability which should be considered: intra-judge reliability and inter-judge reliability.

2.4.1.1 Intra-judge Reliability

Intra-judge reliability is the extent to which an individual judge's decisions are consistent with each other, and so, represent the application of the same standards.

Van der Linden (1982) investigated the issue of intra-judge reliability in applications of the Angoff and Nedelsky procedures and proposed a method for checking for intra-judge consistency by using Latent Trait Theory. His view is that this approach can also be used to select judges, evaluate training programs for judges, or assess the consequences of modifying standard-setting techniques. It is also suitable for use in an interactive way during standard-setting exercises. In such a situation, judges are given feedback on their estimates during the standard-setting stages, allowing them to vary their decisions where they are found to be inconsistent.

Van der Linden further states that if his technique is to be used it is important that the items in an examination fit the Latent Trait Model in use. During the examination-setting stages, items can be trialled and then refined if found not to fit the model. If, on the other hand, it is decided to use Latent Trait Theory in a situation where an examination is already in existence, some items may not fit the model. Van der Linden indicates that in such cases if his technique is applied to set standards, only those items that do fit the model should be used.

Plake, Melican and Mills (1991) identify intra-judge inconsistency as a potential problem in all standard-setting exercises. They indicate that not only is it important to undertake the careful training of judges prior to the exercise, it is essential to monitor intra-judge consistency at various stages throughout the process. They identify a number of factors which can affect the reliability of a judge across the stages, including his/her past experiences, special expertise, perception of the knowledge and skills required of students in order to be classed as minimally competent, and fatigue or lack of concentration during the process. It is also possible that factors to do with the examination itself may affect intra-judge consistency. Such things as the difficulty of the examination, and the view of a judge as to whether it is a suitable instrument for measuring the competence of students in the particular program, may affect consistency. Certain factors to do with the standard-setting process, such as whether an answer key is provided, may also affect intra-judge consistency.

Plake *et al* (1991) suggest a number of strategies which can improve intra-judge consistency. For example, stopping the standard-setting process to provide for retraining of the judges can assist in maintaining in the judges' minds the notion of a

minimally competent student. Providing empirical data can also be of considerable benefit. Such data might relate to the whole candidature, or might be based on the performances of a sub-group of students. Another possible approach is to use ratings based on Latent Trait Theory to compare against judges' item performance ratings and signal cases where there are discrepancies. A further approach is to provide descriptive data on judges' ratings to the whole team. By using an iterative process, each judge is provided with the decisions made by all judges and is given the opportunity to review their own choices. This process can then be undertaken again after adjustments are made. This approach could be used purely for information purposes, or some form of consensus might be expected.

Plake and Impara (1996) show that high levels of intra-judge reliability can be obtained during a structured standard-setting procedure. They conclude that a strong emphasis on training the judges and discussion focused on the skills and characteristics of minimally competent students are very important.

Berk (1996) distinguishes between intra-judge reliability between steps and intra-judge reliability within steps. The former, he argues, will often tend to be quite low in situations where an iterative process is being used and judges make adjustments to their original decisions.

Lack of intra-judge reliability within steps, Berk claims, is a much more serious problem. He identifies a number of strategies that should be used in an attempt to maximise intra-judge reliability. Among these are having judges prepare explicit behavioural descriptions of the achievement levels, training judges to match items or

student work to different levels of performance, providing judges with feedback on their decisions and student performance data, and enabling the judges to discuss and refine their decisions.

In an exercise where judges were asked to identify which test items nurses must answer correctly on items used in a certification test, Engelhard and Stone (1997) applied a Rasch measurement model to examine five categories of rating errors: severity or leniency, halo, central tendency, restriction of range, and inter-rater reliability. They conclude that the application of the Rasch model in such an investigation can provide those with the responsibility of establishing cut-off scores with invaluable information. Such information includes feedback on the items that are included in the test and the views of each judge on what knowledge students need in order to be classed as "minimally competent".

2.4.1.2 Inter-judge Reliability

Inter-judge reliability is the extent to which the decisions made by the judges in the panel are homogeneous or internally consistent.

Jaeger (1988) notes that judgmental standard-setting procedures implicitly presume that judges are equally qualified and precise in providing their recommendations. Judges can differ widely, however, in the standards they recommend. One reason for this is that the judges may not be equally informed and equally confident of their abilities to complete standard-setting tasks. On the other hand, the judges may be equally qualified but simply hold different views of the required standards.

In Jaeger's view,

> "achieving consensus on an appropriate standard for a test is an admirable goal (certainly guaranteed through the use of a single judge), but it should not be pursued at the expense of fairly representing the population of judges whose recommendations are pertinent to the task of establishing a workable and equitable test standard. To eliminate the recommendations of some judges only because they differ from those of the majority is antithetical to the more fundamental goal of seeking the informed and reasoned judgments of one or more samples of judges who represent the population or populations of persons who have a legitimate stake in the outcome of the testing program under study" (p. 29).

Jaeger argues that, rather than eliminate the recommendations of judges whose response patterns are inconsistent with the group as a whole, a better approach is to eliminate them if the pattern of decisions they make are markedly different from the pattern of responses typical of marginally competent students. In such cases it could be claimed that their recommendations were incorrect, rather than simply different from their colleagues'.

The Pass-Fail Consistency Index proposed by Breyer and Lewis (1994) is a procedure for estimating the probability of consistently classifying examinees to mastery or non-mastery states using examination score data from one administration. It can be used with examinations containing a combination of items that are scored dichotomously and those that are scored polytomously. Basically, the index is calculated by using the resulting data when the examination is divided into two comparable (but not necessarily

strictly parallel) half-examinations, each with its own cut-off score. The index then predicts the probability of a consistent classification for the full examination.

Chang, Dziuban, Hynes and Olson (1996) found that when judges use the Angoff procedure to set standards, they tend to set higher and more consistent standards for items they answered correctly themselves and lower and less consistent standards for items they answered incorrectly. They note that because judges have different professional backgrounds, even if they form a relatively homogeneous group, they may not be uniformly familiar with every item on the examination. This may lead to quite different item cut-off scores being proposed. To improve the level of both the intra-judge and inter-judge reliability, Chang *et al* advocate that the judges should be trained or briefed in the content domain for which they are to set the competency standard. This, they claim, may help overcome some of the effects due to judges' different experiences and expertise. Combining this approach with an iterative standard-setting process, where judges discuss their decisions and have the opportunity to modify their standards as a result of those discussions, may lead to an improved level of reliability.

Berk (1996) claims that even when procedures are put in place to improve inter-judge consistency (eg using a multi-stage approach to set standards), it can still be relatively low due to factors such as ambiguity in definitions of achievement levels, differences in the competence of judges and the background characteristics of judges. Nevertheless, even when there is considerable inter-judge variability, Berk states that the mean or median of the cut-off scores proposed by the individual judges can still be used as a cut-off score for an examination.

2.4.2 Validity

Jaeger (1990) states that validity in relation to a standard-setting exercise relates to whether students are classified correctly as a result of the process: those students who are competent should be classified as competent; likewise, those who are not competent should be classified as such. He notes that those students near the cut-off score will be virtually indistinguishable in their achievements and so pass/fail distinctions near the cut-off score will have poor validity.

This view is shared by Shepard (1980), who states:

> "Individuals immediately on either side of the standard will be virtually indistinguishable from one another. With a good test, valid distinctions can be made between those who are well above or well below the standard; but pass-fail distinctions near the cut-off will have poor validity because a continuum of performance has been arbitrarily dichotomised" (p. 448).

Shepard claims that a fundamental problem for judgmental standard-setting methods is the disagreement between judges, which can threaten the validity of the exercise. Simply averaging judges' scores may cover up considerable variation in individual standards. To protect the validity of the standard-setting process, Shepard suggests that attempts should be made to: ensure different value positions and areas of expertise are represented among the judges; collect evidence of important differences of opinion among the judges during the exercise; and take any other measures which might validate the results. For example, one such approach is to determine the cut-off scores independently by using a different standard-setting procedure.

Messick (1994) holds that whether the cut-off scores set are appropriate depends upon the defensibility of the procedures used for determining them. As the process is dependent upon the judgments of the panels, the validity of the standards depends upon the reasonableness of the process and its outcomes and consequences.

Messick believes that a weakness of judgmental procedures such as the Angoff method is that the judgments are made at the item level for each item separately. When each item is considered in isolation, the item-specific variance is large compared to the construct variance. This tends to distort the probability estimates that are supposed to reflect minimal competence. A further weakness, he believes, of the item-by-item approach is that the judgments do not capitalise on the structure of the interrelations among the items, as do Latent Trait Theory scaling or other model-based approaches to developing measurement scales. Messick's concerns can be addressed to a significant degree, however, by having the judges reflect, not only on the individual item cut-off values, but also on the total cut-off score for the examination, once the item cut-off scores are aggregated. If they wish, the judges can modify their examination cut-off scores where they feel it is warranted.

Berk (1996) agrees that the internal validity of a judgmental standard-setting procedure is dependent upon the expertise and experience of the judges and the application of the procedure itself. He claims that, even after a thorough and rigorous standard-setting procedure has been applied, the final standard is "whatever the judges say it is." This, he states, is "certainly not a compelling argument for validity evidence, but the credibility of the group of content experts and procedural fidelity are the only available internal criteria" (p. 230).

Berk believes that only external evidence can indicate whether the correct cut-off scores have been set. One source of such evidence is the consideration of the consequences of a particular cut-off score. Berk's view is that consideration of the political, economic, social or educational outcomes of decisions about examinees is needed to determine to what extent the cut-off scores should be raised or lowered. A second type of evidence that can assist in this process is what Berk refers to as "evidential" or "decision validity evidence". Using this approach, information is collected on the performance of students on an examination and their performance in positions or responsibilities they are given. By analysing this information, a measure can be taken of how well students who achieve a certain cut-off score perform in particular employment situations or courses of study. This provides evidence concerning the consequences of pass/fail decisions. Such an approach, however, is not always practicable.

Norcini and Shea (1997) claim that:

> "rather than speaking of validity, it makes more sense to focus on technical considerations and accumulate evidence to support the use of a particular standard for a particular purpose. Stated another way, it is more important to write of collecting evidence to support the credibility of a standard, rather than to validate it (attempt to establish its correctness) because the latter is not possible" (p. 40).

In their view the type of evidence which would support the use of a particular cut-off score would include the qualifications and experiences of the judges, the rigour of the procedure used, and whether it is generally accepted that the standard is realistic.

2.5 SUMMARY

This review of the literature on equating examinations shows that a number of techniques have been developed. While some methods, such as the use of the same examination paper, or the use of common items, are relatively easy to administer, they are not applicable in all situations. In circumstances where a new examination paper is prepared for each examination with no items being repeated, other methods must be used.

The use of teams of judges to equate different forms of an examination is employed in some high-stakes, curriculum-based examinations. The success of such approaches is dependent upon a number of factors. Among these are how well standards of student performance can be defined and understood by the judges, and the integrity and effectiveness of the procedure used. The judges selected should be capable of performing the task, and should receive appropriate training and advice. They should have the opportunity to discuss and refine decisions they have made individually, and statistical feedback and student scripts should be provided to give the judges a further opportunity to review and refine their cut-off scores.

Many procedures have been used in the past in an attempt to set valid and reliable standards of student performance on an examination. While many of the earlier procedures are suitable for examinations consisting of items which are dichotomously scored, the increased use of items which are polytomously scored and more complex types of examination, means that new or modified procedures are required. Criticisms of traditional standard-setting methods have also led to changes. Many judgmental-

empirical methods employed today use a multi-stage approach. They pay particular care to the selection and training of the judges, facilitate discussion among the judges about their decisions, provide statistical feedback of some form on the performance of the students, and require judges to review student scripts before finalising their cut-off scores.

CHAPTER 3

THE PROCEDURE

3.1 INTRODUCTION

In this chapter, a procedure for equating different examinations in the same course across years is articulated. The manner in which the procedure is used in an initial year to develop performance standards is discussed, and then, how the procedure is applied to impose those same performance standards on a subsequent examination is addressed. Finally, the latent trait models based on the work of Rasch are introduced. These models, particularly the Extended Logistic Model, are a key element in the standard-setting procedure.

The previous chapter introduced a number of different procedures that use professional judgment to establish performance standards. Most contemporary procedures use teams of judges specially trained for the task of establishing standards. These judges first determine cut-off scores independently of each other, and then, as a group, review the decisions they have made. In order to better inform this review, the judges are usually provided with additional material to consider, such as samples of student work or statistical data on student performances.

The procedure for setting standards and equating examinations proposed in this study has the features outlined above, but varies from other similar procedures in a number of

significant ways. First, the judges are brought to a consensus position earlier in the process than is usually the case. In fact, many procedures do not require the judges to reach consensus, simply that they have the opportunity to vary their initial decisions on the basis of discussion and consideration of further evidence. Secondly, the statistical information provided to the judges is linked directly through the use of Latent Trait Theory, specifically the Rasch Model, to the variable defined by the examination. The item level information provided by the use of the Rasch Model is presented, in conjunction with student performance data, to the judges who have the opportunity to use them in refining their decisions.

The procedure is flexible enough to be used with a variety of types of examination. In particular, it is suitable for setting cut-off scores on large-scale, high-stakes, curriculum-based examinations in a range of different courses. Such examinations often consist predominantly of items that are scored polytomously. The procedure does not require any changes to be made to the structure or specifications of the examinations, nor does it place any constraints or limitations on the administration of the examinations.

3.2 THE PROCEDURE USED TO SET STANDARDS

Most high-stakes curriculum-based examinations contain a combination of items, some of which are scored dichotomously and some polytomously. In addition, some of the items in the examination may be optional. Such examinations usually have large candidatures and often cut-off scores are set by people who do not know the students and the standards of performance of which they are capable. Using a student-centred standard-setting procedure in such circumstances is difficult, if not impossible. While

methods such as the Contrasting Groups and Borderline Groups procedures are feasible on a small scale, it is difficult to adapt such methods and use them with confidence in large-scale examination environments. Consequently, the most appropriate type of procedure uses an examination-centred approach - that is, one in which the cut-off scores are determined by consideration of the instrument itself. Such an approach has the added advantage that a significant component of the process of setting the cut-off scores can be undertaken before students attempt the examination.

The Angoff procedure provides the basis for the standard-setting procedure used in this study. Apart from its focus on the test, it is also flexible and simple to use. In addition, procedures based on the Angoff approach have a documented history of producing reliable results in contexts similar to the one described in this study. The key elements of the procedure used in this study are explicated in the following section.

3.2.1 The Initial Steps

In most Angoff-based procedures, once judges acting independently have established how they believe a marginally competent student will perform on each item, they are brought together to compare this information with the other judges. Usually, the judges discuss the decisions made and, if they wish, modify the item cut-off scores they originally proposed. Jaeger (1982), Norcini *et al* (1987), Mills *et al* (1991), Morrison *et al* (1994), Hambleton and Plake (1995) and Berk (1996) all advocate procedures where some form of feedback on the individual decisions of the judges is provided.

The procedure proposed here is consistent with many procedures based on the Angoff approach in that the judges are trained for their task and made fully aware of the purpose of the exercise. When this is achieved, and each judge has followed the first steps in the procedure and estimated a set of item cut-off scores, judges are brought together to discuss the cut-off scores submitted.

The approach used in this study differs from many similar approaches in the purpose of the discussion about the item cut-off scores. In this procedure, where there are differences between the item cut-off scores estimated by the judges, the discussions proceed with a view to reaching consensus. This is contrary to what happens in a number of other standard-setting procedures where, after some discussion, the values nominated by the judges are averaged to give the item cut-off score. The reasons for this variation from common practice are discussed in a later section.

3.2.2 The Statistical Feedback and Its Use

Statistical feedback on the performance of students on each item is provided to the judges. This feedback evolves from an analysis of the examination data using the Extended Logistic Model (Andrich, 1978; Tognolini and Andrich, 1995). This model has a number of features which facilitate the understanding of test-generated data, one of which is that items and students are located on the same variable. As a result, judges can be presented with data showing the expected score students are likely to have obtained on each item, given their performance on the overall examination. As noted in the previous chapter, a number of researchers have explored the use of Latent Trait Theory (LTT) in different ways as part of the standard-setting process. Among these

are Van der Linden (1982), McKinley *et al* (1996), Kahl *et al* (1994), Stephanou in McGaw (1997), Engelhard and Cramer (1997) and Engelhard and Gordon (1997).

In other studies (eg Linn, 1978; Norcini *et al,* 1988; Fehrman *et al,* 1991; Reid, 1991) the statistical feedback has usually taken the form of item facility values or frequency distribution tables. In some cases, account is taken of the proportions of students gaining above a certain score when finalising the cut-off score.

Plake (1998) sounds a note of caution in the application of Item Response Theory (IRT) methods in the setting of standards. She states that such methods assume the unidimensionality of the examination. This is clearly an issue in methods that depend upon the measurement properties of IRT (or Latent Trait Theory (LTT)) models to set the cut-off scores. It will be discussed later that, in the procedure employed in this study, a Rasch model is used to provide data that informs decision-making. These data are only one form of the information made available to the judges. Thus, the issue of unidimensionality is not as critical an issue as it might otherwise be given the manner in which the latent trait model is used in this study. It also needs to be noted that the usual practice of adding the scores obtained on each item to obtain a total score for an examination assumes unidimensionality and accepts, without question, that the component parts are relatively unidimensional.

The Extended Logistic Model (ELM) is used as the method of providing the statistical feedback in this procedure for a number of reasons. Firstly, it is a more appropriate approach for use with items that are scored polytomously. Secondly, as it is based on a

measurement theory, it makes explicit any anomalies in the data. In this way it highlights issues related to the student performance data which the judges should consider when setting standards.

3.2.3 Review of Student Examination Responses

The scripts of some students who achieve scores equal to the proposed cut-off scores are given to the judges. The purpose for this is so that they can satisfy themselves that the standard of knowledge and skills exhibited by these students in the examination is consistent with the "imaginary" borderline students the judges have identified during the process of establishing the cut-off scores. The judges review these scripts individually and then discuss with their colleagues their views as to whether these scripts represent the performances of borderline students. In doing this, the judges take an holistic view of the scripts to ensure the concerns raised by Messick (1994) are addressed.

If there is any doubt as to whether the students' performances as reflected in these scripts are truly "borderline", the judges are given other scripts to review. Depending upon the judges' wishes, these may be a further sample of scripts that were awarded the proposed cut-off score, or ones that received a slightly higher or lower score. The judges have the opportunity to vary their cut-off scores, if they believe it is warranted, after careful consideration of the sample scripts. This step is consistent with the practice advocated by Mills *et al* (1991) and Berk (1996), that has been shown to produce reliable standards. A schematic representation of the standard-setting procedure developed in this study is provided in Figure 3.1.

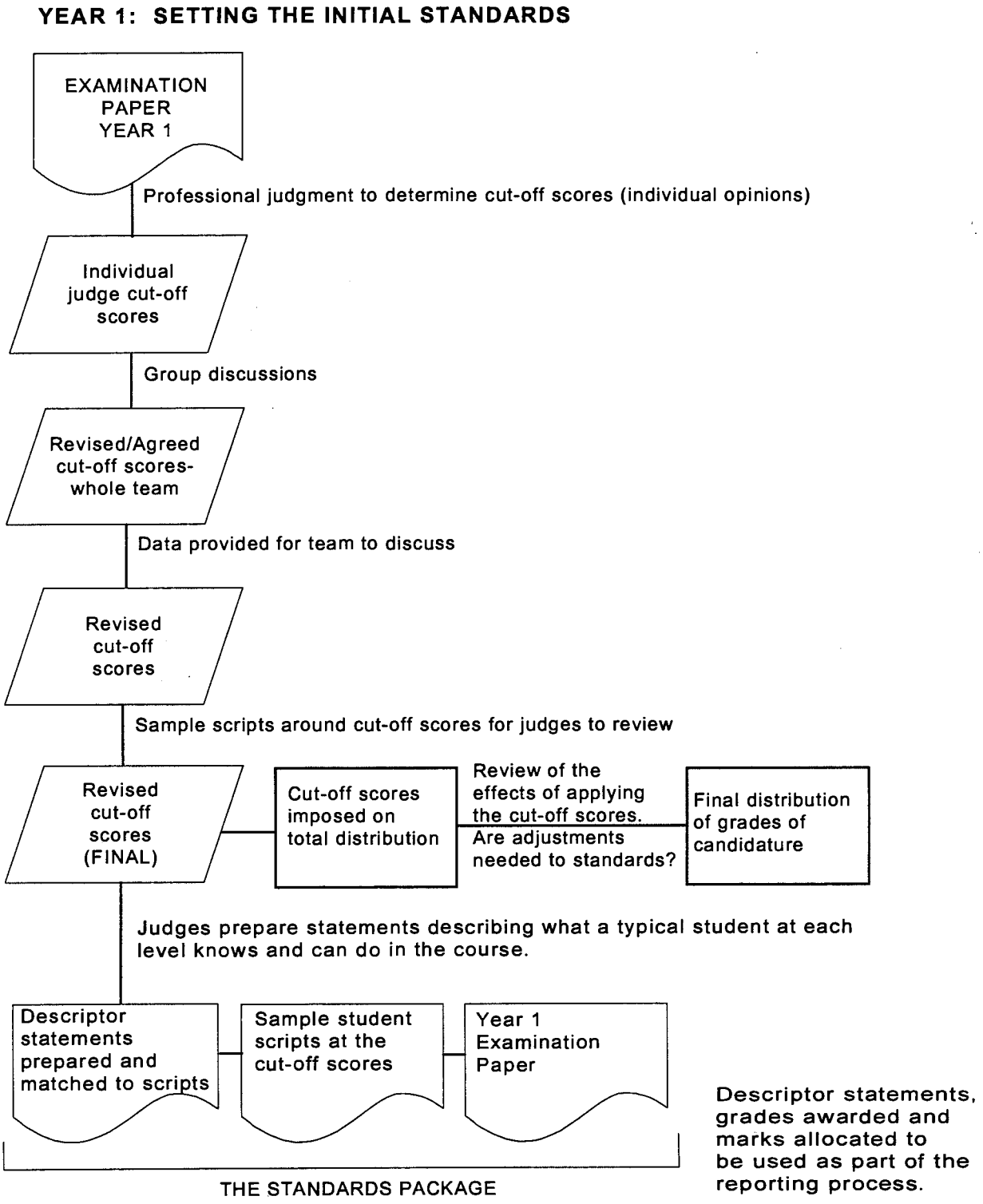

Figure 3.1 The standard-setting model for setting performance standards in an initial year.

3.2.4 Articulating the Standards

The judges prepare a summary descriptive statement of the characteristic performance of students at each standard level. In performing this task, judges consider the expected outcomes of the course, what the items included in the examination are testing, the item difficulties, and the standard of responses in the sample of student scripts chosen. These "descriptors" explicate the standards of performance in the course for that year and are an integral part of the process which enables judges in future years to internalise the standards and perform the equating which enables comparisons of groups of students across years. Kane (1994) and Berk (1996) indicate the value of developing a performance-based interpretation of standards.

These statements of performance (often referred to as grade, level or descriptor statements), the examination papers, and the samples of student scripts awarded the cut-off scores form an integrated package. This package, referred to in this study as the "standards package", is the key to establishing the cut-off scores for future examination papers which correspond to the standards set in the initial standard-setting exercise.

3.3 THE PROCEDURE USED TO EQUATE EXAMINATIONS

The procedure used to equate examinations by setting the cut-off scores for subsequent years is essentially the same as that used in the initial year. Figure 3.2, shows that it contains all the steps included in the initial-year procedure with one major difference.

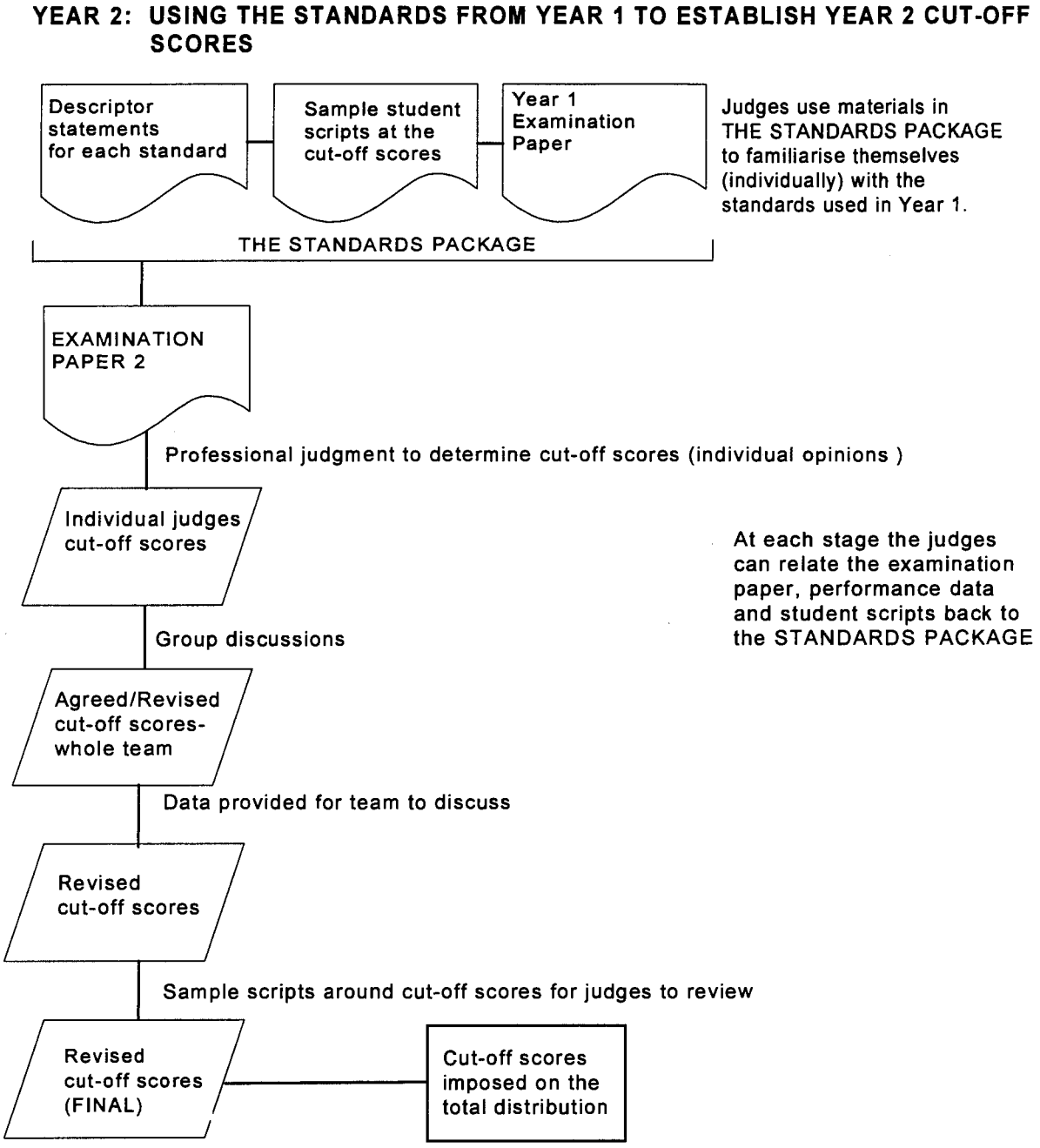

Figure 3.2 **The procedure used to equate examinations by imposing the performance standards developed in an initial year onto a subsequent examination.**

In subsequent years the judges do not create a personal image of students who they would place at the borderline between different performance standards simply based on their own views of what those standards should be. Instead, the judges internalise the performance standards developed in the initial year by studying the standards-related descriptor statements, the examination paper from the initial year, and the samples of examination scripts produced by borderline students from the initial year. They then apply those performance standards to the new examination in the manner described previously, to generate cut-off scores in the metric of the new examination.

3.4 FEATURES OF THE PROPOSED PROCEDURE AND COMPARISONS WITH OTHER PROCEDURES

The features of the procedure adopted in this study for establishing standards and equating examinations is described below.

3.4.1 The Selection of Judges and the Composition of the Panels

In a number of studies, a relatively large panel of judges is assembled. Jaeger (1991), for one, believes that it is wise to use as many judges as possible. Many studies (eg Berk, 1996; Jaeger, 1991) also recommend that a relatively heterogeneous panel be created. Jaeger (1982) used panels consisting of a mixture of registered voters, teachers, school counsellors and principals. Busch and Jaeger (1990) used panels containing school teachers and college/university lecturers.

In the procedure used in this study, the panels selected are relatively small, four to six members, and consist entirely of teachers with extensive experience in teaching and

assessing the course, preparing students for examinations, and, in most cases, scoring the examination responses.

An "expert", rather than "representative", panel is used because of the specific content-based nature of the examinations. In order to maximise the validity and reliability and, hence, the credibility of the outcomes, the judges must have a thorough understanding of the course and its requirements. They must also understand the approaches used and the types of errors made by students who have reached different performance standards, and be able to discuss these matters with their colleagues.

One danger in using an expert panel of subject specialists as judges is that they may be viewed as having a vested interest in the result. This could lead to the perception that the standards are flawed.

Concern about the objectivity of the judges, however, is not really an issue in a large-scale examination environment like the one involved in this study. In such a situation, the results and the standards created will be open to public scrutiny. Standards that are too low, or large groups of students who do not reach expected standards, will result in public criticism of the process. To prevent any criticisms in this regard, it is essential that the process be as open as possible. If used in a high-stakes environment, the judges involved should have credibility both within their discipline and the wider community. Additional checks and audits, such as an independent review of the cut-off scores and the proportions of students reaching the various standard levels, can also be incorporated to give a greater degree of public acceptance.

3.4.2 Training of Judges

Many studies (eg Mills *et al,* 1991; Plake *et al,* 1991; Plake and Impara, 1996) emphasise the need to provide training for those judges involved in a standard-setting procedure. They provide evidence that the effectiveness of a standard-setting (or equating) procedure is enhanced when the judges begin with a clear understanding of what they are doing and why they are doing it. In most cases this training is conducted prior to the application of the procedure. Judges work through and rate samples of scripts, assigning probabilities to individual items according to whether they believe the students at the nominated level could answer the item correctly or not. They then discuss the results with their colleagues and trainers. In the case of items that are scored polytomously, instead of nominating the probability of a correct response the judges nominate an expected score.

The procedure used in this study supports the approach of having all judges in a team attend a formal training session during which they are fully briefed on the task and, if feasible, given the opportunity to apply the first steps of the procedure to some similar items. Where it is not possible to bring the judges together for a training session prior to the exercise, the alternative approach is to brief each judge individually prior to their beginning the task. At the training sessions, judges are provided with detailed briefing notes and have the opportunity to seek further advice and receive answers to any questions they have.

3.4.3 Seeking Consensus amongst the Judges on Item Cut-off Scores

The procedure uses a consensus-based approach that first requires judges to work independently to establish their estimates of the item cut-off scores appropriate for a particular performance standard. The judges then refine these scores as a result of consultation with their colleagues so that the team eventually arrives at an agreed value for each item. These item scores are aggregated to obtain the cut-off score for the examination corresponding to that performance standard.

It is recognised that requiring consensus on the item scores runs the risk of having the panel's decisions determined by a dominant personality, or having judges simply adopt a compromise position. It is also noted that other researchers (eg Jaeger, 1988) have stated that there is value in the differences between judges' decisions which remain after discussion, because these represent different valid professional judgments.

In spite of these concerns, it was decided that, where there were differences between the item cut-off scores proposed by the judges, they would continue discussion until consensus was reached on the appropriate item cut-off score. In the case of items that are scored polytomously there may be a degree of subjectivity associated with the use of the scoring keys, particularly when an holistic approach is used in scoring student responses. In such cases, one judge may succeed in bringing to the attention of his/her colleagues some features of an item or some quality of the likely responses of certain students, which the others had not initially considered. Such an outcome adds to the validity of the standard-setting process.

A second advantage associated with requiring judges to reach a consensus position early in the process is that they are then able to consider each piece of new information presented to them in a cooperative manner and discuss, from a common position, how they should accommodate the information.

3.4.4 The Use of a Compensatory Approach

For this study, a compensatory approach was considered to be more appropriate than a conjunctive one.

The score awarded in an examination is generally the sum of the scores obtained on the individual items. Hence, imposing a further set of conditions for most curriculum-based public examinations, such as requiring students to achieve at least some minimum score on every item, would generally be at variance with the summative nature of the examination. Observations made by teachers and raters over many years indicate that students at all levels can perform above or below expectations on any item under examination conditions, but frequently an unexpectedly poor performance on one item is balanced by an unexpectedly good performance on another item.

Were a conjunctive approach to be used, out of fairness, students would need to be told the minimum score required on each item or task before the examination was administered. Otherwise, students could spend a disproportionate amount of time responding to items where a low item score was required, unaware that they needed, perhaps, to spend more examination time responding to items where a higher minimum score was required. The concerns identified by Plake, Hambleton and Jaeger (1997)

over the use of a conjunctive approach cast doubt on its appropriateness in examinations of general education courses.

3.4.5 The Use of a Compromise Approach

Figure 3.1 shows that one of the last steps in the procedure in the initial year is to review the effects of applying the proposed cut-off scores. Built into the process is an option to consider whether applying the cut-off scores will give proportions of students in each standard level which are acceptable to the stakeholders.

Beuk (1984), De Gruijter (1985) and Hofstee (1983) draw attention to the need to generate "acceptable" proportions of students at each standard level. While they clearly support a standards-based approach, they also conclude that the standards should be set by taking account of what might be referred to as the "reasonableness" of the results.

It is not essential that such a step be taken when using this procedure, rather it is a matter for consideration when given the context and the purposes of the exercise. Hence, in the example developed in this thesis, no attempt was made to adjust any of the cut-off scores finally established by the judges to achieve what might be regarded as "reasonable".

Were this procedure to be used operationally in a high-stakes examination, it would be necessary to take account of such things as expected pass rates in the initial year when the performance scale is being created, without reducing the integrity of the procedure.

3.5 RASCH MODELS

The latent trait model, which is used to provide the statistical feedback in this study, is the Extended Logistic Model (ELM) (Andrich, 1978; Tognolini and Andrich, 1995; Tognolini and Andrich, 1997). It is an extension of Rasch's Simple Logistic Model (SLM) (Rasch, 1960/1980; Wright, 1977; Andrich, 1978; Wright and Stone, 1979; Andrich, 1988).

This particular Rasch model was chosen because its properties permit the analysis and presentation of examination data in a form particularly suitable to the type of approach followed in this study. The model can provide meaningful and useful feedback to the judges on the performance on the examination items by students of different abilities. This form of feedback is useful, both during the process of setting the performance standards in the initial year, and in equating the examinations administered in different years.

3.5.1 The Simple Logistic Model

The Simple Logistic Model (SLM), which is applicable when dealing with items which are scored dichotomously, holds that the probability of a particular outcome when a student attempts an item in an examination is a function of the student's ability and the item's difficulty only. The SLM is one of a number of measurement models proposed by Rasch.

The SLM is commonly represented as follows:

$$P\{X_{ni} = x_{ni};\beta_n,\delta_i\} = \frac{e^{x_{ni}(\beta_n-\delta_i)}}{1+e^{\beta_n-\delta_i}} \qquad \textbf{Equation 3.1}$$

where

$P\{X_{ni} = x_{ni};\beta_n,\delta_i\}$ = the probability of each outcome when student n attempts item i;

β_n = the parameter describing the location (ability) of student n on the variable;

δ_i = the parameter describing the location (difficulty) of item i on the variable; and,

x_{ni} = 1 or 0 depending upon whether student n answers item i correctly or incorrectly

For items which are scored dichotomously, X_{ni} takes the value 1 if the answer is correct, and the value 0 if the answer is incorrect. From Equation 3.1, it can be seen that the probability that student n with ability β will answer item i correctly which has difficulty δ is given by

$$P\{X_{ni}=1\} = \frac{e^{(\beta_n-\delta_i)}}{1+e^{\beta_n-\delta_i}} \qquad \textbf{Equation 3.2}$$

As the denominator $1 + e^{\beta_n-\delta_i}$ is a normalising factor that ensures that $P\{X_{ni}=1\}+P\{X_{ni}=0\}=1$, it follows that the probability of a correct response is governed by the numerator $e^{\beta_n-\delta_i}$, or more particularly, by $\beta_n-\delta_i$.

If β_n is equal to δ_i then $P\{X_{ni} = 1\} = 0.5$

However, if

(i) $\beta_n > \delta_i$ the person is more likely to answer the item correctly

(ii) $\beta_n < \delta_i$ the person is more likely to answer the item incorrectly.

The greater the value of β_n compared to δ_i the more likely the person is to answer the item correctly.

The relationship expressed in Equation 3.2 can be represented graphically as shown in Figure 3.3. Such a curve is generally referred to as an Item Characteristic Curve (ICC). The ICC shows the probability as a function of person ability (β) for an item with location (δ) fixed.

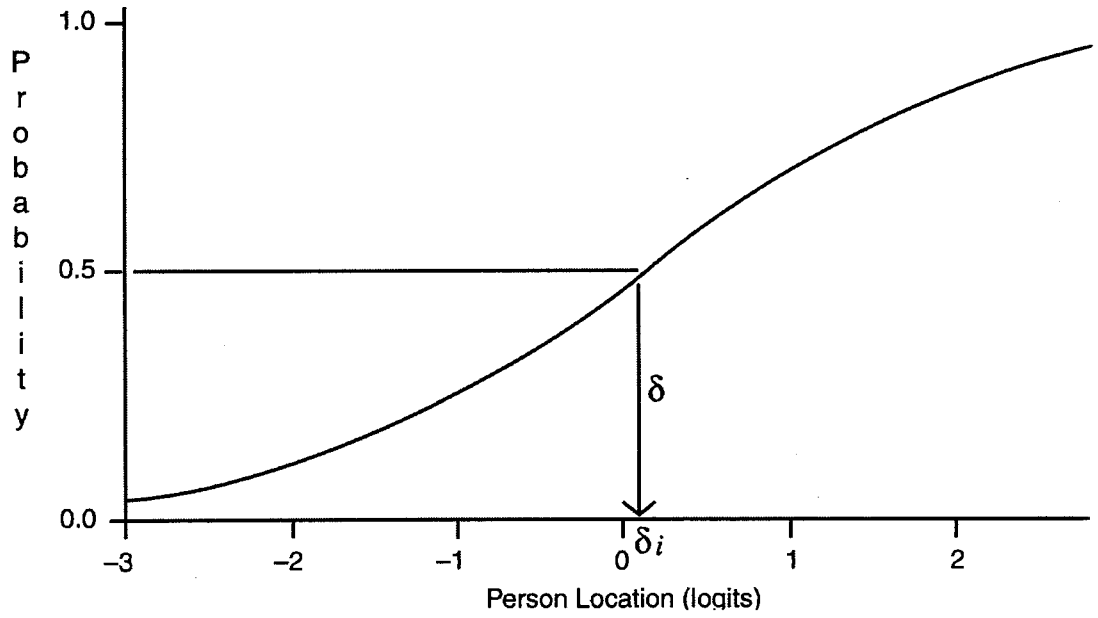

Figure 3.3 Item characteristic curve for dichotomously scored items

As can be seen from Figure 3.3, the greater the ability of the person, the more likely he/she is to obtain the correct answer to the item.

The SLM has the property usually referred to as specific objectivity. This property is a feature of many measurement models used in the physical sciences. Models that have this property satisfy two conditions. First, the calibration of the measuring instrument is independent of those objects used to perform the calibration. Secondly, the measurement of objects is independent of the instrument that happens to be used for measuring (Wright, 1967). In the case of mental measurement, the property of objectivity requires that the calibration of test items is independent of the particular students used for calibration, and that the measurement of student ability is independent of the particular examination items used for measuring. This requires, of course, that the items adequately define the variable and that the group of students being measured has similar characteristics to the group on which the items were calibrated.

This property, which is also shared by the Extended Logistic Model, is particularly useful for providing statistical feedback to the judges in the standard-setting procedure. Once the items have been calibrated, the ability estimates of students can be determined irrespective of whether they responded to the same calibrated items or not.

Another important feature of the SLM is that the raw score is a sufficient statistic. That is, the total score obtained by a student on an examination is a sufficient statistic for calculating the ability of the person (Rasch, 1960/1980; Choppin, 1983).

The SLM also provides a framework for verifying that the data used in the construction of the variable are suitable. If the data do not conform, or cannot be brought to conformity with the model, then the estimates of item difficulties and student abilities will be poor. Hence, if the responses of a student do not fit the model it is not appropriate to use the variable to measure the student's ability.

The issue of how well the data fit the model is of somewhat less importance in this procedure for setting performance standards than in other applications of Latent Trait Theory. When using the procedure developed for this study, the data provided by the Rasch model are used solely to provide feedback, and hence inform debate on the performance of the total candidature of an examination. The Rasch model is not used to create the performance standards, as in some other approaches. This is discussed in the final part of this chapter.

3.5.2 The Extended Logistic Model

Whereas the SLM is only appropriate for use with items which are scored dichotomously, the Extended Logistic Model (ELM) can be used with items which have more than two ordered response categories.

The ELM can be expressed mathematically as

$$P\{X_{ni} = x_{ni}; k_x, \phi_x, \beta_n, \delta_i\} = \frac{e^{[k_x + \phi_x(\beta_n - \delta_i)]}}{1 + \sum_{i}^{n} e^{[k_x + \phi_x(\beta_n - \delta_i)]}} \qquad \textbf{Equation 3.3}$$

where $k = -\tau_1 - \tau_2 - \ldots - \tau_x$;

τ is the threshold between the ordered response categories

ϕ_x = a scoring function associated with each category

In the ELM, each student is still described by the single parameter ability (β) and each item will still have a difficulty (δ). Each item, however, will have at least three ordered response categories. An item for which the values 0, 1, 2, ..., m can be awarded will have $m + 1$ ordered response categories. For example, if the item has six categories, then the possible scores on the item will be 0, 1, 2, 3, 4 and 5. The point of change between one category and another is referred to as a threshold. The difficulty of an item that is scored polytomously is equal to the mean of its threshold values.

Figure 3.4 shows the case where a person with ability β attempts a single item which has difficulty δ and threshold values τ_1, τ_2, τ_3, τ_4.

Figure 3.4 Item with four threshold values and student with ability β represented

In this case the student is more likely than not to obtain the score corresponding to threshold τ_2, but likely not to obtain the score corresponding to τ_3. This relates to an item for which a student could score 0, 1, 2, 3 or 4.

Figure 3.5 shows a Category Characteristic Curve (CCC) reflecting Equation 3.3 for an item with six categories. It demonstrates that the probability of obtaining a high score on an item with difficulty δ increases as the ability β of the respondents increases.

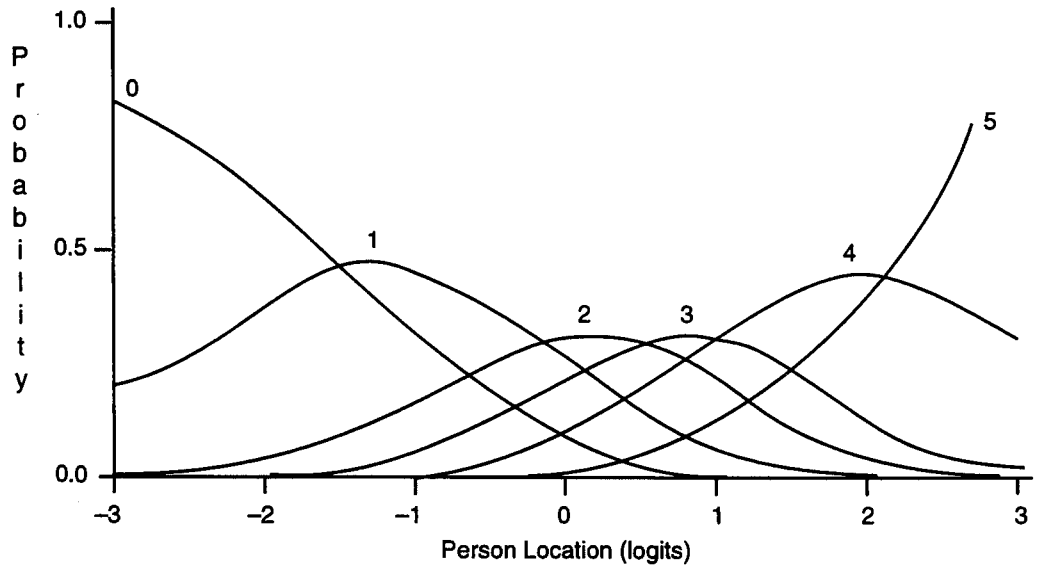

Figure 3.5 Category Characteristic Curve for item with six possible response categories

3.6 THE ROLE OF LATENT TRAIT THEORY IN THE PROCEDURE

In the procedure proposed in this thesis, the results of analysing a sample of student performances in the examination by using the ELM are presented to the judges. This is done in a manner that shows how the ability level associated with a particular cut-off score relates to the individual item cut-off scores the judges had associated with that standard. They use this information to review their expectations of what score students at the cut-off scores would achieve on each item.

The relevance of the ELM to the standards-setting model relates to it's capacity to relate ability and expected score. This is illustrated by Figure 3.6 that shows possible Expected Value Curves (EVC) for three items.

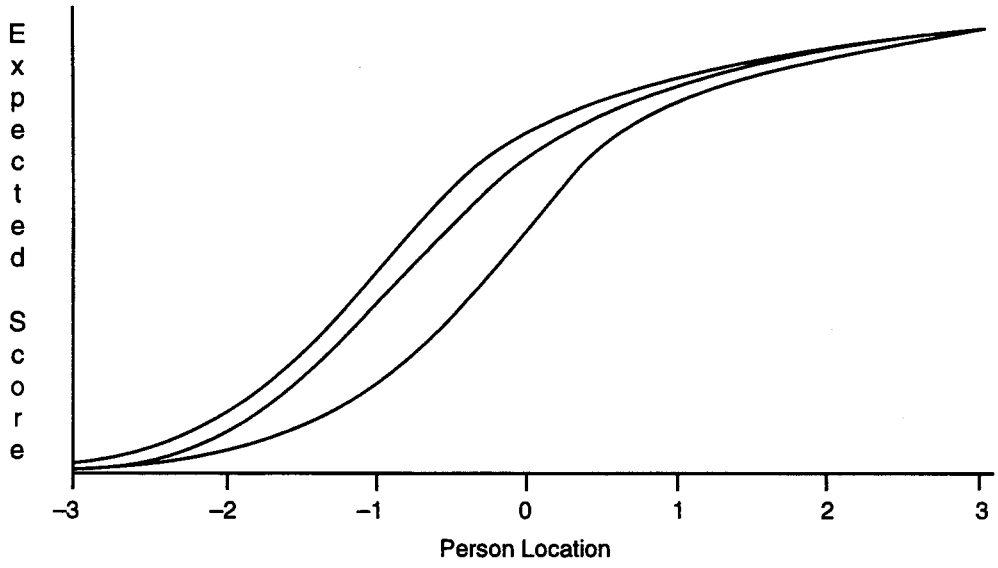

Figure 3.6 Expected values curves corresponding to three items

Using information provided by the analysis of a sample of student performance data the judges reconsider any items where there is a difference between the score they estimated students on a borderline would achieve and the score estimated for such students by the ELM. In these cases, the judges discuss the item and their initial decision in the light of the information provided by the model. After this discussion, if the judges accept that their initial value was inappropriate, they change it. On the other hand, if the judges wish to retain their original value in spite of the score estimated by the ELM, they do so.

3.7 SUMMARY

By taking note of strategies which have been used successfully in the past and incorporating a number of new features, a procedure has been developed for equating examinations by setting appropriate cut-off scores for each examination corresponding to pre-determined standards of performance. Once the standards of performance have been carefully described and exemplified by using student examination scripts at appropriate score points, the procedure can be used by a team of trained and experienced judges to apply those same standards of performance to examinations held in subsequent years. In this way it is possible to compare the performances of groups over time.

The analysis of examination data by the use of the ELM provides statistical feedback to the judges in this procedure. The model is used to estimate the likely score that would be obtained on each item by students with abilities corresponding to each cut-off score. Presenting this information in diagrammatic or tabular form provides judges involved in such an exercise with an indication of what score in each item students with particular abilities are likely to obtain.

CHAPTER 4

THE APPLICATION OF THE PROCEDURE

4.1 INTRODUCTION

In the previous chapter, a procedure was developed to establish standards so that different forms of an examination could be equated. Judges are used to internalise the standards and then determine cut-off scores corresponding to the borderline performance that distinguishes the different performance standards. The procedure builds upon strategies shown to be successful in similar situations in the past. In this chapter the procedure is demonstrated by applying it to examinations for three courses in a major public curriculum-based examination program.

4.2 EXAMPLES

As part of the requirements for the New South Wales Higher School Certificate (HSC), students sit for an examination, at the end of their secondary school studies, in each of the courses they have taken. These courses are based on traditional subject disciplines, such as English, History, Mathematics and French. The examinations, which are held in November each year, test the specific knowledge and skills covered in each course. Results from the examinations are used by tertiary institutions to select students for further education and training programs and by employers to select potential employees.

Every year an entirely new examination paper is prepared for each course. The examinations are usually three hours in duration and contain a variety of different item types. They all contain open-ended items that are scored polytomously, and some

contain a small proportion of objective-type items that are scored dichotomously. The former may require students to solve mathematical problems, write essays to argue a point of view, or produce short written responses to specific stimuli. Although different types of items that are scored dichotomously are used, the most common is the four-option multiple-choice item. In addition to traditional pen-and-paper format, many examinations require students to undertake some form of performance or submit a significant piece of work, such as an artwork or a major project, for assessment.

Students receive a score, expressed as a percentage, for their performance in the examination. No attempt is made to reference this score to any standard of performance when reporting student achievement, other than the performance of the group of students attempting the examination.

4.2.1 The Courses Used

Three different courses from the HSC program (Mathematics, English and Biology) are used in this study to demonstrate the application of the procedure. These courses are chosen because they represent a cross-section of the types of knowledge and skill assessed, and the types of examination used to assess performance. In the subject areas of Mathematics and English, there are several courses from which students can select. One course from each of these subject areas was chosen for the study. The implications of this are considered in Chapter 6.[1]

[1] The courses used are the 2 Unit Mathematics course, the 2 Unit General English course and the 2 Unit Biology course. For simplicity, these are later referred to in this study as Course B in Mathematics and English. This is not necessary in Biology where there is only one course offered.

4.2.1.1 Mathematics

The Mathematics examination consists of ten compulsory items to be completed in three hours. Items are generally divided into at least two parts and most parts are divided into sub-parts. The separate parts in each item are usually based on separate content areas of the course. The examination has a highly prescriptive answer key requiring virtually no interpretation on the part of those scoring the student responses.

4.2.1.2 English

The English examination consists of two papers, each of two hours' duration. The first paper (Paper 1) is entitled *Uses of English and Topic Areas*. It has three parts; each part contains one item. Part A (or Question 1) consists of a reading task. Generally there are five or six sub-parts to an item with maximum possible scores usually ranging from 3 to 5. Part B consists of a writing task. Students are required to produce a piece of writing of 300-500 words in length in response to a short piece of stimulus material. In Part C, students are required to produce an extended piece of writing related to one of a number of "topic areas" they have studied.

The second paper (Paper 2) is entitled *Responses to Literature*. It also has three parts. Part A tests the students' skills related to poetry texts in the course, Part B the fiction texts (or novels) and Part C the drama texts. Students are required to answer one item from each part using an essay format. During the course, students study the works of a number of different authors in each genre. In some years there are optional items within each part of the examination paper relating to different authors. Students can choose from items based on the authors and works they have studied.

The examination consists solely of items that are scored polytomously. Students provide a written response to each item. These responses may vary from a few lines for some items to several pages for others. In some sections of the examination paper, students have a choice of items based on particular texts they have studied. Each of the six items across the two papers carries a maximum score of 20.

4.2.1.3 Biology

The examination paper for Biology in 1994 (the first year of this study) consisted of:

12 multiple-choice items worth one mark each (compulsory) - Section A

6 short free-response items worth three marks each (compulsory) - Section B

6 short free-response items worth five marks each (compulsory) - Section C

12 free response-items worth thirteen marks each (optional-students were required to answer three items).

A major change to the content and scope of the Biology course was introduced in 1995 (the second year of this study). In 1995, the examination paper comprised:

15 multiple-choice items worth one mark each (compulsory) - Section A

10 short free-response items worth three marks each (compulsory) - Section B

6 free-response items worth five marks each (compulsory) - Section C

7 free-response items worth twenty five marks each (optional-students were required to answer one item).

The significance of this change will be discussed in a later section as it did affect the process of equating the examinations across the two years.

4.2.2 Setting the Standards

4.2.2.1 Establishing the Initial Cut-off Scores

A team of judges consisting of teachers who had considerable experience in teaching and preparing students for the examination was established for each course. Each teacher was familiar with the standards of work typically produced by students in the course. In the case of Biology, however, not all judges were thoroughly familiar with every elective (or optional) topic in the course.

It was intended to hold a formal joint training session for each team, prior to having the judges commence their task. Unfortunately, attempts to bring the judges in a team together for such a session were unsuccessful, given their other commitments, so it was determined to brief each judge separately. As the teams used in this study were relatively small, it was feasible to do this. The judges were also given detailed notes on how to proceed and were able to obtain answers to any questions that arose during the exercise. Any issues raised by one judge that had implications for the way the judges were to conduct their tasks were resolved. Advice about the decisions taken was then conveyed to the others who were not present. After the initial steps of the procedure, the judges worked as a team and so it was possible to explain what they needed to do before they undertook each of the remaining steps. A detailed explanation of the nature and meaning of the statistical data was provided at the appropriate point in the study.

The judges first considered the examination paper. Where it was appropriate (in Mathematics and some parts of Biology), they prepared solutions to each item. In other cases they made notes on the features typically required in a response. This process not

only made them familiar with each item, but also enabled them to identify likely sources of error for the less able students.

The judges were next required to visualise the characteristics of students who were, in their opinion, on the borderline between Excellent and Very Good in the course. They then worked independently to estimate the scores that they felt such students would obtain on each part or sub-part of an item. This process was repeated for students on the borderline between Very Good and Good, Good and Satisfactory, and Satisfactory and Unsatisfactory respectively. No attempt was made to define the terms Excellent, Very Good, Good and Satisfactory before the judges made these initial estimates.

A particular complication arose with English. The two papers that comprise the examination are scored in different geographical locations. As the first step of the procedure was to be conducted using as judges, teachers who were involved in scoring the student scripts, it was decided to form two groups of judges initially. One group consisted of four judges involved in scoring the responses to Paper 1, the other consisted of four judges involved in scoring the responses to Paper 2. The two groups worked independently of each other during the early stages of the standard-setting exercise.

Even though the members of each of the English groups had only been associated with the scoring of one paper in 1994, due to their extensive experience in teaching all aspects of the course, each judge estimated item cut-off scores across the whole examination. The two groups were later amalgamated into a single team, and reached agreement on the item cut-off scores, prior to considering the statistical feedback.

4.2.2.2 Reaching Consensus on Cut-off Scores

Once each judge in a team had determined his/her cut-off scores, the team was brought together to reach consensus on the cut-off score for each item depicting each standard level. Starting with the Excellent/Very Good borderline, the cut-off scores determined by each of the judges for each item were considered by the group. The judges considered all items, but focused their discussions on those items where there was a lack of agreement on the cut-off score. By discussing the contents of the item, the knowledge and skills required to answer it, and the characteristics of student performance associated with a particular score, the judges reached a consensus decision on the cut-off score for each item. Through this process of discussion and debate, the judges developed a shared view of the knowledge and skills possessed by students at each borderline.

Item cut-off scores for each borderline were then added to provide the four cut-off scores for the examination. The cut-off scores determined by the judges working with the Mathematics course are provided in Table 4.1. It can be seen that a score of 119 out of a possible 120 was the cut-off score initially set by the Mathematics judges for a borderline Excellent/Very Good performance. The only item that did not receive a perfect score was item 10. According to the judges, a borderline Excellent/Very Good student would be expected to obtain at least 11 out of 12 on this item.

In the case of English, logistical constraints that initially existed meant that the group involved in scoring Paper 1 was brought together at a different time to the group involved in scoring Paper 2. This resulted in the two groups reaching only some

TABLE 4.1

Initial Cut-off Scores for Each Standard Level in Mathematics [1994]

Item	*Excellent/ Very Good*	*Very Good/ Good*	*Good/ Satisfactory*	*Satisfactory/ Not Satisfactory*
1	12	12	10	8
2	12	11	9	6
3	12	11	9	7
4	12	11	8	6
5	12	11	8	5
6	12	9	7	4
7	12	10	7	4
8	12	10	6	2
9	12	9	6	4
10	11	9	4	2
Total Score	**119**	**103**	**74**	**48**

tentative agreement on the item cut-off scores. When it was finally possible to bring the two groups together, it was decided to go back a step and work to get agreement on the item cut-off scores from the values proposed by each judge in the amalgamated team.

Whereas the judges involved with Mathematics and English adopted a similar approach to others in their team in determining their cut-off scores, those involved with Biology used a number of different methods. One of the Biology team specified a score range for each section that a borderline student might be expected to achieve. Another judge used an approach more closely aligned to Angoff's original method of estimating the probability that a borderline student at each level would get the item correct. For those items where the maximum score was greater than one, having estimated the probability that a particular borderline student would answer the item correctly, this judge then

multiplied the probability by the maximum possible score for that item to obtain the expected score. By aggregating the expected scores for each item, this judge was then able to obtain an estimate of the score that borderline students would be expected to obtain in each section of the examination. Judges were permitted to use whatever method they found most comfortable in establishing their initial cut-off scores.

The approach adopted for reaching agreement in Biology was slightly different from that adopted in Mathematics and English. Each judge stated what cut-off scores he/she was proposing for each item. Where there was complete agreement between the judges, no further action was taken. In cases where there was disagreement, discussions preceded consensus. However, as it was decided that the main focus was on obtaining agreement at the section level, it was not considered necessary to get perfect agreement at the item level. Further discussion took place where the cut-off scores produced by the judges for the sections still did not agree, although very little change was required at the section level. Differences revolved around the difficulty of the certain items and the types of errors or lack of skills the judges expected would be demonstrated by students at a particular borderline. Such discussions were effective in assisting the judges to reach consensus quickly about section cut-off scores.

As discussed in a later chapter, it may have been beneficial to require the judges to reach agreement at the item level. This would have strengthened their common understanding of the standards associated with the cut-off scores they established.

The Elective (optional) items in Biology presented some difficulties in the standard-setting process. There were twelve elective items in 1994, from which students were

expected to answer three. Several judges had not taught the sections of the course on which some of these items were based and, so, were not prepared to estimate the scores that borderline students would receive for those items.

After some discussion, it was decided that the cut-off score for each elective item would be determined by averaging the scores proposed by the judges who had submitted a value. Once this was done, an estimated cut-off score was available for each elective item.

Students have a free choice as to which three of the twelve items they respond to. So, for each borderline, it was decided to calculate the average of the cut-off scores proposed for all the elective items. This average was then multiplied by three and added to the cut-off scores calculated for each borderline on the other sections of the examination. While such an approach ignored any differences in difficulty between the items, the judges did not in fact record significant differences in cut-off scores between the elective items, hence there was probably a minimal loss of accuracy.

4.2.2.3 Utilising the Statistical Data

For all courses, samples of approximately 500 students were chosen and the item difficulty and threshold values calculated for each item. Student abilities corresponding to the examination scores were also calculated using the Extended Logistical Model.

Item difficulties, threshold values and student (person) ability measures were plotted on an Item-Student scale. The Item-Student scale for Mathematics is presented in Figure 4.1.

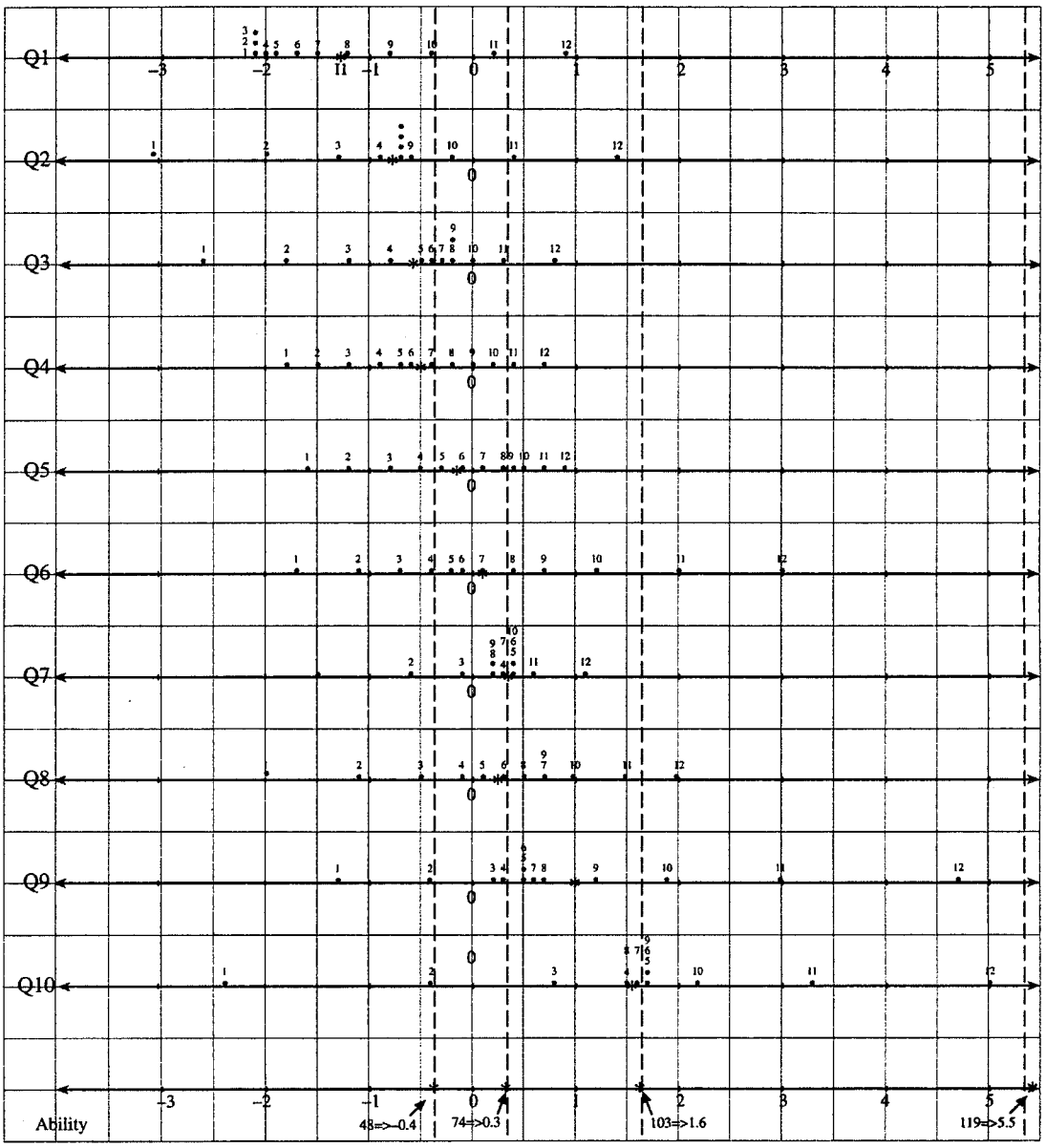

FIGURE 4.1 Item-Student Scale for Mathematics (1994).

Such a scale makes clear to the judges the relative difficulty of each item and the location of each threshold for each item. Vertical lines drawn through the location (ability) corresponding to a cut-off score a team had set showed the judges the score on each item that students who had obtained that particular cut-off score had a 50% chance of obtaining. This diagram is a simpler way of representing the information presented

on a Category Characteristic Curve. From Figure 4.1 we can conclude that a student who scored a total of 48 on the examination is more likely to obtain a score of 10 on item 1 than any other score.

The teams of judges had met to establish their initial cut-off scores soon after the examinations had been attempted. It was several months before it was possible to analyse and prepare the statistical data for their consideration. In order to overcome any deleterious effects due to this time lag, the judges spent time reviewing their previous results. They then considered the statistical information. The judges were told that the scale showed what score a borderline Excellent/Very Good (also Very Good/Good, Good/Satisfactory and might be expected to achieve on each item.

Tables similar to Table 4.2 were provided to the judges. These showed the item cut-off scores (I) previously established by the team, and the expected score (E) for each item produced by the application of the Extended Logistic Model. After being given time to review this information and ask questions, the teams of judges were invited to reconsider their original decisions in the light of the evidence presented in the Item-Student scale and in the table.

Where the statistical information was at variance with the judges' decisions, they discussed once again how borderline students at the particular standard level in question would respond to the item. Sometimes, but not always, they varied their initial decisions on the basis of the data provided by the model. More details of the impact of this information on individual teams are provided in Chapter 5. These discussions also

helped to further refine and consolidate the image of the borderline students that the judges had developed.

TABLE 4.2

Initial Cut-off Score (I) for Each Item and the Expected Score (E) Generated by the Model

	Excellent/ Very Good		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
Item	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**
1	12	12	12	12	10	11	8	9
2	12	12	11	12	10	10	8	9
3	12	12	11	12	9	11	7	6
4	12	12	11	12	8	10	6	7
5	12	12	11	12	8	8	5	4
6	12	12	9	10	7	8	4	4
7	12	12	10	12	7	6	4	2
8	12	12	10	11	6	3	2	3
9	12	12	9	9	6	3	4	1
10	11	12	9	8	4	2	2	1
Total	**119**		**103**		**74**		**48**	

Note: The columns showing the expected scores are not totalled as such a sum has little meaning. The student ability measure corresponding to that sum is not necessarily the same as the student ability measure corresponding to the examination score which produced these values.

4.2.2.4 Reviewing Student Responses

Once the judges had considered the statistical data and modified, or confirmed, their initial cut-off scores, they were provided with the scripts of a sample of students who had received examination scores equal to the cut-off scores they were now proposing. The judges studied the scripts to determine whether they were satisfied that the responses to the items were of a quality they had imagined would be provided by students at the borderlines of the respective standard levels. To confirm their views,

they also reviewed scripts that had been awarded scores near to the values they were proposing.

As the scripts chosen may well have had different scores on some of the items than the item cut-off scores proposed by the judges, the judges took a holistic view of the students' performances. This step gave the team the final opportunity to vary its cut-off scores.

4.2.2.5 Describing the Standards of Performance

The final step in this part of the study was to have the judges prepare descriptions relating to the different levels of performance. Using the objectives and intended learning outcomes of the course, the items in the examination paper, and the image of borderline Excellent/Very Good, Very Good/Good, Good/Satisfactory and Satisfactory/Unsatisfactory students they had developed, the judges wrote statements designed to describe the nature and features of each performance standard. The statements summarised the skills and understandings that a typical student who had reached a particular standard of performance in the course might be expected to demonstrate.

In preparing these statements, the judges once again reviewed the sample of responses at the borderline between one standard of performance and another. The descriptor statements focused attention on what could be expected of a "typical" student. Nevertheless, the judges satisfied themselves that a student whose performance was being classed as borderline Excellent/Very Good (say) had displayed the knowledge and skills required of an Excellent performance, at least to a minimally acceptable degree.

Once these descriptor statements are finalised, they can be used in reporting student achievement in the examination, as they summarise the knowledge and skills generally held by students at each standard of performance. These descriptor statements, however, have a further use. When combined with the examination paper and the student scripts, these statements form a "definition" of the performance standards which are to be applied to subsequent examinations.

4.3 EQUATING EXAMINATIONS ACROSS YEARS

To determine the effectiveness of the procedure for equating examinations across different years, two teams of judges were used for each course in the subsequent year (1995) of the study. The judges who had set the standards in the initial year made up one team. A second team, consisting of other experienced teachers who had not been involved in the study previously, was also established for each course.

The purpose of having two teams for this aspect of the study was to ascertain whether the teams, working independently of each other, could obtain similar results.

It was emphasised to the judges that, in performing this task, they were to apply the same standards of performance established using the 1994 examination to set the cut-off scores on the 1995 examination.

4.3.1 Internalising the Standards of Performance

In order to equate the examination paper from the initial year (1994) and the examination paper in the subsequent year (1995), judges were each given a package of

materials which, taken together, embodied the standards of performance set in 1994. This package comprised:

- the set of descriptor statements, developed as part of the standard-setting exercise in 1994, summarising the types of knowledge and skill from the course typically displayed by students at each of the performance standard levels classified as Excellent, Very Good, Good and Satisfactory;
- the student scripts used in 1994 to illustrate the standard of work typical of that presented by students who were at the borderlines. Each of these scripts showed how students who were at one of the cut-off scores set for the 1994 examinations had responded to each item in that examination. The judges were not told what score had been awarded to each of these borderline responses or what score had been awarded to particular items;
- the examination paper for 1994 showing the items that students were required to respond to in that year.

Judges in the two teams for each course were sent a copy of the above material, written instructions on the steps in the process, and a copy of the 1995 examination paper. They were also briefed to ensure they understood the procedures and were given the chance to ask questions to clarify issues. Once again, unfortunately, it was not possible to brief the judges in a team at the same time.

As had been the case when the standards were being established, the judges worked independently. They were advised to start with the descriptor statements and acquaint themselves with the descriptions of the knowledge and skills typically displayed by students at each of the standard levels defined for their course. They then refamiliarised

themselves with the examination paper from 1994 and the student scripts selected at the cut-off scores. In reviewing these scripts, the judges noted that the cut-off scores were based on the students' total performance on the examination, and that it would be expected that two students who received the same total score would probably receive different scores on each item. The judges also accepted that, while the type of student who would fit the Excellent category would generally be capable of achieving an almost perfect score on each item, this would generally not happen in practice. Such a student, particularly one deemed to be on the borderline between Excellent and Very Good, will still make a number of errors in an examination.

In this study, certain information was withheld from the judges which, if the procedure were being used operationally, might be provided at some stage in the process. The judges were not told the scores awarded to the students' scripts, or even the scores awarded to the individual items in those scripts. Furthermore, while knowing the maximum possible score for each item, the judges were not given the scoring key used.

This information was withheld in order to evaluate the effectiveness of the process, particularly the statistical feedback. Withholding this information eliminated any chance that the cut-off scores from the previous year could influence the judges. It was expected that, without this information, the judges would need to use a process of consideration, discussion and refinement in order to reach agreement on their cut-off scores. While it is possible that members of the original teams may have remembered the cut-off scores from the 1994 examination, there is no way that the second team would have known these values.

4.3.2 Setting the Initial Cut-off Scores

Once the judges felt they had become familiar with the performance standards set using the 1994 examination, they recorded the score for each item that borderline students, as defined by the material in the standards package, might be expected to receive on the 1995 examination paper. This step was performed after the examination was administered, but it could quite easily have been done before the students sat for the examination.

4.3.3 Reaching Consensus on Cut-off Scores

The members of each team of judges were brought together so that they could discuss and reach consensus on their decisions. At this stage, a table similar to Table 4.1, containing the item and total-examination cut-off scores for each standard level, was produced.

During the establishment of the initial cut-off scores, the values being proposed by the individual judges were examined to see that there were no marked differences between the judges within a team or between teams of judges working with the same course. Had significant differences arisen and remained after the discussion sessions, this would have indicated that the judges were applying different standards.

As had been the case in the initial year, the judges were given the materials they were to use and the instructions they were to follow soon after the 1995 examinations were held. The judges then developed their individual cut-off scores, and met to reach consensus on the initial set of cut-off scores. These activities happened some months before it was possible to present the relevant statistical data to the teams. The time lapse permitted a

change to be made in the Mathematics exercise and in the approach used in Biology, for reasons outlined below.

When there was disagreement in the new Mathematics team on the cut-off score for a particular item, attempts were made, in the first instance, to discuss the different views of the likely features of student responses. This discussion did not continue for long, however, nor did the judges go into as much detail as the members of the first team did when they had a disagreement to resolve. The second team was more likely to record the average, or even the most popular, of the scores rather than debate where students at each level might gain and lose marks. It was evident that the judges in this team were not applying the same amount of time and effort as the original Mathematics team when reaching a consensus position.

Another outcome became evident as the new Mathematics team worked through the process. One of the judges had been more thorough than his colleagues in his efforts to familiarise himself with the performance standards inherent in the materials. Having prepared a set of solutions to the items on the 1994 examination paper, he then scored the sample scripts according to a key he felt was appropriate. In this way, he developed a good understanding of the range of scores those students at the borderline of each standard level might be expected to achieve.

It became evident during the discussion process that some of the other judges in the new team tended to impose their own views of what the standards should be when determining their cut-off scores, rather than the standards encapsulated in the descriptor statements and the student scripts. This observation is supported by the fact that one of

the judges, who taught in a school where the typical student performance was generally below average, proposed cut-off scores which were lower than those proposed by other judges. As this team tended to average the individual cut-off scores more readily than the other team, the lower cut-off scores from this judge tended to depress the group scores. The result was that the second team came up with significantly lower cut-off scores for both the Good/Satisfactory and the Satisfactory/Unsatisfactory borderlines than the original team of judges.

Some members of this second Mathematics team were not available to continue with the exercise when the statistical data from the analysis of student performances were available. As a result, and also because of the concern about the integrity of the earlier process followed by this team, it was decided to add new members to the team, and work through the process from the beginning with the reconstituted team. It was also decided, as a further check, to create a third team for Mathematics, consisting of judges who had not been involved in the process before. The results of this action are shown in the next section.

The initial meetings of the Biology teams showed that, as had been the case in the initial year, the members of the teams adopted different approaches in arriving at their cut-off scores. Some focused more closely on the individual items and estimated the probability that the borderline students at each level would obtain a perfect score for the item. Other judges, after having a relatively quick look at the individual items, tended to look more holistically at the total section to which particular items belonged and arrive at a score, or even a likely range of scores, across the section.

From time to time in discussions, reference was made to the type of scoring key used to assess the student responses. This was usually in cases where a judge felt that a particular score was deserved but believed that, in practice, a different value would probably be awarded. The judges commented that, in their experience, it was not uncommon for the scoring key to be set so that a predetermined mean and standard deviation was obtained for each item. In some cases, the design of the item meant that students were required to do a disproportionate amount of work in order to obtain a particular score for an item.

Where there were differences between the opinions of the judges in Biology, they tended to reach agreement simply by a series of quick compromises. An observer commented that there was a tendency for the team members to "split the difference" too quickly, rather than discuss the item and the range of possible student responses to it. This view was illustrated by the fact that some judges did not record the cut-off scores for individual items.

The possibility that some teams may begin to set their own standards of performance in subsequent years, rather than those provided, either by disregarding or misinterpreting the materials given, or by only engaging in superficial discussion before adopting a compromise position, needs to be monitored. If there is any danger of this, the judges need to be reminded that they must focus closely on the standards inherent in the support materials. If there is no independent observer present to monitor the activities of the judges, then the team leaders need to be given extra responsibility in order to ensure the integrity of the process.

Given that so much time had elapsed and there was a need for the team members to refamiliarise themselves with the materials and the process ready for the next stage, it was decided to make a change to the way the Biology judges determined their individual cut-off scores. Instead of estimating the score borderline students would obtain on each item using whatever approach they wished, the team members were asked to estimate the probability that these borderline students would obtain the maximum possible score for each item. For the items that were scored polytomously, these probabilities were then multiplied by the maximum possible score for the item in order to obtain the expected score.

This approach seems to suit the type of items in the Biology examination and assist judges to focus more closely on the individual items, thus providing a basis for more detailed discussion. The probability approach worked well, particularly with the items in the core section of the paper. The approach suited the uniform nature of each of the various sections of the examination paper. The judges were able to compare the relative difficulties of the items within each section quite readily, and so establish their probabilities with minimal difficulty.

While this approach was different from that used with the 1994 examination, the focus of the standard-setting process in 1995 was still the descriptor statements, the statistical feedback and the student scripts chosen at the 1994 cut-off scores. Given the greater focus on the individual items and the more detailed discussions that followed, it was clear that for this course, at least, this approach was an improvement on that used in the previous year.

4.3.4 Using the Statistical Data

When a team agreed on an initial set of cut-off scores, the members were shown statistical data resulting from an analysis of the responses of a sample of students. This information was presented in the same manner as in the process of establishing the standards (see Figure 4.1 and Table 4.2). An explanation of the significance of the information and what it indicated was provided to the judges. After considering the data, all teams made some changes to their proposed item cut-off scores. The judges indicated that they found it to be a useful form of feedback for Mathematics, English and the compulsory items in Biology. To handle the optional items in the Biology examination, a different approach was followed.

For the 1995 examination, the section of the Biology paper examining the electives consisted of a number of items, each with a maximum possible score of 25. Students were required to attempt one item. As had been the case in the initial year, several judges in each team claimed not to have expertise in some of the elective areas of the course, and so felt that, without assistance, they could not accurately estimate likely cut-off scores for the items testing those areas. As a result, a different approach to that used for the core section was needed. One approach would have been to collect whatever estimates were provided for an item and then average these values. This was the method used in the initial year, which is discussed more fully later.

Another approach is to use the cut-off scores determined for the core sections of the examination and multiply these values by an appropriate constant to account for the maximum possible score for the examination. For example, if a cut-off score is set at 48 using the items in the core section which had a maximum possible score of 75, then 64

becomes the corresponding cut-off score for the total examination, which has a maximum possible score of 100. This approach is straightforward and simple to apply. It is recognised, however, it makes the assumption that the elective items are of the same difficulty and discriminate in the same way as the items in the core section. It is unlikely that this assumption will hold in many circumstances. Nevertheless, this approach will enable an initial estimate to be determined for elective items in cases where the expertise of the judges on certain areas of the course may not be thorough and widespread. Its use should be limited to such circumstances, as necessary. Once the initial estimates have been determined, emphasis is then placed on refining the item cut-off scores using the statistical data and the student scripts. In spite of these concerns, it was decided to use this approach when determining the initial cut-off scores in the second year.

While this issue did not create a major problem during the study, a consistent approach will be needed if the standard-setting strategy is used operationally, particularly with courses that have a much smaller compulsory section in their examination. The judges used would need, between them, to have the necessary expertise to make accurate predictions on the basis of all items in the examination.

In this regard, it would seem that the approach adopted in the initial year for Biology - that of averaging the scores agreed by the judges for each of the optional (or elective) items - is a better starting point. Each judge estimates the item cut-off scores for the compulsory items and for those optional items for which he/she claims to have sufficient expertise. The team then reaches agreement on a cut-off score for each item, as before. Then (as in the case of Biology, where students were required to respond to

one optional item), the average of the cut-off scores provided for each optional item is used in determining the initial cut-off scores for the examination. Using this approach, it is relatively simple to accommodate examinations where students are required to respond to more than one optional item, even where such items are from different sections of the examination paper.

The judges consider the statistical data as usual, with the item cut-off scores being refined as they believe necessary. As the ELM takes account of the relative difficulties of items, it is quite possible that students who have responded to different optional items may have different examination scores, but receive the same ability measure. This can mean that students who respond to particular optional items could be advantaged or disadvantaged by the establishment of particular cut-off scores, if they are simply expressed in terms of the examination scores awarded. Consideration needs to be given to any variability in the difficulties of the optional items.

One way to overcome this problem in Biology would be to use the procedure to determine the cut-off scores using the compulsory items and one of the elective items. If the ability measures of the students, in logits, were then adjusted using a suitable linear transformation, it would be possible to obtain cut-off scores to represent the borderlines between each performance standard, which resemble the original cut-off scores. In this way, students who have performed equally well would achieve the same score, irrespective of the optional items selected.

4.3.5 Reviewing Student Scripts

When each team had generated cut-off scores, the judges were given a sample of student scripts, each of which had obtained the cut-off scores the team was proposing. They were asked to verify that these borderline scripts demonstrated the standard of student performance corresponding to the standard level in which they would be placed. Once again, they were given the opportunity to vary their cut-off scores if they wished, based on their review of these scripts.

4.4 COMPARING STUDENT PERFORMANCE LEVELS

A major reason for employing the methodology used in this study is to enable comparisons to be made about the performance of groups of students who sat for the examinations in a course in different years.

In the initial year, the procedure establishes a performance scale. The cut-off scores relating to the borderline performances between the various standards that have been developed can be considered to be the calibrations on this ordinal scale. Underlying this rather crude scale is a more refined one, the units of which are the units of the examination. The scale is established in such a way, then, that the score students receive for the examination will locate them within one of the standards of performance.

In the subsequent year, the performances of the students who sit for the examination in that year are placed on the scale developed in the initial year. In this way, it is possible to compare the performances of groups of students who have sat for different examinations in the same course in different years.

4.5 SUMMARY

The procedure proposed in Chapter 3 was applied to the examinations in the courses of English, Mathematics and Biology to set performance standards and then use them in equating different examinations. A number of issues arose during the application of the procedure which were addressed at the time in a manner considered most suitable. In handling these issues the procedure proved to be sufficiently flexible and adaptable, without incurring any apparent loss of accuracy.

CHAPTER 5

RESULTS OF APPLYING THE PROCEDURE

5.1 INTRODUCTION

In this chapter, the results obtained when the procedure was applied to the 1994 and 1995 examinations in three courses which are part of the New South Wales Higher School Certificate (HSC) are presented.

5.2 THE INITIAL YEAR: ESTABLISHING THE STANDARDS

5.2.1 Mathematics

5.2.1.1 The Initial Cut-off Scores Agreed by the Judges

While discussing the cut-off scores they had established as individuals, the judges commented that they had experienced some initial difficulty in performing their task as a result of not being able to refer to the detailed key used to score the students' responses. This means, for instance, that for a sub-part of an item which had a maximum possible score of two and which required several steps, the judges did not know precisely what series of correct steps or working was required in order to receive a score of one if the final answer was incorrect. It was decided not to provide the scoring key so that the judges would be forced to analyse each item, including preparing solutions and developing what they believed to be an appropriate scoring key. This led to some initial uncertainty on the part of the judges, but they indicated that they were able to overcome it to a large extent by using their knowledge of the range of correct and incorrect steps which students would use when responding to each item. The discussion and refinement of their initial decisions during the implementation of the

procedure further helped to overcome any problems in this regard. The cut-off scores agreed by the judges following their discussions are shown in Table 5.1.

One result of having the judges initially focus on the parts and sub-parts of an item was that, particularly at the Satisfactory/Unsatisfactory borderline, the judges decided that students would be likely to receive a score of zero for parts of an item. Some judges reported an initial uneasiness in allocating a score of zero to these parts, given that the overall performance was judged to be satisfactory. This concern soon disappeared, however, when the scores were aggregated across the whole item.

TABLE 5.1

Initial Cut-off Scores for the 1994 Mathematics Examination

Item	*Excellent/ Very Good*	*Very Good/ Good*	*Good/ Satisfactory*	*Satisfactory/ Unsatisfactory*
1	12	12	10	8
2	12	11	9	6
3	12	11	9	7
4	12	11	8	6
5	12	11	8	5
6	12	9	7	4
7	12	10	7	4
8	12	10	6	2
9	12	9	6	4
10	11	9	4	2
Total	**119**	**103**	**74**	**48**

As can be seen from Table 5.1, the cut-off score for the Excellent category was extremely high. A score of 119 or better on the examination (with a maximum possible

score of 120) was achieved by only 15 students. There is little doubt that in determining these initial cut-off scores, the judges were equating excellent with perfect or near perfect.

This issue was raised, but the judges decided not to make any changes, believing that they would have the chance to review their decisions at a later time when given further information.

5.2.1.2 Refining the Initial Cut-off Scores after Considering the Statistical Data

Once the initial steps of the standard-setting procedure had been completed, a random sample of 500 students was selected from the population of 28 289. The response data were analysed using the ASCORE program, which produced threshold estimates for the ordered response categories of each item. An ability estimate was also produced for each student in the sample and for the score corresponding to each of the cut-off scores. These ability estimates are in the same metric as the item estimates and, as such, can be placed on the same continuum of performance. The information was represented graphically, as shown in Figure 4.1 in the previous chapter, and presented to the judges.

The judges discussed this information and made changes to their initial cut-off scores when they felt such changes were appropriate. Table 5.2 shows the initial cut-off scores (I) agreed to by the judges, the expected cut-off scores (E) generated by the statistical analysis, and the values (F) finally agreed by the judges after considering the statistical data.

TABLE 5.2

Initial (I), Expected (E) and Final (F) Cut-off Scores for the 1994 Mathematics Examination

Item	*Excellent/ Very Good*			*Very Good/ Good*			*Good/ Satisfactory*			*Satisfactory/ Unsatisfactory*		
	I	**E**	**F**	**I**	**E**	**F**	**I**	**E**	**F**	**I**	**E**	**F**
1	12	12	12	12	12	12	10	11	10	8	9	9
2	12	12	12	11	12	11	9	10	10	6	9	6
3	12	12	12	11	12	11	9	11	9	7	6	7
4	12	12	12	11	12	11	8	10	8	6	7	6
5	12	12	12	11	12	10	8	8	7	5	4	5
6	12	12	11	9	10	9	7	8	7	4	4	4
7	12	12	12	10	12	10	7	6	7	4	2	3
8	12	12	11	10	11	9	6	3	6	2	3	3
9	12	12	10	9	9	8	6	3	5	4	1	3
10	11	12	9	9	8	7	4	2	3	2	1	2
Total	**119**		**113**	**103**		**98**	**74**		**72**	**48**		**48**

In considering the cut-off score for the Excellent/Very Good borderline they had initially proposed, the judges came to the conclusion that they were expecting a standard of performance for the Excellent category which was too high. This was reinforced when they examined the item cut-off scores estimated by the model, which indicated that students with a score of 119 might be expected to obtain the maximum possible score on every item. While they still felt that it would be possible for a borderline

Excellent student to receive a perfect score on virtually any item, they agreed that it would be reasonable for such a student to lose some marks and yet, still be classed as Excellent.

The judges reconsidered each item and discussed where a student, who while matching their image of borderline Excellent/Very Good, might make errors. At the end of this discussion, they were prepared to accept that such a student could quite possibly drop one mark on items 6 and 8, two marks on item 9 and three marks on item 10. This led to a reduction in the cut-off score they had originally proposed from 119 to 113. Once they had established the new cut-off score for Excellent/Very Good, the judges then used a similar approach to review the item cut-off scores at each of the other borderlines.

Having reduced their initial cut-off score for excellent by six, the judges also reduced the borderline for Very Good/Good by five to what they considered to be a more realistic value. The borderline for Good/Satisfactory was also reduced slightly, but the Satisfactory/Unsatisfactory score remained the same although some of the item cut-off scores changed.

5.2.1.3 Consideration of Student Scripts

The judges considered the scripts of some of the students who achieved the "new" cut-off scores. This information proved to be particularly useful at the cut-off scores for Excellent/Very Good and Very Good/Good, where there had been considerable reduction from the scores that were originally proposed. By studying the scripts of several students who had scored 113, the judges were able to confirm that those scripts

were the work of students whose performances they would be prepared to class as Excellent. While these students obtained their scores of 113 by receiving different scores on the various items, the judges were satisfied that the students demonstrated, to a sufficient degree, an understanding of and a facility with Mathematics which could be classed as Excellent. The errors made by these students were generally due to carelessness and were not associated with a lack of understanding of the Mathematics being examined. These students also performed well on items which required insight and an understanding of mathematical concepts.

The judges looked at a number of student scripts in the range from 110 to 112 and noted that each showed some lack of understanding of important mathematical knowledge or constructs. Accepting that they had only looked at a small number of scripts and there is very little difference between scores of 112 and 113, the judges were, nevertheless, willing to use this further evidence to confirm the cut-off score of 113 for the Excellent/Very Good borderline.

A similar approach was used at the other cut-off scores of 98, 72 and 48. Sample student scripts were reviewed and the standard and nature of student responses discussed. In this way, the judges used the student scripts to confirm their earlier decisions.

5.2.2 English

5.2.2.1 The Initial Cut-off Scores Established by the Two Groups of Judges

The results in Table 5.3 show the initial cut-off scores established for the 1994 English examination. The members of Group A had been involved in marking Paper 1 (P1), while Group B had been involved in marking Paper 2 (P2) (see Section 4.2.3.1).

TABLE 5.3

Initial (I) Cut-off Scores for the 1994 English Examination Set by Group A (Gp A) and Group B (Gp B)

Item	*Excellent/ Very Good*		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
	Gp A	**Gp B**	**Gp A**	**Gp B**	**Gp A**	**Gp B**	**Gp A**	**Gp B**
P1 Q1(a) /4	4	4	3	4	3	3	2	2
Q1(b) /4	4	4	3	3	3	3	2	2
Q1(c) /3	3	3	3	3	3	2	2	2
Q1(d) /5	5	5	4	4	3	3	2	2
Q1(e) /4	4	4	3	3	2	2	1	2
Q2 /20	20	18	18	16	16	12	12	10
Q3 /20	18	18	16	16	14	12	12	10
P2 Q1 /20	18	18	16	16	14	12	12	10
Q2 /20	18	18	16	16	14	12	12	10
Q3 /20	18	18	16	16	14	12	12	10
Total	**112**	**110**	**98**	**97**	**86**	**73**	**69**	**60**

It can be seen from Table 5.3 that there was a reasonable degree of consistency between the two groups with regard to the Excellent/Very Good and Very Good/Good cut-off

scores. The other two borderlines, however, show less consistency, with the difference at the Good/Satisfactory borderline being 13 marks. This is largely because Group A established 14 and 12 (out of 20) as the respective cut-off scores for the Good/Satisfactory and Satisfactory/Unsatisfactory borderlines in the extended response items (Paper 1, Items 2 and 3 (P1 Q2,3) and Paper 2, Items 1, 2 and 3 (P2 Q1,2,3)). Group B, on the other hand, set 12 and 10 for these values. These items had initially been scored out of 10, meaning that Group A had chosen 7 and 6, whereas Group B had chosen 6 and 5. This difference, when aggregated across the five extended-response items, each of which had been reweighted to give a score out of 20, accounts for most of the difference in scores between Groups A and B at the Good/Satisfactory and Satisfactory/Unsatisfactory cut-off scores.

5.2.2.2 Refining the Initial Cut-off Scores after Considering the Statistical Data

A random sample of 500 students was selected for English from a course candidature of 30 226. The scores obtained by these students were analysed by using the ASCORE program.

Both groups of judges (Group A and Group B) were brought together as a single team for the next part of the exercise. They were each given a copy of the item and total examination cut-off scores proposed by each group, and the cut-off scores each judge had individually estimated. In addition, the judges were given an Item-Student scale similar to Figure 4.1, showing the threshold estimates for each item and the ability estimates associated with each examination cut-off score the groups had proposed. After some discussion about the suitability of the values they had originally established, the combined team decided to use the cut-off scores established by Group A as its initial

values. The judges were also shown the information presented in the I and E columns in Table 5.4 for each borderline performance. The values in Table 5.4 are the initial cut-off scores (I) proposed by Group A, the expected item cut-off scores (E) generated by the statistical analysis, and the cut-off scores (F) established by the combined team of judges after any adjustments were made as a result of reviewing the statistical data.

As had been the case with Mathematics, by considering the item cut-off scores they had set, along with the statistical feedback which emphasised the problem, the judges came to the view that they had set their cut-off scores too high.

TABLE 5.4

Initial (I), Expected (E) and Final (F) Cut-off Scores for the 1994 English Examination

Item	*Excellent/ Very Good*			*Very Good/ Good*			*Good/ Satisfactory*			*Satisfactory/ Unsatisfactory*		
	I	**E**	**F**	**I**	**E**	**F**	**I**	**E**	**F**	**I**	**E**	**F**
Q1(a)	4	4	4	3	4	3	3	2.5	2.5	2	2.5	2
Q1(b)	4	4	3.5	3	3	3	3	3	2.5	2	2.5	2
Q1(c)	3	3	3	3	3	3	3	2.5	2.5	2	2.5	2
Q1(d)	5	4	3.5	4	3	3	3	3	2.5	2	2.5	2.5
Q1(e)	4	4	3	3	3	2.5	2	2.5	2	1	2	2
Q2	20	20	18	18	16	16	16	14	14	12	12	10
Q3	18	20	16	16	16	14	14	14	12	12	10	9
P2 Q1	18	20	16	16	16	14	14	14	12	12	10	8
P2 Q2	18	20	16	16	16	14	14	14	12	12	10	8
P2 Q3	18	20	16	16	16	14	14	14	12	12	10	8
Total	**112**		**99**	**98**		**86.5**	**86**		**74**	**69**		**53.5**

After a discussion about the profile of skills and knowledge they expected of students at each borderline, the judges decided to use the cut-off score Team A had proposed for the Very Good/Good borderline as the new cut-off score for the Excellent/Very Good borderline. They made similar adjustments to the other cut-off scores. The judges then reconsidered each item and, discussed their expectations of students at each borderline in relation to the item and the expected value produced by the Extended Logistic Model (ELM). Where they felt a change to an item cut-off score was appropriate, they made the adjustment.

The judges reported later that they found the discussions particularly enriching. They indicated that, as a result of the process followed, they were better able to understand and appreciate the procedure they were using. In addition, they felt the procedure they had followed resulted in their becoming more committed to the joint decisions they had taken.

5.2.2.3 Consideration of Student Scripts

Before the judges finally settled on the cut-off score for each borderline performance, they were given a sample of scripts from students who had gained scores at or near the cut-off scores being proposed. Focusing on the cut-off score for Excellent/Very Good, the judges discussed whether the scripts of the students who scored a mark of 99, out of a maximum possible score of 120, demonstrated the standard they expected of a borderline Excellent/Very Good performance. In making this decision, the judges considered each student's performance in a holistic manner, as a student may perform unexpectedly well or poorly on particular items. In order to confirm their decision, the

judges then individually considered the scripts of several students whose scores were just below 99. They discussed their views of these scripts and, in each case, agreed that the performances of these students did not quite reach their expectations of an excellent performance.

The judges then considered some student scripts which had been awarded their revised cut-off scores for Very Good/Good, Good/Satisfactory and Satisfactory/Unsatisfactory, and thus settled on final cut-off scores for these other standard levels. For each borderline, the judges discussed their expectations of students at that point, and the features of each student's responses that did or did not demonstrate that the student had met these expectations. The judges reported that they found this stage of the exercise important in confirming the cut-off scores they had set.

5.2.3 Biology

5.2.3.1 The Initial Cut-off Scores Agreed by the Judges

The section cut-off scores agreed by the team of Biology judges are shown in Table 5.5. These sections, referred to as Part A, Part B and Part C, are, respectively, the multiple-choice items (maximum possible value 12), the items scored out of three (maximum value 18) and the items scored out of 5 (maximum value of 30). In addition, each of the items examining the elective topics of the course had a maximum possible score of 13, giving a possible total of 39 for the three such items students were required to attempt. The maximum possible score for the examination was 99.

TABLE 5.5

Initial (I) Cut-off Scores for the 1994 Biology Examination

Section	*Excellent/ Very Good*	*Very Good/ Good*	*Good/ Satisfactory*	*Satisfactory/ Unsatisfactory*
Part A	11	10	9	6
Part B	16	14	12	9.5
Part C	27	24	18	13
Elective	32	26.5	22	17
Total	**86**	**74.5**	**61**	**45.5**

As had been the case with the judges working in English and Mathematics, there was a tendency for the Biology judges initially to set too high an expectation of the standards students needed to reach to be placed in the excellent category. When considering the initial cut-off scores they had set, the judges made the point that, even though they accepted that excellent students would make careless errors, it was difficult to predict where such students would make these errors.

5.2.3.2 Refining the Initial Cut-off Scores after Considering the Statistical Data and Student Scripts

A sample of 500 students was selected from the population of 16 167 students who had undertaken the 1994 Biology examination. Although there were optional items in this examination testing the elective topics, it was considered that a sample size of 500 would be sufficient to enable the items in this examination to be calibrated. The student response data were analysed by using ASCORE and the results presented on an Item-

Student scale in a similar manner to that used for Mathematics and English. The judges were shown the information in the I and E columns for each borderline in Table 5.6.

TABLE 5.6

Initial (I), Expected (E) and Final (F) Cut-off Scores for the 1994 Biology Examination

Section	*Excellent/ Very Good*			*Very Good/ Good*			*Good/ Satisfactory*			*Satisfactory/ Unsatisfactory*		
	I	**E**	**F**	**I**	**E**	**F**	**I**	**E**	**F**	**I**	**E**	**F**
Part A	11	9	10.5	10	9	9	9	8	8.5	6	8	7
Part B	16	11	15	14	11	13.5	12	10	11	9.5	9	9
Part C	27	21	25	24	19	22	18	17	18	13	13	13
Elect.	32	24	31.5	26.5	24	26.5	22	24	23.5	17	22	16
Total	**86**		**82**	**74.5**		**71**	**61**		**61**	**45.5**		**45**

The statistical analysis indicated that there was little discrimination between the item or section scores over the range covered by the ability estimates corresponding to the cut-off scores proposed by the judges. This may have been due to a problem with the sample of data available or possibly due to the scoring key used by the raters and how it was applied. As a result, the judges took note of the analysis but did not vary their initial cut-off scores at this stage.

The judges were also given the examination scripts of a sample of students at and near the cut-off scores they had established. They found this to be useful and, after reviewing these scripts, made changes to a number of their initial cut-off scores.

The judges expressed the view that reviewing the student scripts helped them to clarify the type and extent of the knowledge and skills students at each standard level would

possess, and, as a result, determine how well they would be likely to perform on each item.

After considering the statistical data and the student scripts, the judges concluded that the cut-off score they had set for the Excellent/Very Good borderline was too high. This led them to lower not only that cut-off score, but the Very Good/Good cut-off score as well, using a similar approach to that used by the Mathematics and English judges.

5.2.4 Describing the Standards

Each team prepared statements describing the knowledge and skills typically displayed by students at each performance standard. All three teams commenced by identifying the key objectives of their course, then reflected upon the "performance profile" of students at each standard level that had emerged during their discussions. They then described the extent to which students at each standard level would have achieved those objectives. In developing these statements, the judges made use of the examination paper and the student scripts selected at each borderline. The teams found this material important in clarifying and documenting the extent of the knowledge and skills generally displayed by students who reach each performance standard. The descriptor statements prepared for each course are shown in the Appendices.

5.3 THE SUBSEQUENT YEAR: EQUATING THE EXAMINATIONS

This section reports on the results obtained when the teams of judges from each course familiarised themselves with the standards of student performance set using the 1994 examinations and then used the procedure to establish cut-off scores for the 1995

examinations which they believed were consistent with those standards. For each course, at least two teams were established and worked independently.

5.3.1 Internalising the Standards

At first, each judge in a team worked independently in order to become familiar with the standards of student performance encapsulated in the materials they were sent. By following a set of written procedures, the judges internalised the standards of performance they were to apply and became familiar with the 1995 examination paper. When they had done this, they recorded the score on each item in the 1995 examination that they expected a borderline student at each standard level would receive.

5.3.1.1 Mathematics

The judges from the original team for Mathematics, who had set the standards in the initial year, had no difficulty in using the materials to set initial cut-off scores for the 1995 examination. While there were minor differences between the cut-off scores proposed, each judge seemed to have applied similar "profiles" of student performance when determining his/her cut-off scores.

The members of the newly established team in Mathematics, who had not been involved in developing the descriptor statements, expressed the view that the statements were too general and could be applied to a number of different stages of schooling. They stated, nevertheless, that the sample of student scripts chosen at the cut-off scores was helpful in forming an understanding of the standards to be applied. While it is possible that the process could have been strengthened by some refinement of the descriptors, it is clear that these judges had missed an important point. The descriptor statements, like the

examination paper or the sample scripts, were not designed to stand on their own. Rather, all three components are intended to form a comprehensive package, with each part providing different information to clarify the standards of performance to be applied.

This, and further evidence which emerged during the next stage of the process, (see Section 4.3.3) indicated that some of the judges in this team had not consistently applied the standards defined in the materials they were given when establishing their cut-off scores. This outcome clearly emphasises the need to ensure that the briefing of judges is thorough, and, where possible, that the judges are trained and briefed as a group. Action was taken to eliminate this problem when the statistical data were available.

5.3.1.2 English

The judges from both the original and the newly created team for English reported that they were able to identify the standards of student performance they were to apply quite readily from the materials provided. Again, while there were some differences between the cut-off scores proposed by different judges, each felt that they were able to follow the procedure without any difficulty.

5.3.1.3 Biology

The judges in the newly formed team expressed the view that the descriptor statements were effective in defining what typical students at each of the standard levels created for Biology know and can do in this domain.

The judges noted that the student scripts provided at each borderline sometimes showed a range of standards in the responses for certain items. This was particularly the case

with those scripts at the Good/Satisfactory and Satisfactory/Unsatisfactory borderlines, where the standard of response to items was often inconsistent. The judges stated that in such cases, it was difficult to use the student scripts to establish with any certainty the score which borderline students would be most likely to receive for an item. Notwithstanding this problem, they indicated that the student scripts had helped them to understand the standards they were to apply.

5.3.2 Mathematics: Establishing and Refining the Cut-off Scores

5.3.2.1 The Initial Cut-off Scores

Having independently produced a set of cut-off scores, the members of a team were brought together. Table 5.7 shows the cut-off scores agreed by each team. Team 1 was

TABLE 5.7

Initial Cut-off Scores Proposed for Mathematics for 1995 by Team 1 (T1), Team 2 (T2) and Team 3 (T3)

	Excellent/ Very Good			*Very Good/ Good*			*Good/ Satisfactory*			*Satisfactory/ Unsatisfactory*		
Item	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
1	12	12	12	12	12	12	9	11	10	8	8	8
2	12	12	12	11	11	11	9	9	10	7	8	8
3	12	12	12	11	11	11	9	9	10	7	7	8
4	12	12	12	11	11	11	9	8	9	6	6	5
5	11	11	11	10	10	10	8	9	8	5	6	6
6	11	12	12	9	10	10	7	8	6	4	4	4
7	12	12	12	11	9	11	8	7	9	5	4	5
8	11	10	10	10	9	8	8	6	6	5	4	3
9	10	11	9	9	8	6	5	6	4	1	3	2
10	9	9	8	7	6	5	4	4	3	2	2	2
Total	**112**	**113**	**111**	**101**	**97**	**95**	**76**	**77**	**75**	**50**	**53**	**51**

the group that had set the standards of performance in Mathematics for the 1994 examination and prepared the descriptor statements. Team 2 was a reconstituted team, which contained two of the judges who had earlier met as part of the second team for Mathematics, and three new members. Team 3 was an entirely new group of judges who had not been involved in the process previously.

During the process of negotiation it became evident that some judges, especially those in the new teams, tended to establish scores significantly higher or lower than the other members of their team. From the results obtained and the comments made, this appears to be related to the different schools in which they were teaching. For example, those who taught at schools typically consisting of more able students tended to set higher item cut-off scores than those set by judges teaching at schools where students exhibited a much wider range of ability. These differences were overcome, to a large extent, during the discussion process. As the teams worked through each item, discussing where they felt students at different levels would score marks and the types of errors they would make, a good degree of agreement was obtained. Throughout this process the judges referred to the descriptor statements and assisted each other to focus on the notion of borderline students.

From Table 5.7, it can be seen that the initial cut-off scores for the 1995 HSC examination established by the teams are generally quite close. The greatest difference is at the cut-off for Very Good/Good where the original team set a value higher than the two new teams. An examination of the individual item cut-off scores set by the teams also shows good agreement. The greatest variation between the scores proposed by the teams comes for those items in the latter half of the examination. This is probably

related to the fact that traditionally the Mathematics examination is deliberately set so that the early items are easier than the later ones. This may make it harder for judges to determine how students of different abilities will perform on the more difficult items.

5.3.2.2 Reviewing the Statistical Data and Refining the Cut-off Scores

After each team had settled on a set of cut-off scores, the team members were provided with statistical data, presented in an Item-Student scale, similar to those used in the previous year. The information was also presented as shown in Table 5.8.

TABLE 5.8
Initial (I) and Expected (E) Cut-off Scores for Each Item Proposed by Team 1

	Excellent/ Very Good		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
Item	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**
1	12	12	12	12	9	12	8	8
2	12	12	11	12	9	12	7	8
3	12	12	11	12	9	11	7	8
4	12	12	11	12	9	11	6	6
5	11	12	10	11	8	9	5	5
6	11	12	9	12	7	9	4	2
7	12	12	11	10	8	6	5	4
8	11	12	10	11	8	7	5	3
9	10	12	9	9	5	4	1	0
10	9	7	7	5	4	3	2	1
Total	**112**		**101**		**76**		**50**	

Each team compared the initial item cut-off scores they had set and the expected scores produced by the Extended Logistic Model. The revised cut-off scores reached by each team are shown in Table 5.9.

Table 5.9 shows that the cut-off scores for each standard level established independently by the three teams are relatively close. The biggest difference is at the cut-off for Very Good/Good, where the original team (T1) has set a score higher than the other two teams, who are in close agreement. The two new teams are, in fact, in close agreement for all cut-off scores.

TABLE 5.9

Revised Cut-off Scores for Mathematics Established by Team 1 (T1), Team 2 (T2) and Team 3 (T3) after the Review of the Statistical Data

	Excellent/ Very Good			*Very Good/ Good*			*Good/ Satisfactory*			*Satisfactory/ Unsatisfactory*		
Item	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
1	12	12	12	12	12	12	10	11	10	8	8	8
2	12	12	12	12	11	12	10	10	11	7	8	8
3	12	12	12	12	11	12	10	10	11	7	8	8
4	12	12	12	12	11	11	10	10	9	6	6	5
5	12	11	12	10	11	11	9	9	9	5	6	6
6	12	12	12	11	11	10	8	8	7	4	4	4
7	11	12	12	10	9	11	7	7	7	4	4	4
8	11	10	11	10	9	9	7	6	6	4	4	3
9	11	11	10	9	8	7	4	4	4	1	1	1
10	8	9	8	6	6	5	3	3	3	1	1	2
Total	**113**	**113**	**113**	**104**	**99**	**100**	**78**	**78**	**77**	**47**	**50**	**49**

5.3.3 English: Establishing and Refining the Cut-off Scores

5.3.3.1 The Initial Cut-off Scores

In the case of English, judges in both the original team (Team 1) and the new team (Team 2) reported that they had been comfortable with the process they had undertaken. This is because the scoring key employed in this English course is based on a general notion of standards. Judges familiar with this type of approach find it relatively easy to come to terms with the performance standards expressed in the descriptors and the samples of student work. Members of Team 1, having used the process to set the standards using the 1994 examination, reported that they were able to "visualise" from the descriptor statements the range of responses which students would produce to the items in the 1995 examination, and the scores which they believed would be awarded to those responses.

The judges in Team 2 went about the process in the following way. Having read and become familiar with the descriptor statements, they studied the items on the 1994 examination paper and then scored the sample of student scripts they had been given. In this way, they felt, they were able to get a good idea of the scores awarded to borderline students in each item and across the total examination in 1994. They then considered the items in the 1995 examination paper and estimated what scores students who were on the borderlines in 1994 would score on each item on the 1995 examination. These values are shown in Table 5.10.

TABLE 5.10

Initial (I) Cut-off Scores Proposed for 1995 by Each of the English Teams

	Excellent/ Very Good		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
Item	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
P1Q1(a)	3	4	3	3	2	2.5	1	2
Q1(b)	4	3	3	3	3	2	2	1
Q1(ci)	2	2	2	1.5	1	1	1	1
Q1(cii)	3	3.5	3	3	2	2.5	1	2
Q1(ciii)	5	5	4	4	3	3	2	2
Q2	18	17	16	15	15	13	10	10
Q3	16	17	14	15	12	13	10	10
P2 Q1	17	17	15	15	13	12	9	9
P2 Q2	17	17	15	15	13	13	9	10
P2 Q3	17	16	15	14	13	12	9	8
Total	**102**	**101.5**	**90**	**88.5**	**77**	**74**	**54**	**55**

The judges discussed cases where there were differences between their individual item cut-off scores. These discussions continued until agreement on a cut-off score was reached. There was, however, generally quite close agreement among the judges in each team at the beginning, so not much adjustment was required.

At the end of this stage of the process, the judges in each of the teams were satisfied with the item cut-off scores and the total examination cut-off scores established by their team. Table 5.10 shows the close agreement between the initial cut-off scores produced by the two teams. The individual item cut-off scores are also generally very close.

5.3.3.2 Reviewing the Statistical Data and Refining the Cut-off Scores

Data from analysing the performances of a sample of 500 English students using the ELM were presented to each of the teams. The revised cut-off scores are shown in Table 5.11.

The refinement of the original cut-off scores proposed by the two teams has resulted in close agreement being reached. The biggest difference between the teams is at the cut-off point between the Good and Satisfactory categories. Given this level of agreement, it can be asserted that the different teams of judges were able to determine the standards of performance established for the 1994 examination from the materials provided, and apply these standards consistently to the 1995 examination.

TABLE 5.11

Revised Cut-off scores for English Established by Team 1 (T1) and Team 2 (T2) after the Review of the Statistical Data

	Excellent/ Very Good		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
Item	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
P1Q1(a)	3	4	3	3	2	2.5	1	2
Q1(b)	4	3	3	3	3	2	2	1
Q1(ci)	2	2	2	1.5	1	1	1	1
Q1(cii)	3	3	3	2.5	2	2	1	2
Q1(ciii)	5	4	4	3.5	3	3	2	2
Q2	16	16	14	15	12	12	10	10
Q3	16	16	14	15	12	12	9	8
P2 Q1	17	18	15	15	13	12	9	8
Q2	17	18	15	15	13	12	9	9
Q3	17	18	15	14	13	12	9	8
Total	**100**	**100**	**88**	**87.5**	**74**	**70.5**	**53**	**51**

5.3.4 Biology: Establishing and Refining the Cut-off Scores

5.3.4.1 The Initial Cut-off Scores

After the judges for Biology had produced their individual cut-off scores, each team was brought together so that members could work towards obtaining consensus. The members of both the original team (Team 1) and the new team (Team 2) tended to hurry the discussion process and were quick to compromise on the cut-off scores for each section of the examination, without really addressing the individual items and how students at different standard levels would be likely to respond to them. In addition, as had been the case for the 1994 examination paper, a number of judges were reluctant to propose a cut-off score for items examining those elective (or optional) areas of the course where they felt they did not have sufficient expertise or teaching experience. The initial values proposed by the two teams after their meetings are shown in Table 5.12.

TABLE 5.12

Initial (I) Cut-off Scores Proposed for 1995 by Each of the Biology Teams

	Excellent/ Very Good		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
Section	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
Part A	12	13	11	12	9	9	6	7
Part B	26	26	23	22	18	16	13	12
Part C	24	26	20	22	16	18	12	13
Core	**62**	**65**	**54**	**56**	**43**	**43**	**31**	**32**
Electives	20	21	17	17	15	13	11	11
Total	**82**	**86**	**71**	**73**	**58**	**56**	**42**	**43**

Team 1 only addressed some of the items examining the elective topics (referred to as elective items), whereas Team 2 did consider all of these items and propose a cut-off score for each one. In order to adopt a consistent approach between the two teams in establishing a cut-off score for these optional items, it was decided that for each team, the minimum score agreed by the judges for any elective item would be used as the cut-off score for all such items. This issue, and the implications of this decision, are also discussed in Chapter 4 and Chapter 8.

In all except the cut-off score for Good/Satisfactory, the new team (Team 2) set a cut-off score higher than that set by the original team (Team 1). Nevertheless, the cut-off scores are quite close.

After considering the way the judges had tended to rush the process, rather than undertake the more careful and thorough discussions characteristic of the Mathematics and English teams, it was decided to trial a modification to the way the Biology judges establish their individual cut-off scores.

This change was aided by the fact that there was a delay of some months from the time they had established the cut-off scores shown in Table 5.12, and when the results of the statistical analysis were available. Due to the time gap the judges needed to spend time becoming reacquanted with the standards encapsulated in the materials they were given, hence, it was feasible to require the judges to apply the procedure from the beginning. The approach, which was explained in Section 4.4.2, required the judges to specify the probability that students at the borderline of each standard level would answer the item correctly. This change was intended to make the judges look more closely at the

individual items and make a decision about how well students at each standard level would answer the item, rather than simply arriving at an overall total score for each section of the examination without much in-depth analysis. Given the problems of assigning cut-off scores to the elective items, it was decided to limit this approach to the core sections of the examination. The results of using this new approach for the core sections of the 1995 examination compared with the scores arrived at initially are shown in Tables 5.13 and 5.14.

TABLE 5.13

Section Cut-off Scores Established by Team 1 for the 1995 Biology Core Paper Comparing the Probability Approach and the Original Approach

	Probability Approach				*Original Approach*			
	Ex	**VG**	**G**	**S**	**Ex**	**VG**	**G**	**S**
15 mc items	13	11.4	9.7	7.5	12	11	9	6
10 x 3 mark items	26.2	23	19.1	14.8	26	23	18	13
6 x 5 mark items	25.1	21.6	18	13.3	24	20	16	12
Totals	**64.3**	**56**	**46.8**	**35.6**	**62**	**54**	**43**	**31**

TABLE 5.14

Section Cut-off Scores Established by Team 2 for the 1995 Biology Core Paper Comparing the Probability Approach and the Original Approach

	Probability Approach				*Original Approach*			
	Ex	**VG**	**G**	**S**	**Ex**	**VG**	**G**	**S**
15 mc items	13.5	11.2	8.5	5.9	13	12	9	7
10 x 3 mark items	26.4	22.5	16.5	10.8	26	22	16	12
6 x 5 mark items	25.5	21.0	15	9.5	26	22	18	13
Totals	**65.4**	**54.7**	**40.0**	**26.2**	**65**	**56**	**43**	**32**

Judges from both teams were positive about this new approach. It was also observed that their discussions about the items and expectations of students at each standard level were more thorough than in the earlier exercise. For this reason, it was decided to proceed with the data obtained from this method for the rest of the exercise. It was also decided to focus on the core section of the examination initially, as the issue of the elective items was still to be resolved. The core sections had a maximum possible score of 75.

The cut-off scores for the sum of the multiple choice items (Section A) and the cut-off scores for each item in Sections B and C are shown in Table 5.15. The values reported show that, while there is slightly improved agreement between the two teams for the Excellent/Very Good and Very Good/Good cut-off scores using the probability approach, the differences between the values nominated for the Good/ Satisfactory and Satisfactory/ Unsatisfactory borderlines have increased markedly.

A couple of factors may have been responsible for these differences. First, given the differences in the content and structure of the examinations between the two years, the 1994 examination paper and the student scripts from that year probably did not provide as clear an image of standards to be applied to the 1995 examination as they might otherwise have. The differences are more likely to have occurred at the Good/Satisfactory and Satisfactory/Unsatisfactory points as at the other borderlines students' performances tend to be more consistent and, consequently, the materials from 1994 may have provided better guidance. Secondly, only four of the judges originally assigned to Team 1 and three of the judges assigned to Team 2 were available when each team was required to meet. The numbers of judges in these teams are less than optimal. Where the teams are small it is more difficult to obtain a range of perspectives

on the items and how students are likely to perform. It is also more likely that a dominant personality may influence the views of the other judges.

TABLE 5.15

The Initial (I) Cut-off Scores Set for 1995 by Biology Team 1 (T1) and Team 2 (T2) using the Probability Approach

	Excellent/ Very Good		*Very Good/ Good*		*Good/ Satisfactory*		*Satisfactory/ Unsatisfactory*	
Item	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
mc	13	13.5	11.4	11.2	9.7	8.5	7.5	5.9
Q16	2.7	2.7	2.4	2.1	2.0	1.5	1.5	0.9
Q17	2.7	2.7	2.4	2.1	2.0	1.5	1.5	0.9
Q18	2.5	2.7	2.1	2.4	1.8	1.8	1.4	1.5
Q19	2.7	2.4	2.3	2.1	1.8	1.5	1.4	1.2
Q20	2.4	2.4	2.0	2.4	1.6	1.2	1.2	0.6
Q21	2.7	2.4	2.4	1.8	2.0	1.2	1.5	0.6
Q22	2.5	2.7	2.3	2.4	1.9	1.8	1.4	0.9
Q23	2.7	3.0	2.4	2.4	2.1	2.1	1.7	1.5
Q24	2.6	2.7	2.3	2.4	1.8	1.8	1.5	1.2
Q25	2.7	2.7	2.4	2.4	2.1	2.1	1	1.5
Q26	4.0	4.0	3.2	3.5	2.5	2.5	3	1.5
Q27	4.5	4.5	3.9	4.0	3.1	3.0	2	2.0
Q28	4.4	4.5	3.8	3.5	3.2	2.5	3.5	2.0
Q29	4.0	4.0	3.5	3.0	3.0	2.0	0.5	1.0
Q30	4.2	4.5	3.7	3.5	3.2	2.5	2	1.5
Q31	4.0	4.0	3.5	3.5	3.0	2.5	2.5	1.5
Total	**64.3**	**65.4**	**56**	**54.7**	**46.8**	**40.0**	**35.6**	**26.2**

5.3.4.2 Reviewing the Statistical Data and Refining the Cut-off Scores

The data from the analysis of the performances of a sample of students were presented to the judges in the same manner as for Mathematics and English.

The total examination cut-off scores shown in Table 5.16 were calculated by expressing the core section as a score out of 100. As discussed previously, using the core sections of the paper to estimate performance in the elective items in this manner is only one way of establishing the total examination cut-off score. For example, another method would be to use the minimum cut-off score proposed by the judges for any elective item, as was used for the 1994 examination. This issue is discussed in a later chapter.

TABLE 5.16

Revised Cut-off Scores for Biology Team 1 (T1) and Team 2 (T2) After the Review of the Statistical Data

	Excellent/ Very Good		*Very Good/ Good*		*Good Satisfactory*		*Satisfactory/ Unsatisfactory*	
Item	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**
mc	13.0	13.5	11.4	11.2	9.7	10.0	7.5	8.0
/3	25.4	26.4	21.4	22.5	17.8	17.0	14.5	13.0
/5	25.0	25.5	20.1	21	16.8	17.0	12.8	13.0
Core /75	**63.4**	**65.4**	**52.9**	**54.7**	**44.3**	**44.0**	**34.8**	**34.0**
Total/100	**85**	**87**	**71**	**73**	**59**	**59**	**46**	**45**

Tables 5.15 and 5.16 show that, in spite of doubts already discussed about the Biology data, the judges were prepared to make some changes to their initial cut-off scores based on a consideration of the statistical data. For example, Table 5.15 shows that the initial

cut-off scores for the Core Sections at the Satisfactory/ Unsatisfactory borderline are 35.6 for Team 1 and 26.2 for Team 2. After the teams reviewed the statistical data these values became 46 and 45. While their initial cut-off scores for the Good/Satisfactory and Satisfactory/Unsatisfactory borderlines were quite discrepant, after this review process there is close agreement between the cut-off scores set by the two teams. Indeed, all four cut-off scores proposed by the teams are relatively close.

5.3.5 Review of Student Scripts

Each of the teams for Mathematics, English and Biology next considered a sample of student scripts awarded the cut-off scores it had chosen. This was done with the purpose of ensuring that the standards of performance displayed in the scripts matched the appropriate descriptor statement, and furthermore, that these scripts represented a minimally acceptable performance at that level.

The teams stated that, taken holistically, the student scripts they were given to look at clearly fitted the descriptors for the standard level in which they were being placed. They also confirmed that the descriptors and the sample scripts divided performance in the course into the intended categories.

Members of some teams raised the issue of whether just looking at the total score is sufficient to determine whether a student's performance is satisfactory or not. They questioned whether it was important to consider whether a student had a consistent level of performance in items taken from different sections of the course. There was a view expressed that it may be important to focus on what knowledge and skills the items were actually designed to test, and set expectations of student performance in relation to

these. The judges in these teams discussed this issue and looked further at the samples of student responses. They agreed that a compromise approach was the most suitable outcome. That is, if a student performed below expectations on items testing some aspects of the course, this could be off-set by better than expected performances on other areas of the course. To attempt to do otherwise, they observed, would be too difficult in an examination where the balance between the sections of the course tested could vary from one year to the next.

This position was borne out by the scripts which showed, particularly for high-performing students, that an unexpectedly poor score on an item was often the result of careless errors, rather than a lack of understanding of the knowledge or processes being tested. To classify a student's performance below a level which, based on their total score, they would otherwise have achieved, simply because a careless error produced a score on one item that was below some particular value, would seem to be inappropriate in most curriculum-based examinations. Hence, the view taken was that a conjunctive approach may be appropriate in a situation where the assessment is for licensing purposes where some acceptable basic level of performance is required on certain essential tasks. In a summative examination of a general education course, however, it is more appropriate to take a holistic view, with students given the opportunity to compensate for a relatively poor response to some items by a relatively good response to other items.

5.3.6 A Final Check

In this study, a final check was made on the accuracy with which the judges were able to apply the same performance standards to the 1994 and 1995 examination papers. The

judges were asked to compare a sample of the scripts awarded the cut-off scores set for the 1994 examination with scripts awarded the cut-off scores they had set for the 1995 examination. The purpose of this step was to see whether the students who had produced the scripts in the two different years had demonstrated the same levels of knowledge and skill in the course, even though they had sat for different examination papers.

5.3.6.1 Mathematics

Each judge took a pair of student scripts at the same borderline, one from the 1994 examination and the other from the 1995 examination. They then compared these scripts item by item, looking to see how the two students had performed on their respective tasks. In doing this, the judges noted that often the students were not consistent in how they had performed on the items in the examination. In addition, the corresponding items from the two examinations (ie the two first items, the two second items, and so on) generally did not test the same topics from the course. This meant that in some cases, judges may have been comparing the performance of one student in an item testing the calculus with the performance of the other student in an item testing geometry. These factors tended to make it difficult to draw conclusions about the relative strengths and weaknesses of the two performances.

The teams noted that an alternative approach would have been to look for those items which tested similar skills and knowledge in the two examination papers and to use the students' relative performances on the two tasks to make the comparisons. Using this approach, however, is still not without some problems in that there is no guarantee, for

example, that the items testing the calculus in one year were of the same difficulty as the items testing the calculus in the other year.

In making their decisions about the equivalence of responses at the same borderline in the two years, the judges took into account the type of mistakes the students made, their skills in using mathematical notation and how well they set out their solutions.

While there were some minor differences within and between teams about the equivalence in standard of the responses from the different years, the overall conclusion was that, based on the information at their disposal, the samples of scripts selected at each borderline demonstrated a similar level of mathematical achievement.

5.3.6.2 English

In comparing the student scripts awarded the cut-off scores for the 1994 examination with those scripts awarded the cut-off scores chosen for the 1995 examination, the English judges were looking to see whether the performances in each year were equivalent in terms of the attributes being measured by the examination and summarised by the descriptor statements. After some discussion, the judges expressed the opinion that the sample scripts from the two years did demonstrate the same levels of performance across the whole examination. They concluded that, while there were differences in the level of performance demonstrated on corresponding items, when the total-paper performances were considered the scripts at the cut-off scores in 1994 represented a comparable level of achievement to those scripts selected at the cut-off scores in 1995. The review of the student responses confirmed in their minds that the

cut-off scores they had chosen for the 1995 examination were appropriate and no further adjustment was required.

5.3.6.3 Biology

This final step proved to be particularly difficult for the Biology teams. They discussed how best to perform this task and decided that, given the major change in course content and examination structure between the two years, it was pointless trying to compare the student scripts in total. In an attempt to overcome this problem, they nominated items in the two examination papers that addressed those topics that were common to the 1994 course and the 1995 course. The items they nominated are shown in Table 5.17.

These common items represented only a relatively small part of the whole examination in each year (16%), and the items testing a particular topic in the two years may have had different maximum possible scores. For this approach to be of use, the judges would have needed to compare student scripts which had received the item cut-off scores set for items 20, 23, 15 and 17 from the 1994 examination with scripts which had received the cut-off scores set for items 28, 27, 17 and 16 from the 1995 examination.

TABLE 5.17

Pairs of Items from the 1994 and 1995 Examinations Testing the Same Content Area of the Biology Courses

1994	*1995*
Q20	Q28
Q23	Q27
Q15	Q17
Q17	Q16

After considering this issue, the judges decided that any judgments about the relative performances of students based solely on these items would, at best, simply give them some added confidence that they may also have been able to extend these judgments across the full examinations.

The judges stated that they were satisfied with the procedure they had followed to set the cut-off scores, and the values that had arisen. They indicated that it was unlikely they would wish to change their cut-off scores as a result of the proposed comparison of scripts. Consequently, it was decided not to undertake this exercise.

5.4 COMPARING STUDENT PERFORMANCE LEVELS ACROSS THE TWO YEARS

This study set out to investigate the feasibility of using teams of judges to equate examination papers used in different years. In doing this, the cut-off scores set by the teams for the 1994 and 1995 examinations, will be compared.

It is not expected that the cut-off scores in the two years will be identical. Variations in the relative difficulties of the items in the examination papers will lead to differences in the cut-off scores from year to year. The point is that whatever cut-off scores the judges set, they indicate equivalent standards of performance in the two years.

In Tables 5.18-5.20, the cut-off scores established in the initial year of the study (1994) are shown. Similar figures are also shown for the subsequent year of the study (1995). The cumulative percentages of the total candidature that performed at or above each standard level are provided. These percentages occur as a result of the equating process.

There is no reason to assume that the percentages of students at or above any boundary should be the same from one year to the next. Given the size of the total course candidature and the consistent nature of examinations, it could be expected that generally there would be relatively similar proportions of students in each standard-level from year to year. An interpretation of these figures and an examination of the factors that contributed to these results are provided in the next chapter.

5.4.1 Mathematics

Table 5.18 shows that the cut-off scores for the Excellent category set by all teams are identical for 1994 and 1995. This signifies the judges' view that a score of 113 represents an equivalent standard of performance in the two years, even though different examination papers were used. There is a slight difference, however, in the proportions of the candidatures judged as meeting the standard for Excellent in the two years. A marginally higher proportion of students (1.7%, up from 0.9%) in the Excellent category in 1995. The Good/Satisfactory cut-off score is lower in 1994 than the values

TABLE 5.18

Final (F) Cut-off Scores Proposed by Each Team and the Associated Cumulative Percentages of Students in Mathematics in 1994 and 1995

	1994		*1995 Team 1*		*1995 Team 2*		*1995 Team 3*	
	Score	**Cum %**	**Score**	**Cum %**	**Score**	**Cum %**	**Score**	**Cum %**
Excellent	113	0.9	113	1.7	113	1.7	113	1.7
Very Good	98	12.5	104	10.8	99	17.6	100	16.2
Good	72	42.8	78	46.1	78	46.1	77	47.2
Satisfactory	48	68.8	47	79.0	50	76.5	49	77.4

set by each of the teams in 1995 by 5 or 6 marks, signifying that the 1995 examinations is relatively more difficult at this limit. The proportion of students reaching the standards Good and above in the two years is within 3 to 4%.

5.4.2 English

Table 5.19 shows generally close agreement between the cut-off scores set in 1994 and those set by the teams in 1995. The cumulative percentages of the candidatures at each standard level are also similar between the two years. In 1994, however, approximately 88% of the candidature were deemed Satisfactory or above. In 1995, the cut-off scores set by the two teams placed this figure at around 81% to 84%.

TABLE 5.19

Final (F) Cut-off Scores Proposed by Each Team and the Associated Cumulative Percentages of Students in English in 1994 and 1995

	Year 1		*Year 2 Team 1*		*Year 2 Team 2*	
	Score	**Cum %**	**Score**	**Cum %**	**Score**	**Cum %**
Excellent	99	1.0	100	1.6	100	1.6
Very Good	87	9.2	88	9.6	87.5	9.6
Good	74	37.5	74	33.6	70.5	40.8
Satisfactory	55	87.8	53	80.7	51	83.6

5.4.3 Biology

The values of the cut-off scores established for each borderline for Biology in 1994 and 1995 (shown in Table 5.20) are relatively close. The proportion of the candidature above each borderline, however, shows considerable variation. For example, the cut-off score for the Very Good level was approximately 74 in 1994 compared to 71 and 73 in

1995. In 1994, however, approximately 11% of students were in the combined Excellent and Very Good categories, compared to 31% using the 1995 Team 1 cut-off score and 27% using the Team 2 cut-off score in 1995. This issue is discussed in Chapter 6.

TABLE 5.20

Final Cut-off Scores Proposed by Each Team 1 (T1) and Team 2 (T2) and the Proportions of Students in Each Standard Level in Biology

	Year 1			*Year 2 T1*			*Year 2 T2*		
	Score /75[1]	**Score /100[2]**	**Cum %**	**Score /75[1]**	**Score /100[2]**	**Cum %**	**Score /75[1]**	**Score /100[2]**	**Cum %**
Excellent	63.1	84.1	0.7	63.4	84.5	5.5	65.4	87.2	3.4
Very Good	55.6	74.1	10.8	52.9	70.5	30.6	54.7	72.9	26.5
Good	46.9	62.5	35.9	44.3	59.1	55.1	44	58.7	55.1
Satisfactory	36.2	48.3	65.9	34.8	46.4	75.7	34	45.3	77.0

[1] The maximum possible score for the compulsory core items in 1994 was 60. This has been converted to a score out of 75 to put it on the same scale as the core items from the 1995 examination.

[2] The scores for the core items have been converted to a score out of 100 to give a score comparable to the maximum possible score for the whole examination. In taking this step, the optional items have been ignored due to concerns about the quality of the advice some of the judges were able to provide concerning these items. This issue is discussed elsewhere.

5.5 SUMMARY

The results of applying the procedure to set and compare standards of performance in the NSW Higher School Certificate (HSC) examinations in three courses are presented in this chapter. These results show that the procedure appears to be quite effective in the first of these tasks, namely establishing and describing standards of performance on the types of examination encountered in the HSC program.

The results also indicate that a team of judges can use the procedure to apply the same standards of performance to equate examinations of the same course across different years. The level of agreement between the cut-off scores set by the teams working with each course shows that different teams can make similar interpretations of the standards to be applied when determining cut-off scores on subsequent examinations.

In addition to discussing the evidence to support these claims, the following chapter discusses the figures in the tables in Section 5.4, with a view to determining the implications of these results.

CHAPTER 6

DISCUSSION OF RESULTS

6.1 INTRODUCTION

The previous chapter showed the results obtained when each team of judges applied the procedure, first to establish cut-off scores relating to standards of performance in an examination, and then to equate that examination and subsequent ones by determining cut-off scores for the examination relating to those same standards of performance. Once the examinations have been equated it is possible to make comparisons between the performances of groups of students who have sat for the different examinations.

This chapter considers the results obtained from the application of the procedure. First, a comparison is made of each team's initial and final cut-off scores to ascertain the impact of the statistical data and the student scripts. Secondly, the cut-off scores set by the teams in the second year of the study are compared to determine the extent to which teams working independently can set the same standards. Thirdly, the proportions of students reaching each standard level in a course in the two years are compared. This information is then examined to see what it indicates about the relative performances of the two cohorts.

6.2 COMPARISON OF EACH TEAM'S INITIAL AND FINAL CUT-OFF SCORES

Examining the differences between the cut-off scores set by each team of judges when it first came together, and those it finally settled on at the end of the process, provides an indication of the effectiveness of the statistical feedback and the student scripts. If there is no change, or only very minor adjustments, it might be implied that providing this extra information and going through the iterations in the process add very little and, so, are not worth the effort. Alternatively, it could mean that the judges' initial estimates were correct and the data confirmed these estimates.

If, on the other hand, the team made adjustments to its original values, there is a strong indication that the statistical feedback and/or the student scripts assisted the judges in reflecting on and refining the cut-off scores they had set. Whatever the situation, providing the data will give judges added confidence in the outcome of their deliberations.

6.2.1 The First Year (1994)

The initial and final cut-off scores[2] for each course established by the panel of judges in the first year are shown in Table 6.1. The initial values were obtained after the process of negotiation and discussion when the judges compared the values they had set individually. The final cut-off scores are those agreed by the team after consideration of the statistical data and the student scripts.

TABLE 6.1

Differences between the Initial (I) and Final (F) Cut-off Scores for Each HSC Course Examination in 1994

Mathematics	**Initial (I)**	**Final (F)**	**Difference (I - F)**
Excellent	119	113	6
Very Good	103	98	5
Good	74	72	2
Satisfactory	48	48	0
English	**Initial (I)** [1]	**Final (F)**	**Difference (I - F)**
Excellent	112	99	13
Very Good	98	86.5	11.5
Good	86	74	12
Satisfactory	69	53.5	15.5
Biology	**Initial (I)** [2]	**Final (F)**	**Difference (I - F)**
Excellent	54	50.5	3.5
Very Good	48	44.5	3.5
Good	39	37.5	1.5
Satisfactory	28.5	29	-0.5

[1] The Initial (I) values shown are those set by the judges involved in scoring Paper 1.

[2] The values reported for Biology relate to the compulsory sections of the examination, scored out of 60.

To obtain a measure of the extent of change between the initial and final cut-off scores, the average absolute difference (AAD) was calculated using the formula

$$AAD = \frac{\sum_{i=1}^{n} |I_i - F_i|}{n} \qquad \textbf{Equation 6.1}$$

where I_i = an initial cut-off score for cut-off i;

F_i = the corresponding final cut-off score;

n = the number of cut-off scores used for the examination

[2] In the tables in this chapter the shorter version "Excellent" is used, rather than "Excellent/Very Good" to refer to the borderline between "Excellent" and "Very Good".

In Mathematics, there are substantial differences within two pairs of initial and final cut-off scores. These differences occur at the top two standard levels where the judges realised that, after looking at the statistical data and the scripts, their expectations were too high. The AAD between the initial and final scores is 3.25.

In the case of English, there were large changes made to all the initial cut-off scores. The judges commented during the procedure that their expectations were too high. The AAD between the initial and final scores is 13.

Like Mathematics, the initial cut-off scores for Biology underwent greater change at the top levels. Once again the team's initial expectations of the better students appear to have been too high, as the final cut-off scores, for the two top categories in particular, were lower than the initial scores. The team modified its decisions after considering the statistical data and the student work, resulting in an AAD of 2.25.

The size of the differences between the initial and final cut-off scores suggests that the statistical data and student scripts had an impact on the process of establishing the final cut-off scores.

This finding is supported by the results from the survey conducted at the end of the second year of the study. In the survey, the judges were asked how important they felt the statistical data and student scripts were in assisting them to establish cut-off scores. Judges, both from the original teams created in 1994 and the new teams created in 1995, indicated that these forms of information were effective ways of providing feedback,

and were important steps in the procedure. This issue is further discussed in the next chapter in the summary of the judges' responses to the survey.

6.2.2 The Second Year (1995)

Similar information relating to the second year of the study is shown in Table 6.2. In each case the teams worked independently and so were not aware of the values set by the other team(s) working with their course. In each case, Team 1 is the team that had been involved in setting the performance standards in the initial year.

The difference between the approaches used in the two years is that in the second year, instead of the judges using their own "images" of Excellent, Very Good and so on, these standards of performance were reflected in material provided to the judges. This package consisted of the descriptor statements, the examination paper and a sample of student scripts awarded the cut-off scores set in the first year.

If the differences between the initial and final values proposed by each team are relatively small, it may be inferred that the materials in the standards package and the way they are used in the procedure are effective in helping judges to internalise the standards.

It can be seen from Table 6.2 that each of the Mathematics teams had relatively small differences between the initial and final estimates. The AADs for Team 1, Team 2 and Team 3 are 2.25, 1.5 and 2.75 respectively, compared to 3.25 for Team 1 in the initial year. The biggest value for any difference between an initial and final cut-off score is 5.

TABLE 6.2

Differences between the Initial (I) and Final (F) Cut-off Scores for Each HSC Course Examination in 1995

Mathematics	*Team 1*			*Team 2*			*Team 3*		
	Initial (I)	**Final (F)**	**I - F**	**Initial (I)**	**Final (F)**	**I - F**	**Initia l (I)**	**Final (F)**	**I - F**
Excellent	112	113	-1	113	113	0	111	113	-2
Very Good	101	104	-3	97	99	-2	95	100	-5
Good	76	78	-2	77	78	-1	75	77	-2
Satisfactory	50	47	3	53	50	3	51	49	2
English	*Team 1*			*Team 2*			*Team 3*		
	Initial (I)	**Final (F)**	**I - F**	**Initial (I)**	**Final (F)**	**I - F**			
Excellent	102	100	2	101.5	100	1.5			
Very Good	90	88	2	88.5	87.5	1			
Good	77	74	3	74	70.5	3.5			
Satisfactory	54	53	1	55	51	4			
Biology [1]	*Team 1*			*Team 2*			*Team 3*		
	Initial (I)	**Final (F)**	**I - F**	**Initial (I)**	**Final (F)**	**I - F**			
Excellent	64.3	63.4	0.9	65.4	65.4	0			
Very Good	56	52.9	3.1	54.7	54.7	0			
Good	46.8	44.3	2.5	40	44	-4			
Satisfactory	35.6	34.8	0.8	26.2	34	-7.8			

[1] The initial and final cut-off scores for Biology are based on the compulsory items which had a maximum possible score of 75.

The differences are also quite small for English. The biggest difference between an initial and final value is 4, with the AAD for Team 1 and Team 2 being 2.0 and 2.5 respectively compared to 13.0 for Team 1 in the initial year.

In the case of Biology, the AADs are 1.6 and 2.95 respectively, compared to 2.25 for Team 1 in the initial year. Team 2 made no change to its top two cut-off scores, but increased the cut-off scores for Good/ Satisfactory and Satisfactory/ Unsatisfactory, which were much lower than the values set by Team 1. It would appear that the statistical data and student scripts were effective in assisting Team 2 to adjust its cut-off scores for the bottom two levels.

The relatively small AAD for each course and each team in the second year compared to the initial year suggests that the packages of materials were effective in assisting judges to understand the standards they were to apply to the 1995 examination paper. For English, in particular, the descriptor statements, student scripts from the 1994 examination and the review of the 1994 examination paper were quite effective. In 1995, the judges generally established initial cut-off scores using the standards package, which were quite close to the values on which they finally settled, unlike the situation in 1994 when the differences were relatively large. Similar patterns of results were found in Mathematics and Biology.

6.3 COMPARISON OF THE CUT-OFF SCORES SET BY THE YEAMS IN EACH COURSE IN THE SECOND YEAR

6.3.1 The Initial Cut-off Scores Set for Each Course

An important question which this study has posed is whether different teams of judges, working independently, can use the procedure and arrive at very similar cut-off scores, thus demonstrating that they are applying the same standards of student performance in the types of examination used for the NSW Higher School Certificate (HSC).

The differences between the initial cut-off scores set by the teams working within a course during the second year of the study are shown in Table 6.3.

In most cases the initial cut-off scores set by a team are close to the values set by the other team(s) for that course. In Mathematics there appears to be substantial differences, at the Very Good/Good borderline, between Team 1 and the other two teams. Teams 2 and 3, however, have relatively consistent cut-off scores.

The English teams generally are in close agreement, with the biggest difference being 3. Team 1 has set 77 for the Good/Satisfactory cut-off score, whereas Team 2 has proposed 74 as its value.

In Biology, while the teams set very similar cut-off scores for the Excellent/Very Good and Very Good/Good borderlines, there is a substantial difference between the values set for the Good/Satisfactory and Satisfactory/Unsatisfactory borderlines.

The differences for Biology are probably due, to some extent, to changes in the course and its examination paper between 1994 and 1995. These changes seem to have made it a more difficult task for the judges to internalise the standards set using the 1994 materials, and then apply these same standards to the 1995 examination, which had a different structure and differences in its content domain. The fact that the teams of judges were able to reach a greater level of agreement for the Excellent/Very Good and Very Good/Good cut-off scores might be expected given the observation made by judges during the process that the better students are often more consistent in their

performances. This makes it easier for judges to predict the item scores these students will receive.

TABLE 6.3

Comparison of Initial Cut-off Scores Set by Teams for all Courses in the 1995 Examinations

Mathematics	*Team 1*	*Team 2*	*Team 3*	*T1 - T2*	*T1 - T3*	*T2 - T3*
Excellent	112	113	111	-1	1	2
Very Good	101	97	95	4	6	2
Good	76	77	75	-1	1	2
Satisfactory	50	53	51	-3	-1	2
English	*Team 1*	*Team 2*		*T1 - T2*		
Excellent	102	101.5		0.5		
Very Good	90	88.5		1.5		
Good	77	74		3		
Satisfactory	54	55		-1		
Biology	*Team 1*	*Team 2*		*T1 - T2*		
Excellent	64.3	65.4		-1.1		
Very Good	56	54.7		1.3		
Good	46.8	40		6.8		
Satisfactory	35.6	26.2		9.4		

The values shown in Table 6.3 support the claim that the materials embodying the standards set in the initial year (the descriptors, examination paper and student scripts) are effective in assisting judges to internalise the standards to be applied when setting cut-off scores on examinations of the type used for the HSC.

6.3.2 The Final Cut-off Scores Set for Each Course

A comparison of the final cut-off (F) scores proposed by the teams in the second year will indicate whether, by following the full procedure, the teams are consistent in establishing the same or very similar cut-off scores for an examination paper. The cut-off scores proposed by each team in the 1995 examinations are shown in Table 6.4.

In this case the average absolute difference (AAD) can be used to measure the similarity between cut-off scores proposed by the different teams. It is calculated using the formula

$$AAD = \frac{\sum_{i=1}^{n} |F_i - F_j|}{n} \qquad \textbf{Equation 6.2}$$

where F_i = a final cut-off score proposed by Team i;

F_j = a corresponding cut-off score proposed by Team j;

n = the number of cut-off scores used for the examination

It can be seen from Tables 6.3 and 6.4 that for Mathematics there is a slight increase in agreement between the initial and final cut-off scores proposed by the teams after the statistical information and the student scripts have been considered by the judges. The average absolute difference (AAD) between the initial cut-off scores set by Team 1 (T1) and Team 2 (T2) is 2.25, while the AAD between the final cut-off scores set by these teams is 2.0. For Team 1 and Team 3 the corresponding values went from 2.25 to 1.75.

TABLE 6.4

Comparison of Final Cut-off Scores (F) Set by Teams for all Courses in the 1995 HSC Examinations

Mathematics	*Team1*	*Team 2*	*Team 3*	*T1 - T2*	*T1 - T3*	*T2 - T3*
Excellent	113	113	113	0	0	0
Very Good	104	99	100	5	4	-1
Good	78	78	77	0	1	1
Satisfactory	47	50	49	-3	-2	1
English	*Team 1*	*Team 2*		*T1 - T2*		
Excellent	100	100		0		
Very Good	88	87.5		0.5		
Good	74	70.5		3.5		
Satisfactory	53	51		2		
Biology	*Team 1*	*Team 2*		*T1 - T2*		
Excellent	63.4	65.4		-2		
Very Good	52.9	54.7		-1.8		
Good	44.3	44		0.3		
Satisfactory	34.8	34		0.8		

While there are some changes to individual cut-off scores proposed by the English teams, the AAD between the teams is 1.5 for both the initial and final cut-off scores. For Biology, there is a relatively large decrease from the AAD of the initial cut-off score to that of the final values. The initial value of 4.7 has dropped to 1.2. This difference is largely attributable to the increase in agreement between the teams on the final values for the Good/Satisfactory and Satisfactory/Unsatisfactory cut-off points.

The information presented in Tables 6.3 and 6.4 suggests that teams of judges are able to internalise the standards they are to apply from the package of materials provided. The results also indicate that the statistical feedback provided by the Extended

Logistical Model, and the sample of student scripts chosen at the cut-off scores, are important elements in the procedure. As this information tends to bring the cut-off scores proposed by the teams closer together, it is suggested that this information helps to improve the reliability of the decisions made by the judges.

For the 1995 examinations, while each team carefully considered the student scripts selected at the cut-off scores, no team made adjustments to any of their values based on these scripts. Instead, the judges used this information to confirm their decisions by satisfying themselves that the scripts were at the lower end of the standard level in which they were being placed. In this way they used the scripts as a validation of the values they had set.

All teams commented that the procedure had enabled them to produce final cut-off scores that they felt were appropriate for the standards they were expected to apply. This was further reinforced by comments the judges made in the surveys they completed.

6.4 COMPARING STUDENT PERFORMANCE LEVELS IN THE DIFFERENT YEARS

By setting the cut-off scores on the 1995 examinations using the standards established in 1994, the judges are equating the examinations. This section discusses factors which pertain to the issue of how well the judges were able to internalise the standards set in the 1994 examinations and then apply them to the 1995 examinations.

Evidence that the same standards of performance were being applied in the two years comes from the comparisons of student scripts. The procedure is designed to give the

judges the opportunity to review and refine decisions they made earlier in the process. Reviewing student scripts gives the judges the opportunity to compare pairs of scripts from the two years at the same borderline, with a view to deciding whether they represent the same standards of performance in the course. The English teams found this to be a relatively simple task and so confirmed that the standards of performance demonstrated by the pairs of scripts were the same.

In Mathematics and Biology, however, the judges found that the comparison of scripts in this way was not an easy task, given the variation between the two years in examination content, order and emphasis previously discussed. Having reviewed pairs of scripts and discussed their impressions with their colleagues, however, the judges concluded that there was no evidence to suggest that the scripts represented different standards of performance.

One approach to collecting further confirmatory evidence of the accuracy of the equating process would be to create an additional team of judges who had not previously been involved in the exercise. These judges could be given the task of performing pair-wise comparisons on the samples of scripts from the two years selected at the borderlines. These judges would not follow the full procedure, but would simply ascertain whether the scripts at each of the borderlines represented equivalent standards of performance. This would not be an easy task in some cases, as discussed above, but a structured procedure could readily be established to support such an operation.

Further evidence provided by the study that the cut-off scores for the two years identify the same standards of performance comes from the fact that the cut-off scores

established by each team for the 1995 examinations are similar to those established by the other team(s) working on that course. The teams, working independently, used the materials from the 1994 examinations to "learn" the standards to apply to the 1995 examinations. The fact that, in doing so, they produced similar cut-off scores for the examination supports the claim that the standards applied in the two years are the same.

Two factors which can impact on the accuracy of the equating process - variation in the difficulty level of the examination papers, and variations in the stringency of the scoring keys - have been taken into account during the application of the procedure.

The first of these factors, the relative difficulty of the examination papers, can be managed during the application of the procedure itself. When judges internalise the standards of performance from the package of materials, they use this information to generate images of "borderline students". They then decide how such students would perform on each item in the new examination. In this way, they nullify the effects due to variations in the difficulty level of the items comprising the examinations. To a large extent, being able to account for this variable is dependent on the quality and effectiveness of the training given to the judges, the experience and quality of the judges themselves, and the careful application of the procedure.

The effects of the second source of variation, differences in the scoring keys, can be overcome by providing the judges with the keys. In this study, the judges were not provided with the scoring keys, nor were they told the cut-off scores established in the previous year. This information was not provided to ensure they focused on the items and on the materials which demonstrated the standards to be applied. It was decided to

withhold this information in this part of the study to eliminate the possibility that the judges would be unduly influenced by the raw score values from the previous year.

In practice, it would seem reasonable to give the judges access to the scoring keys used and, probably, the scores awarded to the sample scripts they are given. This step should assist the judges to internalise the standards they are to apply to the new examination paper. Judges would need to be informed of how they are to use this information, and not simply seek to "short cut" the process by adopting the values from the previous year. Provided the judges follow the procedure as specified, the provision of this information should not adversely affect the integrity of the process.

In order to consider the results of equating the 1994 and 1995 examinations, a single set of cut-off scores for each course for the 1995 examinations was created. This was achieved by averaging the cut-off scores established by the teams. The scores that will be used to further analyse the equating of the 1994 and 1995 examinations are shown in Table 6.5.

TABLE 6.5

Cut-off Scores for Each Course in the Two Years of the Study

Mathematics	*1994*	*1995*
Excellent	113	113
Very Good	98	101
Good	72	78
Satisfactory	48	49
English	*1994*	*1995*
Excellent	99	100
Very Good	87	88
Good	74	72
Satisfactory	55	52
Biology	*1994*	*1995*
Excellent	84	86
Very Good	74	72
Good	63	59
Satisfactory	48	46

In the following sections the cumulative proportions of the students in each performance standard in the two years, based on the cut-off scores in Table 6.5, are shown. Consideration is given to whether the proportions of students in the two years represent any significant variation in the relative performances of the student groups.

6.4.1 Mathematics

The cut-off scores established in Mathematics for the two years, and the corresponding proportions of the candidatures in each performance standard, are reported in Table 5.18. Those proportions are summarised as cumulative percentages in Table 6.6.

TABLE 6.6

Cumulative Proportions of Students in Each Performance Standard in the HSC Mathematics Examination in 1994 and 1995

Standard	*1994 Cum. %*	*1995 Cum. %*
Excellent	0.9	1.7
Very Good	12.5	14.8
Good	42.8	46.1
Satisfactory	68.8	77.4

On the basis of the values in Table 6.6, is it reasonable to conclude that the 1995 group of students performed better than the 1994 group?

A Kolmorgorov-Smirnov (Siegel, 1956) two-tailed test was conducted on the proportions in Table 6.6. The null hypothesis, that there was no significant difference between the 1994 and 1995 candidatures, is rejected at the 0.05 level. It is, thus, likely that the two candidatures differed in their mathematical ability. The higher proportions

of students within each standard level above the Satisfactory/Unsatisfactory borderline suggest that the 1995 students performed better than the 1994 student group.

In seeking evidence to support or refute this conjecture, consideration was given to the nature of the candidatures of all Mathematics courses in the two years.

In Chapter 4 it was noted that the course which is part of this study is one of several Mathematics courses examined for the NSW Higher School Certificate. All courses fit within a hierarchy of difficulty. For reasons of simplicity, they will be referred to as courses A, B, C, and D. The course which is the focus of this study is course B, which is the second most difficult. The candidatures of these courses in 1994 and 1995 are shown in Table 6.7.

Based on data collected over many years, it can be assumed that the better students of Mathematics generally select the more difficult courses - that is, courses A and B. There has been a drop between 1994 and 1995 of over 900 in the number of students who took course A. As most students take a Mathematics course as part of their HSC studies, a relatively large number of able Mathematics students most likely elected to take course B in 1995 rather than course A. The candidature of course C increased between 1994 and 1995. This increase would have come predominantly, although not exclusively, from the weaker students who would, in previous years, have elected to take course B. Given these changes in the relative course candidatures, it could be implied that the 1995 course B candidature, as a whole, was more able than the 1994 candidature. This is consistent with the results of the Kolmogorov-Smirnov test that indicate that the 1995 cohort performance was superior to that of the 1994 cohort.

TABLE 6.7

Candidature of the Mathematics Courses Examined for the NSW Higher School Certificate

Course	*1994*		*1995*	
A	3 403	6.1%	2 495	4.6%
B	28 289	50.8%	26 040	48.3%
C	20 246	36.4%	21 060	39.1%
D	3 725	6.7%	4 323	8.0%
Total	**55 663**		**53 924**	

The change in the size and nature of the candidature in course B between 1994 and 1995 is consistent with the results obtained from the Kolmogorov-Smirnov test which indicates that the performance of the 1995 candidature was superior to that of the 1994 candidature. This would imply that the judges for Mathematics have successfully applied the same or very similar standards of performance in determining the cut-off scores in the two years. The judges appear to have been able to take account of variations in examination paper difficulty, scoring keys and ability of candidatures between the two years. Hence, the higher proportions of the 1995 candidature in each standard level, compared to 1994, is probably due to the superior performance of the 1995 cohort.

6.4.2 English

From Table 6.8 it can be seen that the cumulative percentages of students allocated to each performance standard by the teams are close. The biggest difference occurs in the Satisfactory category. Whereas 37.5% of the 1994 cohort and 38.3% of the 1995 cohort

were deemed at least "Good", the proportions of the cohorts deemed at least "Satisfactory" were 87.8% in 1994 compared to 82.1% in 1995.

A two-tailed Kolmogorov-Smirnov test indicated that the difference between the two groups was significant at the 0.05 level. Hence, given the values in Table 6.8, it can be asserted that the performance of the 1994 candidature is superior to that of the 1995 candidature, to the extent that a higher proportion of students were placed in the Satisfactory category in 1994. In the categories Excellent, Very Good and Good, the proportions in the two years are very similar.

TABLE 6.8

Cumulative Proportions of Students in Each Performance Standard in the HSC English Examination in 1994 and 1995

Standard	*1994 Cum. %*	*1995 Cum. %*
Excellent	1.0	1.6
Very Good	9.2	9.6
Good	37.5	38.3
Satisfactory	87.8	82.1

As was the case for Mathematics, the English course used in this study is one of several which students can take. For simplicity of explaining the relationship between the courses, they will be referred to as courses A, B and C. Course A is the most difficult. It is course B that is used in this study. The numbers of candidates taking these courses are shown in Table 6.9.

TABLE 6.9

Candidature of the English Courses Examined for the NSW Higher School Certificate

Course	*1994*		*1995*	
A	10 091	17.8%	8 723	15.8%
B	30 226	53.3%	28 937	52.4%
C	16 434	28.9%	17 511	31.7%
Total	**56 751**		**55 171**	

There was a decrease of nearly 1400 students in the candidature of the highest level course between 1994 and 1995. Many of these students would have taken course B. There was also a drop of approximately 1300 students in the candidature of course B. Many, but not all, of these students would probably have come from among the weaker students who would previously have taken course B.

Unlike the situation with Mathematics, however, able students of English do not necessarily choose to study the more difficult courses. Evidence provided by teachers and examiners over recent years indicates that many very capable English students elect to study courses B and C. Equally, some students more suited to study course C insist for a variety of reasons on taking course B. As a result, there is a relatively large group of students of mediocre ability in the lower tail of the distribution of course B.

Hence, given the changes in the size and nature of the candidatures in course B between 1994 and 1995, and from the results obtained, it would seem reasonable to conclude that the performance of the more able group of students in 1995 was similar to that of the 1994 students, but that the performance of the less able group in 1994 was superior to that of the 1995 group.

6.4.3 Biology

Table 6.10 shows that the proportions of students assigned to each performance standard in each year are very different. Given the discrepancies between these proportions, is it reasonable to conclude that either the Biology judges have not been able to apply the same standards, or that the performances of the students in 1995 were markedly superior to that of the 1994 students?

TABLE 6.10

Cumulative Proportions of Students in Each Performance Standard in the HSC Biology Examination in 1994 and 1995

Standard	*1994 Cum. %*	*1995 Cum. %*
Excellent	0.7	4.4
Very Good	10.8	28.5
Good	35.9	55.1
Satisfactory	65.9	76.4

In addressing these questions, certain factors need to be reiterated. First, the examination paper in 1995 had a very different structure to that in 1994. In 1995, students attempted compulsory items worth a total of 75 marks and a single item, worth 25 marks, based on the elective area of the Biology course they had studied. In 1994 the compulsory items had a total of 60 marks and students needed to provide responses to three items based on different elective areas. Each of these items had a maximum possible score of 13.

Secondly, a number of the judges in the teams used in this study indicated that they were not confident in predicting what scores borderline students would receive for some

of the optional items. Given that the elective items made up 39% in 1994 and 25% in 1995 of the total examination score, the procedures used to calculate cut-off scores for these elective items may have added to the variation in the proportion of students at each performance standard in the two years.

The level of agreement between the two teams about the cut-off scores that they set for the 1995 examination suggests they were applying the same standards to that paper. It is possible, then, that despite the changes between the two years, the teams did apply the same standards, and it was the use of the cut-off scores established for the compulsory items in 1995 to create total examination cut-off scores which was responsible for increasing the discrepancies. The need to establish a consistent and appropriate means of taking into account both core and elective items in establishing cut-off scores is discussed elsewhere in this thesis.

The question, then, of whether the performance of the 1995 students, as a group, was superior to that of the 1994 students is difficult to answer. The Kolmogorov-Smirnov test rejects the hypothesis of no difference between the two groups at the 0.05 level. In addition, the level of agreement between the two teams in 1995, and the variation in the size and probable ability of the two groups implies that the 1995 group performed better. Given the difficulties experienced by the judges in addressing the issue of the optional items, and the profound change to the course content and examination structure between the two years, the extent of any improvement would be difficult to quantify accurately.

Two further points arise from the consideration of the Biology data. First, as already discussed alternative approaches to handling the optional items in an examination need to be explored. These items are an important part of the examination and add a significant amount of information on the performance of the students. The outcomes of this aspect of the study show that care needs to be exercised in the selection of judges to ensure that a sufficient number with expertise in preparing students for examination in these areas of the Biology course are included in the team. In Chapter 5 an approach was suggested which is consistent with the standard-setting procedure, and also makes use of the capability of the ELM for putting optional items on the same scale. Whatever approach is used, the selection of judges should ensure, as far as possible, that the judges within a team, collectively, must have expertise across all areas of the course. This will mean that professional judgment will be used as the main factor in establishing the cut-off scores.

A second issue that emerged relates to cases where there is a major change to the composition and structure of the examination, as was the case for Biology between 1994 and 1995. Such changes seem to make it more difficult for judges to apply the standards they have internalised by considering materials from one examination when equating that examination and the new one. To overcome this problem one approach might be to have the judges spend additional time considering those items that examine similar content in the two courses. This could be achieved if a meeting of the judges was held prior to their setting their individual item cut-off scores. In this way they could share their views about the items in the two examinations to which they should give particular consideration. In this way they can build up a common understanding of the similarities and differences in the two examinations.

6.5 SUMMARY

The cut-off scores developed for each course during this study were examined in several ways.

First, the cut-off scores set by the judges following their initial discussions were compared with the final scores they set after they had had a chance to consider the statistical data and the student scripts. This comparison was conducted for both years, and showed that the statistical data and the student scripts were effective in assisting judges to refine their initial cut-off scores. Furthermore, as the differences between the initial and final cut-off scores were smaller in the second year, it can be concluded that the package of materials the judges were given, which exemplified the standards to be applied, was effective in helping them to internalise those standards.

Secondly, a comparison was made of the cut-off scores produced by the different teams in each course operating in the second year of the study. These comparisons showed that, in virtually all cases, the amount of difference between the teams on the final cut-off scores was smaller than the difference on the initial cut-off scores. Further, the teams produced very similar cut-off scores. This exercise provided further evidence that the multi-stage standard-setting procedure used in this study is effective in assisting judges to become familiar with the performance standards to be applied and, thus, to establish cut-off scores reflecting those same standards on subsequent examinations.

Finally, the proportions of students in a course placed in each performance standard in the two years were compared. All else remaining the same, significant variations in

these proportions across different years would imply a change in the standard of performance of the student groups. This study has shown that the standard-setting procedure can take account of differences in the degree of difficulty of the examinations and significant changes in the composition of the student group. The provision of the scoring key and the scores awarded to borderline scripts from the previous year may assist in this step. However, where there are major changes in the structure and content domain of the examination paper, it is difficult to accurately compare standards across different years.

In situations like that of English and Mathematics, where the structure of each examination was the same in 1994 and 1995, the procedure appears to provide a viable means of comparing how students from different years, taking different examinations within the same course, performed in relation to the same course performance standards. In such a situation, it is then possible to determine whether the performance of the group in the second year is better, poorer or the same as in the initial year.

CHAPTER 7

ISSUES OF RELIABILITY AND VALIDITY AND THE VIEWS OF THE JUDGES

7.1 INTRODUCTION

The application of the procedure using the 1994 and 1995 examinations in Mathematics, English and Biology courses shows that panels of experienced teachers can use the procedure to produce cut-off scores related to standards of performance in comprehensive curriculum-based examinations containing items which are scored polytomously. Further, the results presented in Chapters 5 and 6 provide evidence that a team of suitably qualified judges can use materials and data relating to performance standards to equate two different examinations for the same course. The key questions relating to the reliability and validity of the results emanating from such a procedure are addressed in this chapter.

7.2 RELIABILITY EVIDENCE

In Chapter 2, approaches which have been used in the past (eg Jaeger, 1990; Plake and Impara, 1996; Breyer and Lewis, 1994) to measure both intra-judge and inter-judge reliability were discussed. A number of these approaches are used to provide evidence of the reliability of the procedure advocated in this paper.

7.2.1 Intra-judge Reliability

Intra-judge reliability concerns the degree of variation in the decisions of individual judges within and across the various steps of a multi-stage standard-setting process.

The smaller the degree of variation, the more reliable the judge in the application of standards.

Studies similar to this one required the judges to refine their individual decisions throughout the process. These individual decisions are averaged at the end to obtain the cut-off scores. In contrast, the procedure advocated in this study requires that individual decisions made by the judges be discussed and refined in such a way as to reach a consensus decision on the part of the team at an early stage in the process. From that point on, it is the consensus values that the team works with and refines collectively throughout the rest of the process. Hence, problems in establishing cut-off scores, which can be an issue in some exercises due to low intra-judge reliability, were largely overcome in this study by the use of the consensus approach.

It is possible that individual judges may have contributed to the team discussions in such a way that their input throughout the application of the procedure was inconsistent. This is largely irrelevant, however, as the discussions held within the teams gave each judge a chance to reflect upon his/her recommendations and discuss them to the point where each member of the team was satisfied with the outcome.

In order to obtain some measure of the intra-judge reliability, the cut-off scores initially proposed by the individual judges in both English and Mathematics in 1995 were considered. This information could not be determined for Biology as the item cut-off scores for the individual judges were not available. For each course, the reliability was measured by calculating the correlation coefficients between the item cut-off scores proposed by the judges and the item cut-off scores agreed by the whole team at the end

of the first round of discussion and review. It might be argued that this is more a measure of "intra-team reliability". Nevertheless, it could be suggested that if the cut-off scores individually determined by the judges are in reasonable agreement with the cut-off scores agreed by the team, it is unlikely that they will demonstrate significant drops in the levels of intra-judge reliability between the different stages of the procedure. Such an outcome, however, can not be guaranteed.

For the English judges, these correlation coefficients were all greater than 0.95. The values for Mathematics, with the exception of one judge whose values were around 0.5 to 0.6, were generally in the range 0.85 to 0.95. This suggests that by and large, the judges were able to independently set cut-off scores that were not too dissimilar from the cut-off scores initially established by their team. In the second year of the study, then, it seems that the judges were able to develop common "images" of borderline students from the materials provided and determine cut-off scores quite close to the consensus scores. This information does not provide a comprehensive picture of the intra-judge reliability. Nevertheless, these results give some confidence that, having been relatively close to the team's agreed scores at this early stage, judges are unlikely to vary to any great extent during the rest of the process, their understanding of the standards they are applying.

The correlation coefficients between the cut-off scores initially proposed by the judges in the first year of the study were not calculated. It would be expected that they would be lower than the values from the second year, however, as the judges set their initial cut-off scores based on their individual concepts of Excellent, Very Good, and so on, rather than interpreting the standards from concrete materials.

7.2.2 Inter-judge Reliability

Inter-judge reliability is the measure of the homogeneity of the final decisions made by the judges (Berk, 1996).

Because the judges achieve a consensus position on item cut-off scores at the end of the first period of discussion, this removes many of the potential problems associated with inter-judge reliability. Differences certainly existed initially, as would be expected, but these were resolved through discussion and review. Differences also arose as a result of considering the statistical data and the student scripts, but again the judges resolved any differences during their discussions.

In order to obtain a measure of the inter-judge reliability in the early stages of the procedure, the correlation coefficients were calculated using the initial item cut-off score recommendations made by each possible pair of judges in a team.

While it was believed that the team discussions throughout the process would greatly improve the reliability of the exercise, it is useful to know to what extent the judges were in general agreement after they had made their individual decisions at the beginning of the process. To assess the extent of this agreement between pairs of judges from each team, correlation coefficients were calculated between the item cut-off scores for each performance standard initially proposed. For English, these values were all greater than 0.97. In the case of Mathematics, the values ranged between 0.7 and 0.9.

The generally high values of the correlation coefficients show the strong agreement between the judges at the early stage of the process. This is supporting evidence that, even before the process of discussion and review took place, the judges had been able to quite accurately internalise the standards and apply them to the 1995 examination paper to produce cut-off scores which were consistent with those proposed by other judges.

7.3 VALIDITY EVIDENCE

Kane (1994) identifies three sources of evidence for validating a standard-setting procedure and the cut-off scores that it produces. These consist of the procedure itself, the data generated within the standard-setting process, and comparisons with external sources of information.

7.3.1 Evidence Associated with the Procedure Itself

The judges who set the cut-off scores in the courses in this study had for many years taught their course and prepared students for the examination. All had experience in teaching the relevant course to students of varying abilities, and so on the whole, were aware of the full range of performances of students in the examinations. In some cases in the initial year, judges who taught in schools where the levels of performance tended to be skewed compared to the total candidature, took a little time to adjust their expectations so that they took account of the whole range of performances. This was not really a problem in the second year where the judges had the descriptor statements, examination paper and student responses to guide them. Most of the judges had been involved in the process of scoring the examinations for a number of years, and many currently hold leadership positions in this process. All judges have experience in teaching and assessing their subject. In spite of the fact that some of the Biology judges

did not regard themselves as having expertise in some of the elective areas of the course, they were nevertheless aware of the overall standard of work produced by the strongest and weakest students in this course.

The "procedural fidelity" (Berk, 1996) of the standard-setting procedure itself also contributes to the validity of the process. Chapter 3 shows that the procedure is highly structured and contains a series of steps designed to permit the review and refinement of earlier decisions. The briefing of judges, including the provision of written instructions, and the guidance and support provided to the judges throughout the process, all contribute to the rigour and order of the procedure.

7.3.2 Evidence Provided by the Application of the Procedure

7.3.2.1 The Variability of the Item Scores Achieved by Borderline Students

Kane (1994) states that one way of testing the validity of a cut-off score is to ascertain whether students who achieve that examination cut-off score also achieve scores for each item similar to the item cut-off scores proposed by the judges.

This test was performed by comparing the item score profile of all students in the examination who had achieved one of the cut-off scores with the item score profile for that cut-off score as finally determined by the judges. The "item score profile" is the ordered set of scores a student is awarded for the items in an examination. For each cut-off score within a course, a distribution of differences was created, showing the proportion of students awarded that cut-off score whose individual item scores varied from the corresponding item cut-off scores determined by the team of judges. A "difference" was considered to have occurred when the score for an item awarded to a

student varied from the score for that item determined by the judges by at least one quarter of the maximum possible score for that item. For example, for an item with a maximum score of 20, a "difference" would be recorded if the difference between the score achieved by a student and the value set for that item by the judges was 5 or more. All students who had been awarded a total score equal to a particular cut-off score were grouped according to whether they had 0, 1, 2, ... differences between their item score profile and the judge-determined item score profile. The maximum number of possible differences that could be obtained is equal to the number of items in the examination. In both Mathematics and English, this value would be 10.

This particular test does not directly measure the validity of the actual cut-off score, but rather the validity of the item score profile set by the judges which created that cut-off score. If, however, there is a high degree of agreement between the item score profile set by the judges and the patterns of item performances submitted by the students, this would lend weight to a claim that the judges were well attuned to the standards of performance of students in the course. The statistical analysis provided to the judges as part of the standard-setting procedure, if heeded, would contribute to ensuring that most of the score profiles were consistent with what would be estimated for a "typical" student at the cut-off score.

The results of using this test for each Mathematics team at the Satisfactory/ Unsatisfactory, Good/Satisfactory and Very Good/Good borderlines are shown in Table 7.1. For each team, the first column shows the categories for the number of differences that were recorded. The second column shows the proportion of the students in each category. The third column contains the cumulative proportions. For example, there

are no "differences" for 9.9% of the students who scored 47, the cut-off score for the Satisfactory/Unsatisfactory borderline established by Team 1. For the same team and same cut-off score, 82% of students had three or fewer differences.

TABLE 7.1
Proportions of Borderline Students Demonstrating Item Score Variability in 1995 HSC Mathematics

Cut-off	Team 1		Team 2			Team 3	
Satisfactory	47		50			49	
0[1]	0.099[2]	0.099[3]	0	0.127	0.127	0	0.118 0.11
1	0.239	0.338	1	0.224	0.351	1	0.236 0.354
2	0.320	0.658	2	0.237	0.588	2	0.214 0.568
3	0.162	0.820	3	0.25	0.838	3	0.245 0.813
4	0.108	0.928	4	0.114	0.952	4	0.123 0.936
5	0.054	0.982	5	0.031	0.983	5	0.045 0.981
6	0.018	1.000	6	0.004	1.000	6	0.018 1.000

	Team 1		Team 2			Team 3	
Good	78		78			77	
0	0.203	0.203	0	0.209	0.20	0	0.115 0.11
1	0.359	0.562	1	0.339	0.54	1	0.249 0.36
2	0.284	0.846	2	0.293	0.84	2	0.339 0.70
3	0.116	0.962	3	0.119	0.96	3	0.204 0.90
4	0.032	0.994	4	0.017	0.97	4	0.070 0.97
5	0.003	0.997	5	0.017	0.99	5	0.019 0.99
6	0.003	1.000	6	0.003	1.00	6	0.003 1.00

	Team 1		Team 2			Team 3	
Very	104		99			100	
0	0.612	0.612	0	0.401	0.40	0	0.268 0.26
1	0.304	0.916	1	0.331	0.73	1	0.338 0.80
2	0.075	0.991	2	0.220	0.95	2	0.294 0.90
3	0.006	1.000	3	0.034	0.98	3	0.078 0.97
4			4	0.011	0.99	4	0.023 1.00
5			5	0.003	1.00	5	

1 This column gives the number of instances where the difference between the item cut-off score proposed by the judges and the item score obtained by a student who obtained the total examination cut-off score was greater than or equal to one quarter of the maximum possible score for the item. This is referred to as a difference.

2 This column shows, for the particular examination score, the relative frequency of the number of differences as a proportion which were observed in the item scores of students who obtained that examination score.

3 This provides the cumulative frequencies as proportions of the previous column.

The item cut-off scores proposed by all teams for the Very Good/Good borderline are consistent with the item score profiles obtained by the students who were awarded those scores. For example, the proportion of students with two or fewer differences for the examination cut-off scores established by each of the teams are 99%, 95% and 90%. This is in accord with expectations, as students at that level of performance tend to be consistent in their performance across the whole examination.

The greater variability for the Good and Satisfactory performance standards indicates that, at these levels, student performance is relatively inconsistent. Students' knowledge is not as broad, or as deep. Thus, some students do not perform as well as expected in easier items because they do not have a sufficient knowledge of certain aspects of the course. On the other hand, the same students may have a good understanding of other sections and, so, can perform above expectations on items based on these sections.

When this test was applied to the scores established by the English teams, the results shown in Table 7.2 were obtained. Unlike Mathematics, where a value of three or more signalled a difference in every item, the items in English had different maximum possible scores. In some items, where the maximum possible score was four, a difference of one between the judges' item cut-off score and the score obtained by a student would be recorded.

As was the case for Mathematics, the item scores obtained by those students who achieved one of the cut-off scores were similar to the item scores set by the judges for that cut-off score. For all three borderlines, the item scores of at least 90% of students

scoring one of those values had a "difference" from the item scores set by the judges in four or fewer items.

TABLE 7.2

Proportions of Borderline Students Demonstrating Item Score Variability in 1995 HSC English

Cut-off	Team 1		Team 2`	
Satisfactory	53		51	
Differences	Freq.	Cum.	Freq.	Cum.
0	0.027	0.027	0.014	0.014
1	0.119	0.189	0.119	0.133
2	0.331	0.520	0.288	0.421
3	0.286	0.806	0.332	0.753
4	0.112	0.918	0.147	0.900
5	0.051	0.969	0.074	0.974
6	0.026	0.995	0.017	0.991
7	0.003	0.998	0.004	0.995
8	0.002	1.000	0.005	1.000

	Team 1		Team 2	
Good	74		71	
0	0.030	0.030	0.044	0.044
1	0.162	0.192	0.238	0.282
2	0.352	0.544	0.328	0.610
3	0.275	0.819	0.228	0.838
4	0.143	0.962	0.114	0.952
5	0.030	0.992	0.039	0.991
6	0.006	0.998	0.008	0.999
7	0.002	1.000	0.001	1.000

	Team 1		Team 2	
Very Good	88		88	
0	0.070	0.070	0.076	0.07
1	0.322	0.392	0.304	0.38
2	0.341	0.733	0.298	0.67

For all teams, and in both Mathematics and English, the students' performance profiles are generally consistent with those established by the judges. Even at the borderline between Satisfactory and Unsatisfactory, around 80% of students obtaining the cut-off score had a difference recorded for three or fewer items. These results give support to the validity of the cut-off scores.

7.3.2.2 Comparability of the Cut-off Scores Produced by the Different Teams Working Within Each Course

In the second year of the study, multiple teams were used to determine cut-off scores for the examination on each course. A measure of the validity of the procedure, and the cut-off scores it produced, is the level of agreement between the cut-off scores produced by the teams. These values are shown in Tables 7.3, 7.4 and 7.5.

TABLE 7.3

Final Cut-off Scores for Mathematics in 1995

Standard	*Team 1*	*Team 2*	*Team 3*
Excellent	113	113	113
Very Good	104	99	100
Good	78	78	77
Satisfactory	47	50	49

TABLE 7.4

Final Cut-off Scores for English in 1995

Standard	*Team 1*	*Team 2*
Excellent	100	100
Very Good	88	87.5
Good	74	70.5
Satisfactory	53	51

TABLE 7.5

Final Cut-off Scores for Biology (Core Items Only) in 1995

Standard	*Team 1*	*Team 2*
Excellent	63.4	65.4
Very Good	52.9	54.7
Good	44.3	44
Satisfactory	34.8	34

These tables show that for all three courses, the cut-off scores determined by the teams are relatively close. For Mathematics, the greatest difference is between Team 1 and Team 2 at the Very Good/Good cut-off, where a difference of 5 represents 4.2% of the maximum possible score. For English and Biology, the greatest differences of 3.5 and 2 represent 2.9% and 2.7% respectively. Given that the teams worked completely independently, these results contribute further evidence to the validity of the procedure.

7.3.3 Evidence Based on Comparisons with External Sources of Information

Another way of validating either the standards set, or the process of equating the examinations by applying the same standards from one year to the next, is to use some form of measure or process outside and independent from the process employed by this study.

In relation to the standard-setting process itself, the fact that it shares some similar characteristics with other popular procedures provides supporting evidence of its validity. Many variants of the Angoff procedure have been proposed and used over the years, with researchers usually concluding that an Angoff-type approach generally gives

more valid results than other procedures (eg Brennan and Lockwood, 1980; Cross *et al*, 1984). The procedure used in this study builds upon the earlier approaches and incorporates other techniques designed to ensure the validity of the process, and hence the validity of the standards set.

To validate the use of the procedure for equating the examinations by linking standards across different years, sources of evidence external to the procedure were sought. In Chapter 2, the common approaches to performing this task were identified. The results of attempts to use these methods in this study are discussed below.

7.3.3.1 Using Common or Linking Items

One method often used for comparing performance standards over time where different versions of an examination are administered is to incorporate common or link items in each examination. By using the performances of the students in each year on the common items, it is possible to obtain an overall measure of the relative performances of the two groups of students.

This approach, however, was not possible in this study. The Higher School Certificate examinations on which this study was based are new examinations every year. The examinations are in the public domain after they have been administered, and it is common practice for students who will sit for the examination to use past examination papers to practise. In subjects like Mathematics and Biology, it is also possible to purchase worked solutions. It is policy that no specific item be included in examinations in two different calendar years.

7.3.3.2 Using Common Students

Traditional approaches to the use of common students to equate two examinations involve having a relatively large group of students attempt both examinations. Measures of student performance on the two examinations can then be used to create either a linear or curvilinear relationship between the scores on the two examinations. A number of methods have been developed in an attempt to overcome the need to use a large group of students in such an equating exercise. For example, Angoff's Design I (Angoff, 1971) involves selecting two groups of students at random from a larger group and administering one examination to each group.

Common student methods for equating two examinations using latent trait models have been developed. One advantage of such methods is that a much smaller number of common students can be used (Wright and Stone, 1979). By calculating the mean ability of the students on both tests, a simple relationship can be established which will enable the performances of other individuals or groups of students to be compared, irrespective of the examination they have taken.

During this study, consideration was given to seeking to administer the 1994 and 1995 HSC examinations to students from another educational system. Had this been done, the results of these students could have been used to calibrate the items from the two examinations, and to obtain a measure of the abilities of the students as measured by the two examinations. Once these ability measures were known a procedure could then have been used to put the two examinations onto a common scale by adjusting their item difficulties.

This approach is not without its problems. Only items examining content areas in which these external students had had sufficient exposure could be included in the examination. That is, the items included in the examinations would need to represent the same challenge to the external students as to the New South Wales students. As well as issues to do with the comparability of the curriculum, other factors - such as ensuring the use of standardised administrative procedures and determining whether all students involved will perform to their maximum potential in an exercise for which they receive no personal reward - need to be considered in such an approach. Nevertheless, this technique has been used successfully in other situations, and so is worthy of further consideration. It is suggested that such an exercise may be the focus of further research.

There were a number of students who, having sat for the 1994 examination, decided to "repeat" the course, and so sat for the 1995 examination. It was decided to consider the performances of these students in the two years to see what information might be provided about the relative performances of the two cohorts. If it is appropriate to equate the two examinations by using these common students, it would be possible to obtain evidence related to the accuracy of the procedure used in this study.

The methods for equating examinations using common persons must be careful to ensure that there is nothing in the process that can distort the results. For example, when a group of students is given both examinations (Form X and Form Y), it is usual for half the group to do Form X followed by Form Y, and the other half to attempt the examinations in the reverse order. Such an approach nullifies the effects of fatigue.

The issue that needs to be addressed here is whether repeat students will tend to improve their performances over the two years. If their performances remain stable over this time, it would seem reasonable to use these common students to equate the examinations.

Taylor (1979) studied the performances of nearly 300 students who, having sat for the NSW HSC examinations in 1976, did not perform well enough to be offered a place at university. These students then took the option of repeating all courses at a College of Technical and Further Education (TAFE). He found that the majority of students (73%) tended to improve their overall performances, but that the extent of this improvement varied. For example, of those students in the first quartile (0-25%) in 1976, 46% remained in that group, while 23% improved sufficiently to be placed in the third quartile (50-75%). Overall, 42% improved their performances sufficiently to be accepted into university on the second attempt. Taylor does not report whether students whose performances deteriorated offset these improvements. Given the data reported below, although collected some fifteen years later, it is unlikely that this is the case.

Smith (1994) surveyed students from three Australian states - NSW, Queensland and Western Australia - who had completed their secondary school studies in 1991, had applied for admission to a university but had been unsuccessful in gaining a place. He found that of those who elected to repeat their Year 12 studies, over 93% reapplied for university in 1992. Of these, over 90% gained admission. Smith notes that these data do not tell whether these students just failed to gain entry in 1991 and so, may have changed some of the courses they took in 1992 to make it easier to gain entry. The data also do not indicate what part added maturity and motivation, and a greater awareness

of the system, played in the students' improvement. Smith also reports that the students who responded to his survey generally came from high socioeconomic backgrounds, attended non-government schools, lived at home, were probably more motivated and interested in education, and came from homes where English was the main language.

In spite of the shortcomings of this study due to the sample of students who responded to the survey, it would seem that repeat students tend to improve their performances over the two examinations. This improvement would seem to limit the use of common-person equating methods as a means of validating the equating performed using professional judgment.

Henriksson (1993) reports that students who repeat the Swedish Scholastic Aptitude Test (SweSAT) tend to improve their results on all sub-tests, and that the amount of improvement can vary from sub-test to sub-test. If this is the case, variations in standards of performance of repeating students will be hard to separate from variations in the difficulty of the examinations and the stringency of the scoring key.

In spite of the apparent problems with the use of repeat students, if some other measure of student performance, largely independent of the course being considered, can be used, it may be possible to quantify the relative difficulties of the two examinations.

Students seeking the New South Wales HSC credential can repeat individual courses if they wish to improve their overall results in the total program. That is, they can "top up" their results by repeating only one course. To ensure that data from the two years were as comparable as possible, only the results of those students who had presented at

least five courses in both years were included in the comparison of the 1994 and 1995 examinations. This was done in an attempt to eliminate the possible inflationary effects of students doing much better in the second year by only presenting for a single course. A Tertiary Entrance Score (TES) is calculated for students who present for at least five courses. This index is used as a measure of these repeating students' relative overall performances in the two years. The TES comprises the weighted course scores for every course the student has presented for. It has a maximum possible value of 500 and is a measure of a student's achievement in his/her total program of study.

Table 7.6 shows the mean HSC examination marks scored by those students who sat the Mathematics examination in both 1994 and 1995. These values have been converted to percentages to enable comparisons with the mean Tertiary Entrance Scores.
The difference in mean scores (expressed as a percentage) of repeat students in Mathematics is 13.8 higher in 1995 than in 1994. The Tertiary Entrance Score mean in 1995 is 44.5 higher than in 1994. This difference is 8.9 when converted to a score out of 100.

TABLE 7.6

Mean Examination Marks as Percentages and Tertiary Entrance Scores of Repeat Mathematics Students

Maths(/100)		*TES(/500)*	
1994	**1995**	**1994**	**1995**
43.3	57.1	215.15	259.65

Hence, consistent with Henriksson's findings, the repeat students, as a group, have improved both their overall performances and their performances in Mathematics. Across their whole program, the repeat students improved on average by 8.9 marks per

course, whereas in Mathematics they performed slightly better, improving on average by 13.8 marks.

A similar analysis of the English examination provided the results shown in Table 7.7. The difference in mean scores (expressed as a percentage) of repeat students in English is 2.23 higher in 1995 than in 1994. The Tertiary Entrance Score mean is 47.88, (or 9.58 if expressed as a percentage), higher in 1995.

TABLE 7.7

Mean Examination Marks as Percentages and Tertiary Entrance Scores of Repeat English Students

English(/100)		*TES(/500)*	
1994	**1995**	**1994**	**1995**
51.55	53.78	212.85	260.73

The repeat students in 1995 have tended to improve on their performances in the 1994 examinations across their total program. These results are consistent with the findings of Taylor (1979), Smith (1994) and Henriksson (1993).

Henriksson (1993) also found that repeat students tend to improve more in mathematics-based sub-tests of the SweSAT than in verbal-based sub-tests. The results provided above are consistent with this finding. Hence, given that the evidence seems to indicate that, as a group, students repeating an examination like the HSC will tend to improve, and do so differentially in various courses, it would appear that using these common students to equate the two examinations is not viable.

Given the major changes to the Biology course and examination paper in the two years, it was decided not to conduct a similar analysis of the data for that course.

7.4 REPLICATION OF THE EQUATING PROCEDURE

Examinations conducted in 1996 in Mathematics, English and Biology were equated with those from 1995 in order to seek further evidence regarding the validity of the procedure.

Once again, two teams of judges were created for each course. There was only one major difference between the way the procedure was applied in 1995 and in 1996. The judges were given the descriptor statements, the 1995 examination paper and a sample of student scripts which had been awarded the 1995 cut-off scores, as had been the case when they were equating the 1994 and 1995 examinations. In addition, however, they were told the cut-off scores established for the 1995 examinations and were given the key used in scoring the 1996 student responses.

Although from the results of equating the 1994 and 1995 examinations it would appear that the judges can perform the task quite satisfactorily, it was decided to provide this additional information to the judges. This was done in an attempt to make it easier for judges to internalise the standards to be applied when equating the 1995 and 1996 examinations.

The judges were advised that they should use this information in conjunction with the other information and materials they had been given, and not simply take a "short-cut"

and use the 1995 values. The statistical feedback for each course was provided to the judges in the same manner as before, and they were given student scripts to consider.

The examinations were equated using the same procedure as that used to equate the 1994 and 1995 examinations. The initial cut-off scores established by the teams for each course are shown in Tables 7.8(a), 7.9(a) and 7.10(a). The related tables, Tables 7.8(b), 7.9(b) and 7.10(b), show the final cut-off scores determined by each team and the proportion of the total candidature within each standard level based on the use of these final cut-off scores.

In ordinary circumstances, it would be usual to equate the 1996 examination to the 1994 examination, as it was the one on which the standards were originally set. Perhaps the 1995 materials could be used to provide confirmatory evidence. Given the changes to the Biology course and examination between 1994 and 1995, however, it was decided to equate the 1996 and the 1995 examinations for all courses.

7.4.1 Mathematics

TABLE 7.8(a)

A Set of Cut-off Scores from the 1995 HSC Mathematics Examination and the Initial Cut-off Scores for the 1996 Examination

Borderline	*1995*	*1996 Team 1*	*1996 Team 2*
Excellent/Very Good	113	113	111
Very Good/Good	100	100	99
Good/Satisfactory	78	75	75
Satisfactory/Unsatisfactory	48	49	47

TABLE 7.8(b)

Final Cut-off Scores Set for the 1996 Examination and the Cumulative Proportions of Students in Each Standard Level in 1995 and 1996

Borderline	*1995*		*1996 Team 1*		*1996 Team 2*	
	Score	**%**	**Score**	**%**	**Score**	**%**
Excellent/Very Good	113	1.7	113	0.8	111	1.5
Very Good/Good	100	16.2	101	9.6	101	9.6
Good/Satisfactory	78	46.1	75	42.5	76	41.1
Satisfactory/Unsatisfactory	48	78.2	49	73.2	48	74.2

7.4.2 English

TABLE 7.9(a)

A Set of Cut-off Scores from the 1995 HSC English Examination and the Initial Cut-off Scores for the 1996 Examination

Borderline	*1995*	*1996 Team 1*	*1996 Team 2*
Excellent/Very Good	100	101	101
Very Good/Good	88	83	86
Good/Satisfactory	71	67	70
Satisfactory/Unsatisfactory	53	48	53

TABLE 7.9(b)

Final Cut-off Scores Set for the 1996 Examination and the Cumulative Proportions of Students in Each Standard Level in 1995 and 1996

Borderline	*1995*		*1996 Team 1*		*1996 Team 2*	
	Score	**%**	**Score**	**%**	**Score**	**%**
Excellent/Very Good	100	1.6	101	0.3	100	0.3
Very Good/Good	88	9.6	82	6.8	86	4.0
Good/Satisfactory	71	40.8	67	32.8	70	25.4
Satisfactory/Unsatisfactory	53	80.7	49	79.2	50	77.2

7.4.3 Biology

TABLE 7.10(a)

A Set of Cut-off Scores from the 1995 Biology HSC Examination and the Initial Cut-off Scores for the 1996 Examination

Borderline	*1995*	*1996 Team 1*	*1996 Team 2*
Excellent/Very Good	87	87	87
Very Good/Good	71	74	72
Good/Satisfactory	59	60	55
Satisfactory/Unsatisfactory	45	46	40

TABLE 7.10(b)

Final Cut-off Scores Set for the 1996 Examination and the Cumulative Proportions of Students in Each Standard Level in 1995 and 1996

Borderline	*1995*		*1996 Team 1*		*1996 Team 2*	
	Score	**%**	**Score**	**%**	**Score**	**%**
Excellent/ Very Good	87	3.4	88	1.8	88	1.8
Very Good/ Good	71	30.6	75	21.3	74	21.3
Good/ Satisfactory	59	55.1	60	52.9	59	54.8
Satisfactory/ Unsatisfactory	45	77.0	44	77.6	43	78.7

In the case of Mathematics and Biology, there are some relatively small differences between the two years in the proportions of students in each standard level. For Mathematics, this occurs at three borderlines, whereas for Biology the only difference of any note is at the Very Good/Good borderline.

A Kolmogorov-Smirnov two-sample test rejects, on the Mathematics data, the null hypothesis of no significant difference between the performances of the 1995 and 1996 candidatures. Examining the data leads to the conclusion that the 1995 group's performance is probably superior to that of the 1996 group. Performing a similar test on the 1994 and 1996 cohorts again rejects the null hypothesis. In this case, however, it

seems that the more able candidates in 1994 may have performed slightly better than the similar group in 1996, but that a higher proportion of students in 1996 were considered to have reached the satisfactory standard or above.

When a similar Kolmogorov-Smirnov test was performed on the 1995 and 1996 Biology data, the null hypothesis was also rejected. This was primarily due to the higher proportions of students in the Excellent and Very Good categories in 1995. The proportions of students classed as Satisfactory and above in 1995 and 1996, show virtually no difference.

In the case of English, there was a change to the way in which scores were calculated for the 1996 examination from that used in 1995. It was decided by the examining authority to place on the same scale the results achieved by the students in the English course used in this study (course B) and the results achieved by students who had taken the more difficult English course (course A). To do this, two common items were included in what were otherwise two different examinations. The discrete items in the two examinations were scored separately using the same type of approach as in past years. The common items, however, were scored using the same key for both courses. The students of course A tended to perform better in the common items.

The distributions of marks obtained by the candidature of each course were adjusted, using an equipercentile approach, to have the same distribution shape as that candidature had received on the common items. After this adjustment, the final distribution of scores obtained by the students in the English course that is the focus of this study (course B) was less spread than in previous years. The judges setting the cut-

off scores for the 1996 examination were unaware of the impact of this scaling process, and so determined their item cut-off scores in accordance with the materials exemplifying the standards applied in 1995. The impact of the equipercentile scaling was also not reflected in the data used to create the statistical feedback to the judges, nor would it have been readily apparent from the sample of scripts provided. This scaling of the scores almost certainly contributed to the lower proportions of students in the Excellent, Very Good and Good categories in 1996.

To undertake such a comparison it would be necessary to examine a distribution of 1996 examination scores of the students in course B prior to the adjustments that were made. Even if that information had been available, it must be noted that those items in common with the course A examination paper were scored using a more stringent key than would have been applied to the corresponding items in the 1995 course B examination paper. So, no firm conclusions can be drawn about the relative performances of the groups in 1995 and 1996.

The cut-off scores proposed by the two teams in 1996 reported in Table 7.9(b) show that the teams were in very close agreement for the cut-off scores for the Excellent/Very Good and Satisfactory/Unsatisfactory borderlines. However, at the Very Good/Good borderline there was a difference of four marks and a difference of three marks at the Good/Satisfactory borderline. While these differences are not large given the maximum possible score is 120, it would be hoped that the two teams would be able to get within one or two marks of each other on all values. There was a relatively slight change made to the structure of the examination paper in 1996 to accommodate the common items for the course A and course B examinations. It is possible that this change and the sample

student scripts from 1995 at those borderlines made it a little more difficult for the judges to determine their cut-off scores. By comparison, Tables 7.8(b) and 7.10(b) show that the two teams in each of Mathematics and Biology were able to arrive at very similar cut-off scores for all borderlines as their colleagues.

From a consideration of the results for each subject, it is not evident whether providing the judges with the scores awarded to the student scripts in the standards package and the scoring keys has made any difference to the accuracy of the decisions made. Provided that the judges use such information in conjunction with other data and follow the procedure as required, it would seem reasonable to continue to provide it. When setting the cut-off scores for the 1996 examinations, the judges were aware of the cut-off scores established for the 1995 examinations. However, it is evident from the 1996 cut-off scores, and from the way the judges went about the process of establishing these scores, that they did not simply copy the 1995 values.

When establishing the cut-off scores for the 1996 examinations, the judges used materials from the previous year. This has the potential problem that, over time, any slight errors introduced during one equating process will be passed on, and even compounded. To minimise the chance of such an outcome, it would seem a sound approach would be to provide the judges with materials relating to the year when the performance standards were first established, in addition, to the materials relating to the previous year. The provision of this material should strengthen the process.

The results obtained from using the procedure to establish cut-off scores on the 1996 examinations provide evidence that supports the claim that the procedure is valid. The

different teams of judges for each subject, after using materials encapsulating the standards used for the 1995 examinations and applying the procedure independently of each other, arrived at similar cut-off scores.

7.5 VIEWS OF THE JUDGES

Following the last step taken by each team in 1995, individual judges were asked to complete a questionnaire aimed at giving them the opportunity to express their views on the process and make suggestions for improvement. Throughout the process, the judges had made comments that were noted, but it was considered important that they be given the opportunity to reflect on the exercise as a whole and provide comment on an individual basis.

Two different questionnaires were produced. The judges who had been involved in the process in both years were given a questionnaire that sought their views on the process in 1994 (the initial year) and 1995 (the subsequent year). Those judges who had been involved in only the second year were asked the same questions as the original judges concerning the second year of the study. A copy of the questionnaire administered to the judges involved in the study during 1994 and 1995 is provided in the Appendix.

7.5.1 The Views of the Judges Concerning the Initial Year

Irrespective of the course with which they had been involved, the judges all expressed the view that they were satisfied with the outcomes of the process. They stated that being required to discuss differences until consensus was reached was a strength of the procedure. Some indicated that while calculating the average of their individual recommendations may have given similar results, it was far more meaningful and

satisfying to reach a point of agreement through professional discussion. This enabled them to listen to their colleagues and, perhaps, become aware of issues they might have overlooked in their own considerations. Continuing this discussion to the point where the team reached consensus gave them a feeling of having a shared position in which they had confidence.

The judges claimed that other teachers of their course would be able to understand the performance standards they had established. They did point out, however, that as this way of judging and reporting student achievement in the NSW Higher School Certificate examinations would be new to many teachers, care would need to be taken in explaining the process used and what it implied about student achievement.

The strengths of the procedure identified by the judges included the emphasis on the use of professional judgment in the process of negotiation and decision making, the support that the procedure gave them in focusing on the achievements of students, and the consideration of student scripts.

The concerns identified related mainly to the fact that the sample of student scripts they reviewed showed that two students who had gained the same total score in the examination could often have quite different patterns of scores for the items. While this helped them to appreciate the nature of student performance, it added some difficulty to the process of establishing cut-off scores. Some judges also indicated that the descriptor statements, while adequate, would have benefited from further review and refinement.

7.5.2 The Views of the Judges Concerning the Subsequent Year

The members of the original teams who responded to the questionnaires claimed that, having been involved in the process in the initial year, they had no difficulty in using the materials provided to refamiliarise themselves with the performance standards to be applied. Opinion varied as to which part of the standards package was most useful. Some indicated that the descriptor statements were most effective in helping them, others stated that it was the student scripts. Even amongst members of the same team, there was variation in what the judges found to be most useful in internalising the standards. This could well be a strength of the procedure, in that it may result in judges bringing different perspectives to the discussion process.

A number of judges, particularly those involved with Mathematics, commented that the scoring key and the scores awarded to each item on the sample scripts would have been particularly useful. It is quite likely that, for some judges, the sample scripts may have provided even greater support had this information been supplied.

Judges from the newly created teams expressed the view that it was relatively simple to understand the standards that were to be applied from the materials in the standards package. Once again, however, claims were made that having the scoring key and item values awarded to the sample scripts would have provided assistance.

Irrespective of whether the judges had been involved in the exercise in the initial year, most respondents indicated that all stages of the process were important. Each stage gave them the opportunity to reflect, discuss and refine earlier decisions.

The judges' reactions to the statistical data varied. Members of the original teams were generally very comfortable with it. Some of the judges in the new teams, however, admitted to some uncertainty when first shown this information. Once they came to understand it, they were prepared to take into account what it indicated, particularly after they saw that it generally matched their own judgments. Overall, there was agreement that these data were an effective form of feedback and an important ingredient of the procedure.

As had been the case in the initial year, the view was expressed by several judges that the sample of student scripts they were given to consider may have been more useful if they had been closer to the profile of item scores typical of students at the borderlines. The judges indicated that "out-of-character" performances on some items by borderline students made it harder to obtain an unambiguous image of the capabilities of such students.

Opinion varied as to what the judges considered to be the strengths of the model in equating the examinations. Some identified as the main strength the structured approach, which made use of a variety of different forms of information; others felt that the discussion and refinement of original decisions were the strength. Some judges identified the use of experienced teachers with a strong understanding of the examination and the capabilities of the students as the main strength of the procedure. Clearly, all of these contributed to the effectiveness of the procedure.

The judges indicated that a number of potential problems might need to be overcome when seeking to compare the performances of groups of students across different years.

These include significant changes in the nature of the candidature, changes in the structure of the examination, large variations in the difficulty of the examination from one year to the next, and the potential for individual judges to make very different interpretations of the standards to be applied. The judges commented, however, that provided the procedure is applied carefully, each of these problems can be overcome, or at least their impact minimised.

When given the opportunity to make further comment, a number of respondents expressed the view that they had gained personally from participating in the exercise through the opportunity to engage in discussion with their colleagues about issues of student performance in their course.

7.6 SUMMARY

The evidence presented in this chapter supports the claim that the procedure employed in this study can produce reliable results. The emphasis on discussion and review, supported by the data on student performance provided through the use of the statistical model, and the focus on consensus, seem to have been effective in delivering an acceptable degree of inter-judge reliability.

Several different forms of validity evidence are provided. These include those associated with the suitability of the judges and the fidelity of the process, those associated with the item score profiles of the cut-off scores, and those associated with the use of multiple teams to set the standards for an examination. Other possible approaches to gathering information to test for validity are discussed. From the

information available, it can be claimed that the performance standards that are created have an acceptable degree of validity.

The replication of the study to equate the 1995 and 1996 examinations provides further confirmation of the validity of the procedure. Setting aside the results for English due to the changes in calculating the scores introduced in 1996, the procedure has given results for the equating which are consistent with expectations. The variations in the proportions of students at each standard level between the two years would seem to be due to real differences in the performances of the candidatures.

The comments made by the judges in response to the questionnaires were generally similar to those made during the application of the procedure. The opportunity, however, to reflect upon the whole process led to some useful statements. The procedure used had strong support from some judges and qualified support from others. Even those who expressed qualified support indicated that the procedure had potential as a means of setting standards and equating examinations across different years.

CHAPTER 8
SUMMARY AND CONCLUSIONS

8.1 SUMMARY OF THE STUDY

An important issue of concern for those with the responsibility for assessing and reporting student achievement in examinations using a standards-based approach is addressed in this study. Namely, once valid and reliable standards of performance are set, can these standards be applied by a team of judges to equate different examinations administered at different times, so that judgments can be made as to whether the performance standards of the different groups of students are the same?

The usual approach to equating two examinations is to use common or link items, or employ designs that use common students. Once two examinations have been equated, it is possible to compare the performances of students who have taken the different examinations. Thus, questions relating to whether standards of student performance are changing or static can be easily addressed. Such well-established procedures, however, can not be used in all situations. In many high-stakes, curriculum-based examinations it is not possible to equate different versions of an examination by using link items, or by having the same students attempt both examinations. In such circumstances, it is common to use an entirely new version of the examination paper every year, so that no item is ever used in more than one examination. This type of examination often consists of a variety of different item types, and at least some of the items have more than two response categories. In fact, it is quite common for the majority, or indeed all, of the items to be scored polytomously.

The approach used in this study was to first use teams of suitably trained judges to set cut-off scores corresponding to the borderlines between different standards of performance on examinations. This involved the development of a procedure that can be used to set cut-off scores, corresponding to a number of different standards of performance, in large-scale public examinations. Importantly, this procedure was designed to give acceptable results when applied to examinations that contain different item types, including a variety of items that are scored polytomously.

Once these standards were developed, the key issue - whether teams of judges can equate the original examination and a new one by applying the same performance standards to the new examination - was addressed. If the procedure can be used for this purpose, then the question of whether there are any differences between the performances of groups of students who have studied the same course, but who have undertaken different examinations, can be answered.

The context for this study is the NSW Higher School Certificate. The examinations used are comprehensive measures of the achievements of students in courses of study based on traditional areas of knowledge, such as Mathematics, History and Music. A new examination is prepared and administered every year. Most items are polytomously scored, and in many examinations students have a choice of items based on the optional sections of the course they have studied. While these examinations often contain a number of multiple-choice or other objective items, they consist predominantly of different forms of extended-response items, such as those requiring an essay response, or a detailed solution to a mathematics problem. Some examinations also require students to submit a project, or deliver a performance, or participate in a

conversation in a foreign language. This study uses the examinations administered at the end of 1994, 1995 and 1996 in the courses of Mathematics, English and Biology.

In order to address the problem posed in the study, a review of the literature on equating examinations and on setting performance standards on examinations was conducted. Techniques that had been used successfully in the past were considered to be a good starting point for developing procedures which would suit the circumstances of this study. Various methods for equating different forms of an examination were considered. Most of these methods are based on statistical designs that involve, in various ways, either items which are common to both examinations or groups of students who attempt both examinations. Such techniques are of no use in the situation where an entirely new examination paper is administered every year. The third category of equating techniques, namely approaches where teams of judges are used to make the decisions, appeared to be the only one which offered any promise in these circumstances. The review of the literature on standard-setting procedures reported in Chapter 2 also shows that, given the nature of the examinations being considered and the way they are administered, the most suitable techniques for setting standards are those based on the use of panels of judges to establish cut-off scores. Account was also taken of the various criticisms of standard-setting practices made in the past. As a result of the review, it became clear that the procedures which offer the flexibility required, both for setting performance standards and equating different forms of an examination, are those based on the use of informed professional judgment.

Building on the findings from the review, Chapter 3 puts forward a procedure to equate examinations by first, establishing a set of standards of performance in each course and

then, applying those same standards to determine cut-off scores on a subsequent examination. The procedure is an adaptation of the Angoff (1971) procedure, and allows for a team of suitably qualified and trained judges to set cut-off scores relating to standards of performance. Performance standards are defined, and described in terms of the types of knowledge and skill possessed by students who have achieved each standard. The standards are further clarified by using samples of examination responses produced by students at each standard in the course. Once this is done, the procedure can be used to set cut-off scores corresponding to those standards on entirely different examinations of the course.

The procedure is sufficiently flexible to enable cut-off scores to be set in examinations involving a variety of different item types that are scored polytomously. The procedure also employs a multi-stage approach, where the judges are given the opportunity to review and refine their earlier decisions as a result of discussion and feedback.

A review of some aspects of modern measurement theory is included in Chapter 3. In particular, the Simple Logistic Model and the Extended Logistic Model, which are Rasch-based measurement models, are discussed. It is considered that these models can provide an effective form of statistical feedback on the performance of students of different abilities in the various items in the examination. Such techniques can provide rich information on the performance of students on items, and offer the potential to uncover aspects of student performance that the judges may not otherwise discover.

The application of the procedure to set performance standards, and then to equate examinations administered in different years, is discussed in Chapter 4. The use of the

procedure with the examinations set in three different courses from the NSW Higher School Certificate is explained. The manner in which each step of the procedure is applied is revealed.

The results of using this procedure to set performance standards in the initial year of the study are provided in Chapter 5. The way in which the judges in the team that was created for each course went about their task, and the results obtained, are presented. In the subsequent year of the study, at least two teams independently used the procedure to equate the examinations. The results obtained by each team are reported.

In Chapter 6, the results reported in the previous chapter are discussed. Various comparisons are made between the cut-off scores proposed by the teams at different stages of the process. Comparisons are also made between the decisions of the different teams in the second year of the study. On the whole, the decisions made by the judges are quite close. This suggests that suitably trained judges are able to use this procedure to establish comparable cut-off scores. The issue of how well the teams of judges are able to apply the same performance standards when determining cut-off scores in different years is also discussed. Taking account of certain circumstances which existed at the time (changes to the student population between the different years and changes to the course content and structure of the examination paper in the case of one course), the results are promising.

Questions of whether the process can be used to obtain reliable and valid standards of performance are addressed in Chapter 7. Different tests are applied and different

statistics calculated. The outcomes of attempts to use the performance data of students who sat for both the 1994 and 1995 examinations are reported.

The application of the procedure to equate the examinations administered in 1996, in each of the three courses, with the examinations from the previous years is also discussed. In this regard, once allowance is made for certain situations outside the control of the study that nevertheless affected it (such as an adjustment of scores in one course by the credentialling authority), the results can be considered encouraging. The application of the procedure to the 1996 examinations provides results that support those obtained in the earlier years.

In the final section of Chapter 7, comments made by the judges in response to a survey are summarised. On the whole, the comments made are very supportive of the procedure used. While some comments suggested different ways in which the procedure could be strengthened, the judges nevertheless claimed that the process they had followed enabled them to set cut-off scores appropriate to the standards of performance.

8.2 IMPLICATIONS OF THE STUDY

The results obtained in this study from using the procedure to equate examinations are encouraging. They indicate that the procedure can be used by suitably qualified judges to set cut-off scores that accurately reflect a set of performance standards on large-scale, curriculum-based examinations containing a variety of items that are scored polytomously. The level of agreement between the teams establishing cut-off scores for a course, and the agreement between the cut-off scores set and the proportions of

students within each performance standard in the different years of the study, show that judges are able to internalise the performance standards to be applied. Furthermore, given the fact that the level of agreement between the judges in a team improved throughout the process, it can be concluded that the procedure was successful in assisting judges to refine and improve their initial decisions.

Even though the results are satisfactory, there are a number of changes, including modifications to the way the procedure is used, which should be considered in any further application of the procedure. Such changes, while relatively minor, should lead to an even greater measure of comfort in the results.

8.2.1 Duration of the Application of the Procedure

In this study, circumstances beyond control meant that the procedure was applied over a lengthy period of time. In most cases, the judges established their individual cut-off scores soon after the examinations were administered. The first meeting, when they discussed their individual decisions, was usually held a week or two later. It was then, generally, some three or four months before the judges came together again to consider the statistical data and the sample of student scripts. By this time, the "sense" of the standards of student performance they had developed was dim and time was needed to build up this understanding again. It would clearly be more effective, and would presumably lead to better decision making, if the procedure was applied over a far shorter period of time. In this way, the judges would retain a much clearer idea of the examination items and the standards of students' responses. A timeframe of one or two days would seem sufficient for the judges to set the initial cut-off scores, discuss and reflect on them, and finalise their decisions.

If this procedure were to be used operationally, with the judges relieved from other duties for this period and with sufficient resources and support mechanisms provided, it would be quite possible to perform this task over such a timeframe.

8.2.2 Size of the Teams and Background of the Judges

A number of the studies reported in the literature use quite large teams to set cut-off scores. Often these teams consist of people from quite diverse backgrounds. Disagreements are discussed, and then averages of the judges' revised item cut-off scores are calculated.

In this study, relatively small teams were used; however, the judges selected were deliberately chosen because of their experience in preparing students for the examination. Many of the judges involved in the study had also been involved in scoring the student scripts. The use of large heterogeneous teams may be suitable for examinations which test general knowledge or basic skills. It is important when working with curriculum-based examinations where the knowledge and skills are more specialised, however, that the judges have a thorough understanding of the course material and are able to provide quality responses to the examination items themselves. If they cannot do this, their capacity to contribute to the team is limited. Considering the outcomes of this study, and observing the application of the procedure, it is suggested that the optimum size for a team applying this procedure is five or six. A team of this size permits thorough discussion enabling all judges to air their views, while ensuring that a range of opinions is provided. It is also important that the judges

understand the full range of performances typically demonstrated by students in the course.

A particular problem arose in this study in relation to Biology. Even though the judges used were experienced and well qualified for the task, as the examination gave students a relatively large range of optional items from which to choose, some of the judges admitted that they did not believe they had sufficient knowledge of the course content on which some of the optional items were based. As a result they declined to express an opinion on the appropriate cut-off scores for these items.

To overcome this problem, it is necessary to ensure that the team has sufficient knowledge of all aspects of the course. If, due to the extent of the range of course options, this could only be achieved by making the team too large and unwieldy, strategies need to be employed to enable people with expertise in these areas to provide appropriate advice to the judges. The issues concerning the most suitable way of handling optional items in a standard-setting context need to be further investigated.

8.2.3 Defining the Standards for the Judges as the Initial Step

In this study, when the teams of judges were required to set cut-off scores related to different standards of performance in the initial year, they were not given any detailed descriptions of what these different standards implied about the knowledge and skills students possessed. Instead, the judges were simply advised to establish five broad categories of examination performance typically achieved by students in their course and relate these to profiles of students they had taught. The labels of Excellent, Very Good, Good, Satisfactory and Unsatisfactory, while open to wide interpretation, gave

the judges some limited indication of where these standards of performance should be pitched. The fact that even the most experienced judges will vary in their individual opinions of what constitutes an excellent performance is not major cause for concern. Provided the judges are all familiar with the content of their course, and the range of performances produced by students in the examinations, it is not difficult for experienced judges to reach agreement, through discussion, on the profile of knowledge and skills of students at different levels of performance.

An alternative approach would have been to provide the judges with a more detailed profile of students at each of the performance standards. Using an approach more akin to that used in the subsequent year of the study, it would have been possible to develop descriptions of the knowledge and skills usually demonstrated by students at these different standards and exemplify them with samples of student work. The evidence from the second year of this study would indicate that, were this approach to be used in establishing their cut-off scores, the judges' initial cut-off scores would probably be closer to their final values than those resulting from using the more open approach.

There may be circumstances where this second approach would be preferable. For example, if it is possible that without strong direction the judges may develop a performance scale which is too stringent, then it would be wise to remove this risk. On the other hand, if the judges are well attuned to the range of standards of performance of students, and provided that steps are built in to help them to refine their initial decisions, the former approach can be effective. The judges will tend to have a greater acceptance of standards of performance that they have set themselves.

8.2.4 Training and Briefing the Judges for their Task

Circumstances did not permit the step of bringing all the judges in a team together for an initial briefing and training session prior to beginning their task. Instead, notes were prepared for the judges to follow when they were determining their individual cut-off scores, and discussions were held with judges individually to ensure they understood the procedure they were to follow.

While this approach had limited detrimental effect on the appropriateness of the cut-off scores the judges finally set, problems such as those initially encountered with the second Mathematics team in the subsequent year of the study may have been prevented had the judges been briefed together. Where possible, it would be highly desirable to bring the judges together prior to the commencement of the process. At such a meeting they would be instructed in the process they were to follow, and given some samples of student work to grade and discuss. This step would have the effect of giving the judges greater confidence in applying the procedure, particularly when they were working on their own in the initial stage, and probably save time in the early stages of the discussion process. As much of the validity of the standard-setting procedure comes from the rigor and uniformity of the procedure itself, it is an advantage if all judges follow exactly the same approach.

During the process it is also important that the judges do not let themselves fall into the habit of making quick compromises on the item cut-off scores, simply to avoid thorough discussion. Through the various discussion stages, the judges build up a comprehensive, shared "image" of the standards of student performance in their course. When this procedure is applied properly, judges are required to work hard in performing

their tasks. Observations made during the study indicate that the "pay-offs" from this effort on the part of the judges are greater confidence in the outcomes and an increased level of credibility.

8.2.5 The Use of the Rasch Measurement Model

Using the Extended Logistic Model to provide statistical feedback to the judges to assist them in reviewing and refining their decisions, seems to be an effective strategy. The evidence provided in the body of the report indicates that the judges were often prepared to adjust their initial decisions after considering this information.

The use of a latent trait model in this manner is quite different from the way it is usually used in the assessment and reporting of student performance. Traditionally, latent trait models are used to equate examinations by the use of link items or pre-calibrated items. They are also used in the development of a measuring instrument by providing feedback on the results of trialling, thus assisting in the improvement or selection of individual items. This in turn leads to improvement in the total instrument.

In this study the latent trait model performs more of an informing role, rather than a true measurement role. The Extended Logistic Model is used to analyse the results from an examination which has already been administered, not to develop the measuring instrument itself. There will be cases where the student performance data do not show fit to the model. This is not surprising, nor, provided the information produced is used in an advisory capacity, is it a cause for concern. Throughout this study, the judges were told that the decision as to what weight they should place on the data provided from the analysis was theirs.

The use of the latent trait model with the type of examinations included in this study, and in this manner is new. The nature of the examinations, including the item score ranges and variation in item types, means that the model is being used to analyse and report on performance data in a way that has not been attempted in quite the same way in the past. Only in very recent times have attempts been made to use the results of latent trait analyses to assist in providing information about student examinations similar to the type encountered in this study. More research is required in this area. Used in this way, the Extended Logistic Model strengthens the process of establishing and linking standards of performance. Further studies could investigate the use of the model with a wider range of examinations, especially those with a small number of items which may be awarded a wide range of possible item scores. The issue of whether there are better ways to deal with optional items in the examination also needs further study.

8.2.6 The Use of Samples of Student Scripts

An important step in this procedure is to provide the judges with a sample of the examination scripts of students who achieved the cut-off scores. This activity is integral to the equating procedure, both at beginning, when developing an understanding of the standards to be applied to a new examination based on the performances of students in a previous examination, and at the end, when reviewing and confirming the cut-off scores proposed.

While the judges in this study realised the value of these student responses, comment was made that they might have been more helpful had they been selected differently.

The scripts provided to the judges were chosen simply because they had received the same total score as one of the cut-off scores set by the judges. In the second year of the study, in particular, the judges stated that they sometimes found it difficult to appreciate the performance standards they were to apply in determining a cut-off score if a student whose script they were given did unexpectedly well or poorly in some items. To overcome any concern in this regard, it would seem wise to provide the judges with a sample of scripts where the scores awarded for each item are the same as, or close to, the individual item cut-off scores proposed by the judges. If time did not permit an exhaustive search for student responses which matched this pattern, an alternative, although probably less satisfactory, approach would be to provide student scripts which matched this pattern, but consisted of the individual item responses of a number of students.

8.3 CONCLUDING REMARKS

The results obtained in this study support the contention that, by following the structured procedure proposed, it is possible to use a team of judges to apply standards of performance established in an initial year using one examination, in order to set standards on a new examination which equate with those originally established. that and another examination by using these standards to determine appropriate cut-off scores on the new examination. Once the judges have internalised the performance standards they are to apply by using materials from the original examination, they then apply those standards to the subsequent examination using basically the same procedure as that used initially.

The procedure is sufficiently flexible to be used with a variety of different examinations, including comprehensive curriculum-based examinations containing a variety of different item types. The procedure used has a strong theoretical base and adapts techniques, used successfully in other circumstances, to suit the type of examinations encountered in this study.

APPENDICES

Note

In the following material the term "question" is often used in place of "item" and the term "mark" is used in place of "score". This has been done because the terms "question" and "mark" are more commonly used in describing the NSW Higher School Certificate program and so, will have been more meaningful to the judges involved in this study.

Appendix A: Sheet of Initial Written Instructions Given to Judges to Supplement Verbal Briefings

STANDARD-SETTING PROCEDURES – ADVICE TO JUDGES

You are part of a team chosen to establish and describe five levels of student performance in the course you are marking at the 1994 HSC examinations.

The purpose of the exercise is to develop and test a model for equating the performances of students in subjects presented for public examination across different years within the context of the current HSC examination procedures.

There are a number of steps to be followed in this process, some of which you will do individually, some will be done as a team. You will be asked to make notes and some recordings at various stages of the process on the sheets provided. I will be present to coordinate the team sessions (Steps 2 to 5), and I am available to assist with any questions prior to or during these steps.

You will be provided with a copy of the relevant 1994 HSC examination paper and a copy of the syllabus objectives and outcomes. You will be provided with other materials and information at the appropriate point in the process.

The steps in the process are as follows:

Step 1: ***Making Individual Decisions***

Working independently each team member will go through the examination paper and determine what minimum mark a student would need to score for a question (an item) in order for their performance in that question to be classed as

excellent
very good
good
satisfactory.

At this stage the determination of what makes a response *excellent, very good, good* and *satisfactory* is left up to you.

You are asked to record your cut-off marks on the sheet provided as well as noting any comments you would wish to make.

Step 2: Comparing and Obtaining Agreement

At this step the team will be brought together and will compare the decisions each member has made. The members will discuss any differences in cut-off scores with a view to achieving a consensus decision for each question (item).

If an individual member wishes to change his/her decision on a particular question this new value should be recorded in the appropriate place on his/her original sheet.

The agreed values of the team will be recorded on a separate sheet along with any comments the team feels it wants to make.

Step 3: Reviewing Decisions on the Basis of Statistical Data

Some statistical data will be presented to the team. The team will be asked to review the proposed cut-off scores in the light of the information provided.

When, after discussion, the team decides to change a cut-off score it had originally set, it will adjust that value. If, on the other hand, the team wishes to retain its value it is free to do so.

Step 4: Checking the Decisions Against Student Responses

At this stage the group will be given some student responses which have been awarded the minimum examination score needed to be included in each level. The members of the team will satisfy themselves that the standard of work being presented by the sample of students is appropriate for students at the borderline of the two levels.

Adjustments can be made to the team's decisions if, when having reviewed student work, the team feels that the cut-off scores established at Step 3 are not appropriate.

Step 5: Describing the Levels of Performance

The team will next describe in terms of the objectives and outcomes of the course the general characteristics of the performances of students whose examination score would place them in each performance level in each level. These statements will attempt to describe the performances, across the total examination, of students placed in each level.

Appendix B: Sheet Used by Judges to Record their Individual and Revised Cut-off Scores - Mathematics 1994

NAME:

QUESTION	*MINIMUM (OR CUT-OFF) MARK REQUIRED TO OBTAIN EACH LEVEL IN EACH QUESTION*							
	INITIAL INDIVIDUAL DECISIONS				*REVISED DECISIONS (IF APPROPRIATE*			
	Excellent	*Very Good*	*Good*	*Satisfactory*	*Excellent*	*Very Good*	*Good*	*Satisfactory*
Q1 (a) /2								
Q1 (b) /2								
Q1 (c) /2								
Q1 (d) /2								
Q1 (e) /1								
Q1 (f) /3								
TOTAL Q1								
Q2 (a) /1								
Q2 (b) /1								
Q2 (c) /2								
Q2 (d) /1								
Q2 (e) /2								
Q2 (f) /1								
Q2 (g) /3								
Q2 (h) /1								
TOTAL Q2								
Q3 (a) (i) /1								
(ii) /2								
(iii) /2								
Q3 (b) (i) /2								
(ii) /2								
Q3 (c) (i) /1								
(ii) /2								
TOTAL Q3								

QUESTION	MINIMUM (OR CUT-OFF) MARK REQUIRED TO OBTAIN EACH LEVEL IN EACH QUESTION							
	INITIAL INDIVIDUAL DECISIONS				REVISED DECISIONS (IF APPROPRIATE			
	Excellent	*Very Good*	*Good*	*Satisfactory*	*Excellent*	*Very Good*	*Good*	*Satisfactory*
Q4 (a) /3								
Q4 (b) /4								
Q4 (c) /2								
Q4 (d) (i) /1								
(ii) /2								
TOTAL Q4								
Q5 (a) /2								
Q5 (b) /3								
Q5 (c) (i) /1								
(ii) /2								
(iii) /1								
Q5 (d) /3								
TOTAL Q5								
Q6 (a) /1								
Q6 (b) /03								
Q6 (c) (i) /3								
(ii) /2								
(iii) /2								
(iv) /1								
TOTAL Q6								
Q7 (a) (i) /1								
(ii) /2								
(iii) /3								
Q7 (b) /2								
Q7 (c) (i) /2								
(ii) /2								
TOTAL Q7								

QUESTION	MINIMUM (OR CUT-OFF) MARK REQUIRED TO OBTAIN EACH LEVEL IN EACH QUESTION							
	INITIAL INDIVIDUAL DECISIONS				REVISED DECISIONS (IF APPROPRIATE			
	Excellent	*Very Good*	*Good*	*Satisfactory*	*Excellent*	*Very Good*	*Good*	*Satisfactory*
Q8 (a) (i) /1								
(ii) /2								
Q8 (b) (i) /3								
(ii) /3								
(iii) /2								
(iv) /1								
TOTAL Q8								
Q9 (a) (i) /3								
(ii) /2								
(iii) /3								
Q9 (b) /4								
TOTAL Q9								
Q10 (a) (i) /1								
(ii) /1								
(iii) /2								
(iv) /2								
Q10 (b) (i) /1								
(ii) /2								
(iii) /1								
(iv) /2								
TOTAL Q10								
TOTALS								

Comments:

Appendix C: Further Instructions to Judges in the Initial Year of the Study (1994) - Mathematics

STANDARD-SETTING PROCEDURES
ADVICE TO JUDGES RELATED TO STEP 3, STEP 4 AND STEP 5

In earlier steps in this procedure you worked through the examination paper independently and decided what minimum mark in each question you felt a student should score in order for their performance on that question to be classed as

excellent

very good

good

satisfactory.

You then met with the other judges in your team and through a process of discussion and negotiation arrived at an agreed minimum mark for each level in each question. These cut-off marks in each question were then totalled to give cut-off marks across the whole examination for each level of performance.

Since that time the marks of the students in the examination have become available and have been processed ready for your consideration.

A random sample of 500 students was selected from the total candidature of the examination paper. Their marks in each question, as well as their total marks in the paper, were collected. These marks were then analysed using a Rasch model called the Extended Logistic Model (ELM).

The ELM, which is an extension of Rasch's Simple Logistic Model (SLM), determines a *difficulty level* for each item (question). It also determines *threshold values* for each possible mark a student can obtain for that item. That is, it places each item and the threshold values for that item on a scale. In this way, it is possible to see the relative difficulties of each question, including how difficult it is to obtain each mark on each question.

The model also enables the *ability level* of students, based on the examination score they receive, to be placed on the same scale. In this way we can see what score on each item the statistical model estimates students with a particular ability level will obtain.

The *difficulty level* and *threshold values* for each item as well as the *ability level* corresponding to each of the four cut-off you identified have been placed on the same scale.

Step 3: Reviewing Decisions on the Basis of Statistical Data

At this stage the data will be presented in graphical form. This will enable you to see what minimum mark in each question a student would be most likely to get if their *ability level* corresponded to that of each examination cut-off scores the team established.

You will then be asked to discuss and decide whether the cut-off scores in each question the team agreed on in Step 2 are still appropriate, or whether, as a result of examining the statistical information you think a greater or smaller value should be taken for a question.

In this way the team will either modify or confirm each question cut-off value. This will lead to the team either confirming the current total paper cut-off values or establishing new ones.

Step 4: Reviewing Decisions on the Basis of Student Responses

You will be given a sample of student scripts which have been awarded the examination cut-off scores your team has now established. You will be asked to read through these scripts and decide whether you believe that they demonstrate the minimum level of knowledge and skills a student needs to demonstrate to be placed in that particular performance level. You will discuss your views with your fellow judges who have read the same scripts.

Step 5: Describing Levels of Performance

Once the examination cut-off scores are established you and your fellow team members will then examine the student scripts awarded each cut-off point. Then, using the outcomes of the course syllabus, your team will prepare statements indicating in general terms the types of mathematical knowledge and skills the students in each level typically have demonstrated.

These statements will be fairly general and will not refer to the specific examination paper. They will, however, describe in terms of the outcomes of the syllabus what students at each level can typically do. When you have written these statements you will need to check that the scripts at each cut-off score match the statements for the standard level in which they will be placed, albeit, just.

Appendix D: Further Instructions to Judges in the Subsequent Year of the Study (1995)

(Issued after the Initial cut-off Scores were set by the Teams of Judges)

The Task

We are up to the final stages of this part of the study. All that remains for us to do is to:

- have you re-familiarise yourselves with the standards of performance embodied in the materials. i.e. the 1994 examination paper, the descriptor statements and the sample scripts;
- have you check through the question and total examination cut-off marks we set for the 1995 examination paper to see that you are still happy with them;
- come together with the others in your team to review the statistical feedback relating to the performances of a sample of students on the 1995 examination paper to see whether you want to vary any of the initial cut-off marks we have established;
- look through samples of students' scripts at and around the total paper cut-off marks to see whether you want to vary any of the initial cut-off marks we have established;
- look through samples of students' scripts at and around the total paper cut-off marks to see whether you are satisfied that they would be placed in the correct standard level.

A Reminder

Throughout the process it is important to remember that what we are trying to do is to relate the standards developed using the 1994 examination paper, statistical data and student scripts, and described in the statements written for each standards level, to the 1995 examination paper and student performance. So in some cases you may need to put your own wishes and expectations aside and adopt the standards embodied in the materials provided as your own.

Background

In the Initial Year (1994) of the study a team (Team 1) applied a model which had been developed for setting standards in examinations. Applying the model to the 1994 Mathematics examination paper the team established a set of cut-off marks corresponding to their opinion of performances which could be described as *borderline excellent/very good/good/satisfactory/unsatisfactory*. That is, by applying a step by step structured approach, first individually and then as a group, the team came up with the

mark they felt a student would score in the examination paper if their performance could be described as *borderline excellent/very good.* Similarly for the other three borderlines.

In addition to establishing these cut-off scores the team wrote a set of statements describing the standards of performance of **typical** students at each level. These statements describe, in terms of the objectives of the syllabus and/or the knowledge and skills tested in the examination, the type of things which students at each standard level typically know and can do. To further illustrate these standards the team collected sample of student examination scripts which had scored the **cut-off marks**.

In the Subsequent Year (1995) of the study other teams were created. Each team used the *standards-defining materials* developed by Team 1 in the initial year. Using these materials which exemplify the standards each team worked to establish cut-off marks in the 1995 examination paper corresponding to the same standards. Initially, team members individually worked through the examination paper and listed the mark in each question they felt that a *borderline excellent/very good* student would received. They then used a similar approach for *very good/good/satisfactory/unsatisfactory* students. The team then met and, through a process of discussion and negotiation, arrived at an agreed set of cut-off marks, firstly for each question and then, for the total examination. ***This is the point we have reached at the present time.***

The Final Steps

1. *Review the Standards Materials, the 1995 Examination Paper and the Initial Cut-off Marks Established Once Again*

   Before we come together again you will need to spend some time reviewing the Standards Materials. That is, you will need to look through the 1994 examination paper noting the questions which were asked and their difficulty/complexity; you will need to read the descriptor statements carefully to ascertain what is expected of students at each standard level; you will need to read through the sample student scripts chosen at the borderlines to note the standard of work submitted by students at each level. In doing this you will need to adopt an holistic approach as students will not necessarily be consistent in the marks they score in each question.

   You will also need to look through the 1995 examination paper once again nothing the questions which were asked and their complexity and difficulty. You have already done this when you were determining your individual cut-off marks.

   Finally, you will need to have a look at the cut-off marks agreed to by your team to see whether you are still satisfied with them.

   When we first come together again we will spend some time discussing the cut-off values you have proposed. It is particularly important that we ensure they are consistent with the standards as defined by the descriptor statements and exemplified by the student scripts.

2. *Consider the Statistical Data from the Analysis of a Sample of Student Responses*

A random sample of 500 students was selected from the total 1995 candidature of the Mathematics examination. Their marks in each question, as well as their total marks in the paper, were collected. These marks were then run through a computer program which determines a *difficulty level* for each question. It also determines *threshold values* for each possible mark a student can obtain for that question. This enables us to place a measure of the difficulty of obtaining each possible mark in each question on the same scale. Hence, we can see the relative difficulties of each question, including how difficult it is to obtain each mark on each question.

The model also enables the *ability level* of students to be placed on the same scale as the questions. In this way we can obtain the total mark in the examination which would correspond to any student's ability level.

The *threshold values* for each item as well as the *ability level* corresponding to each of the four total examination cut-off scores you identified have been placed on the same scale. When we meet you will be shown this data in graphical form. It will allow you to see what minimum mark in each question a student would be expected to get if their *ability level* corresponded to that of each cut-off total.

You will then be asked to decide whether those particular cut-off values in each question are appropriate, or whether a greater or smaller value should be taken. In this way the group will either modify or confirm each question cut-off value. This will lead to the team either confirming the current total examination cut-off values or establishing new ones. Whether you mark a change or not on the basis of the statistical data is entirely a matter for the professional opinion of the team.

3. *Consider Student Scripts at the Cut-off Marks*

The final step is to look at a sample of student scripts which were awarded the total marks corresponding to each examination cut-off mark. You will need to read through these scripts to see if these performance of these students is consistent with the descriptor statements relating to the standard level to which they would be assigned on the basis of the mark they had received. The team will also be given some scripts above and below these cut-off marks to check that they are satisfied with the cut-off value.

If the members of the team have any doubts they can further discuss the decisions they have made and, if they wish, make an adjustment to the cut-off values proposed.

Once these three steps have been undertaken the cut-off marks for each standard level for the 1995 examination will have been finalised.

John Bennett

Appendix E: Item Difficulty Values and Person Ability Values Corresponding to Initial Cut-off Scores - 1994 and 1995

Tables E1 Item Threshold Difficulty Values and Initial Cut-off Score Ability Values for Mathematics Examination in 1994

Item	Loc'n	Difficulty associated with each threshold											
		1	2	3	4	5	6	7	8	9	10	11	12
1	-1.229	-2.10	-2.11	-2.08	-2.02	-1.91	-1.75	-1.52	-1.22	-0.83	-0.35	0.23	0.92
2	-0.749	-3.08	-1.98	-1.27	-0.87	-0.70	-0.67	-0.69	-0.68	-0.56	-0.24	0.37	1.36
3	-0.564	-2.65	-1.78	-1.16	-0.75	-0.51	-0.37	-0.30	-0.24	-0.15	0.02	0.32	0.80
4	-0.510	-1.81	-1.47	-1.18	-0.94	-0.74	-0.55	-0.38	-0.21	-0.04	0.16	0.39	0.66
5	-0.140	-1.65	-1.20	-0.80	-0.52	-0.27	-0.07	0.10	0.25	0.39	0.53	0.69	0.87
6	0.141	-1.87	-1.25	-0.82	-0.53	-0.33	-0.17	0.01	0.24	0.60	1.11	1.85	2.85
7	0.249	-1.42	-0.53	0.04	0.36	0.48	0.48	0.41	0.34	0.33	0.44	0.75	1.30
8	0.258	-1.97	-1.12	-0.52	-0.12	0.14	0.30	0.42	0.55	0.73	1.03	1.49	2.17
9	1.017	-1.29	-0.37	0.17	0.43	0.53	0.55	0.60	0.77	1.17	1.89	3.04	4.72
10	1.528	-0.77	0.14	0.68	0.94	1.04	1.06	1.11	1.28	1.68	2.41	3.55	5.23

Borderline	Initial cut-off score	Ability associated with each cut-off score
Excellent/Very Good	119	5.53
Very Good/Good	103	1.58
Good/Satisfactory	74	0.33
Satisfactory/Unsatisfactory	48	-0.38

Tables E2 Item Threshold Difficulty Values and Initial Cut-off Score Ability Values for English Examination in 1994

Item	Loc'n	Difficulty associated with each threshold											
		1	2	3	4	5	6	7	8	9	10	11	12
1	-0.308	-2.03	-1.36	-0.75	-0.24	0.18	0.47	0.63	0.63				
2	-0.095	-1.32	-0.90	-0.53	-0.20	0.11	0.41	0.70	0.98				
3	-0.526	-1.94	-1.05	-0.54	-0.23	0.07	0.53						
4	0.034	-1.86	-1.40	-0.93	-0.45	0	0.42	0.78	1.08	1.29	1.41		
5	0.178	-0.63	-0.48	-0.23	0.06	0.37	0.82	0.89					
6	0.080	-0.16	-0.41	-0.46	-0.35	-0.14	0.13	0.39	0.60	0.70	0.64		
7	0.127	-0.32	-0.33	-0.26	-0.14	0.02	0.23	0.39	0.57	0.73	0.85		
8	0.160	-0.28	-0.25	-0.18	-0.07	0.07	0.22	0.39	0.56	0.72	0.86		
9	0.156	-0.35	-0.29	-0.19	-0.08	0.07	0.22	0.40	0.58	0.76	0.94		
10	0.194	-0.28	-0.25	-0.19	-0.08	0.05	0.22	0.41	0.62	0.86	1.11		

Borderline	Initial cut-off score	Ability associated with each cut-off score
Excellent/Very Good	112	1.18
Very Good/Good	98	0.69
Good/Satisfactory	86	0.47
Satisfactory/Unsatisfactory	69	0.19

Tables E3 Item Threshold Difficulty Values and Initial Cut-off Score Ability Values for Biology Examination in 1994

Item	Loc'n	Difficulty associated with each threshold											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Pt A	-1.424	-5.69	-4.08	-2.88	-2.03	-1.44	-1.06	-0.79	-0.58	-0.34	0	0.51	1.27

Pt B	-0.967	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
		-3.64	-3.09	-2.60	-2.17	-1.80	-1.47	-1.19	-0.94	-0.73	-0.55	-0.38	-0.24
		13	**14**	**15**	**16**	**17**	**18**						
		-0.10	0.03	0.15	0.29	0.43	0.59						

Pt C	-0.855	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
		-3.86	-3.26	-2.82	-2.43	-2.08	-1.78	-1.52	-1.29	-1.10	-0.94	-0.81	-0.70
		13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
		-0.61	-0.54	-0.48	-0.44	-0.41	-0.38	-0.36	-0.33	-0.30	-0.27	-0.22	-0.16
		25	**26**	**27**	**28**	**29**	**30**						
		-0.09	0	0.12	0.26	0.43	0.63						

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
A1	0.290	3.04	1.49	0.35	-0.43	-0.90	-1.09	-1.05	-0.81	-0.43	0.06	0.61	1.18	1.74
A2	0.062	2.56	1.00	-0.07	-0.71	-1.01	-1.03	-0.85	-0.54	-0.17	0.19	0.45	0.56	0.42
A3	-0.200	0.73	0.02	-0.47	-0.75	-0.86	-0.84	-0.71	-0.51	-0.27	-0.12	0.21	0.39	0.48
A4	0.212	2.60	0.97	-0.13	-0.78	-1.07	-1.05	-0.82	-0.43	0.02	0.47	0.85	1.07	1.06
A5	0.219	2.55	1.08	0.04	-0.63	-0.98	-1.06	-0.92	-0.62	-0.21	0.26	0.73	1.15	1.47
A6	-0.202	1.91	0.61	-0.26	-0.75	-0.94	-0.91	-0.74	-0.49	-0.23	-0.04	0	-0.17	-0.63
A7	0.236	3.62	1.49	0.03	-0.86	-1.28	-1.32	-1.08	-0.65	-1.13	0.40	0.82	1.05	0.99
B1	0.775	3.49	1.69	0.37	-0.50	-0.99	-1.12	-0.97	-0.56	0.05	0.81	1.68	2.60	3.54
B2	-0.314	0.74	-0.02	-0.50	-0.76	-0.84	-0.79	-0.65	-0.45	-0.26	-0.10	-0.03	-0.09	-0.32
B3	0.189	4.14	1.58	-0.12	-1.10	-1.50	-1.44	-1.07	-0.52	0.07	0.58	0.89	0.78	0.21
B4	0.784	3.49	1.42	0.03	-0.78	-1.09	-1.02	-0.64	-0.07	0.62	1.32	1.94	2.39	2.57
B5	1.194	3.37	1.88	0.62	-0.38	-1.08	-1.47	-1.49	-1.14	-0.37	0.85	2.54	4.74	7.47

Borderline	Initial cut-off score	Ability associated with each cut-off score
Excellent/Very Good	82	-0.28
Very Good/Good	71	-0.34
Good/Satisfactory	61	-0.41
Satisfactory/Unsatisfactory	45	-0.57

Tables E4 Item Threshold Difficulty Values and Initial Cut-off Score Ability Values for Mathematics Examination in 1995

Item	Loc'n	Difficulty associated with each threshold											
		1	2	3	4	5	6	7	8	9	10	11	12
1	-1.123	-1.73	-2.22	-2.4	-2.31	-2.03	-1.62	-1.13	-0.63	-0.18	0.17	0.34	0.27
2	-0.667	-1.8	-1.52	-1.28	-1.07	-0.87	-0.69	-0.53	-0.37	-0.21	-0.06	0.11	0.28
3	-0.512	-1.38	-1.16	-0.98	-0.83	-0.7	-0.58	-0.47	-0.35	-0.21	-0.05	0.15	0.4
4	-0.505	-1.81	-1.58	-1.32	-1.03	-0.73	-0.45	-0.19	0.04	0.20	0.30	0.30	0.20
5	-0.062	-1.28	-0.90	-0.63	-0.43	-0.29	-0.18	-0.07	0.06	0.25	0.51	0.87	1.35
6	0.076	-0.56	-0.39	-0.27	-0.18	-0.11	-0.05	0.02	0.11	0.24	0.42	0.67	1.00
7	0.231	-1.97	-1.26	-0.73	-0.34	-0.05	0.18	0.38	0.59	0.85	1.19	1.66	2.29
8	0.296	-0.79	-0.59	-0.4	-0.22	-0.05	0.12	0.31	0.51	0.74	1.00	1.29	1.64
9	0.738	-0.17	-0.07	-0.7	0.25	0.45	0.65	0.87	1.07	1.25	1.40	1.51	1.57
10	1.528	-0.38	-0.17	0.15	0.55	0.99	1.45	1.90	2.31	2.65	2.89	3.01	2.97

Borderline	Initial cut-off score			Ability associated with each cut-off score		
	Team 1	Team 2	Team 3	Team 1	Team 2	Team 3
Excellent/Very Good	112	113	111	2.30	2.44	2.16
Very Good/Good	101	97	95	1.30	1.09	0.99
Good/Satisfactory	76	77	75	0.34	0.37	0.31
Satisfactory/Unsatisfactory	50	53	51	-0.28	-0.21	-0.25

Tables E5 Item Threshold Difficulty Values and Initial Cut-off Score Ability Values for English Examination in 1995

Item	Loc'n	Difficulty associated with each threshold											
		1	2	3	4	5	6	7	8	9	10	11	12
1	-0.277	-1.83	-0.83	-0.28	-0.05	-0.01	0.04	0.17	0.55				
2	-0.212	-1.06	-1.28	-1.08	-0.63	-0.05	0.56	0.90	0.99				
3	-0.189	-0.49	-0.74	0.09	0.38								
4	0.211	-0.08	-0.59	-0.57	-0.20	0.32	0.81	1.07	0.93				
5	0.211	-0.71	-0.68	-0.57	-0.38	-0.16	0.09	0.35	0.60	0.82	0.99	1.08	1.09
6	-0.032	-0.54	-0.68	-0.65	-0.49	-0.25	-0.05	0.35	0.62	0.81	0.90		
7	0.080	-0.45	-0.43	-0.35	-0.22	-0.01	0.14	0.35	0.57	0.79	1.00		
8	0.082	-0.32	-0.29	-0.23	-0.13	-0.02	0.12	0.28	0.44	0.61	0.78		
9	0.085	-0.34	-0.35	-0.29	-0.17	-0.01	0.17	0.35	0.51	0.64	0.72		
10	0.039	-0.55	-0.42	-0.27	-0.12	0.03	0.18	0.32	0.46	0.59	0.71		

Borderline	Initial cut-off score		Ability associated with each cut-off score	
	Team 1	Team 2	Team 1	Team 2
Excellent/Very Good	102	102	0.73	0.73
Very Good/Good	90	89	0.48	0.46
Good/Satisfactory	77	74	0.26	0.21
Satisfactory/Unsatisfactory	54	55	-0.09	-0.08

Tables E6 Item Threshold Difficulty Values and Initial Cut-off Score Ability Values for Biology Examination in 1995

Item	Loc'n	Difficulty associated with each threshold											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
mc	-0.221	-0.88	-0.89	-0.87	-0.81	-0.73	-0.63	-0.5	-0.35	-0.19	-0.01	0.19	0.39
		13	**14**	**15**									
		0.6	0.82	1.04									
		1	**1.5**	**2**	**2.5**	**3***							
16	-0.099	-0.66	-0.20	0.5	0.59	-0.8							
17	0.121	-0.52	-0.34	0.21		0.59							
18	0.463	-0.47	-0.40	0.79	1.26	-0.81							
19	-0.108	-1.21	-0.68	0.15	0.92	1.27							
20	0.364	-0.66	0.33	1.64	1.70	-1.09							
21	-0.203	-0.43	-0.36	0.08	0.61	-1.25							
22	0.208	-0.10	0.35	0.56		0.64							
23	-1.537	-4.12	-2.99	-0.39	0.97	-1.59							
24	0.265	-0.63	-0.06	0.39	1.11	2.48							
25	0.369	-0.16	0.20	0.48	0.88	1.58							
		1	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5***			
26	0.284	0.49		-1.34	-0.99	-0.12	0.70	1.16	0.89	-0.49			
27	0.228	-0.49		-0.18		0.19		0.72	1.08	1.53			
28	-0.531	-1.44		-0.94	-0.68	-0.41	-0.14	0.13	0.39	0.64			
29	0.132	-0.42		-0.04		0.38		0.59	0.54	0.33			
30	0.189	-0.33		-0.31	0.04	0.44	0.76	0.88	0.66	-0.03			
31	0.076	-1.10		-0.29	-0.22	0.14		0.74	1.29	2.09			

* As items 16 to 31 could be scored using half marks some of these thresholds are shown.

Borderline	Initial cut-off score		Ability associated with each cut-off score	
	Team 1	Team 2	Team 1	Team 2
Excellent/Very Good	64.3	65.4	1.06	1.13
Very Good/Good	56	54.7	0.62	0.56
Good/Satisfactory	46.8	40.0	0.30	0.07
Satisfactory/Unsatisfactory	35.6	26.2	-0.07	-0.38

Appendix F: Sample Fit Statistics - Data from the 1994 Mathematics and English Examinations

Mathematics 1994

The total data did not fit the model well (Chi-square = 52·396 with 27 degrees of freedom and a probability = 0·0024). While this fit improved markedly after deleting the poorest fitting item from the analysis, it was decided to retain all ten items. This approach was quite acceptable in this case given the purpose of the study and the interpretation to be placed on the results of the analysis.

The *Location Order* for each item is shown in Table F1 below. The items are shown in order of their position on the examination, not in order of fit to the model.

Table F1 *Item–Trait Interaction Test of Fit*

Item Number	Location Value	Chi-square	Probability
1	-1.229	2.720	0.420
2	-0.749	6.501	0.062
3	-0.564	8.152	0.014
4	-0.510	0.208	0.976
5	-0.140	10.264	0.000
6	0.141	8.219	0.012
7	0.249	3.723	0.271
8	0.258	2.773	0.410
9	1.017	0.400	0.938
10	1.528	9.436	0.000

English 1994

The total data did not fit the model well (Chi-square = 107.490 with 27 degrees of freedom and a probability = 0.0000). As with the Mathematics data, given the purpose to which the data were being put, it was decided to retain all items in the analysis.

The *Location Order* for each item is shown in Table F2.

Table F2 *Item - Trait Interaction Test of Fit*

Item Number	Location (logits)	Chi-square	Probability
1	- 0.308	3.995	0.239
2	- 0.095	2.156	0.527
3	- 0.526	7.744	0.022
4	0.034	29.053	0.000
5	0.178	12.286	0.000
6	0.080	43.5	0.000
7	0.127	1.990	0.561
8	0.160	1.427	0.690
9	0.156	4.427	0.195
10	0.194	1.913	0.578

Appendix G: Descriptor Statements Developed for Mathematics, English and Biology

G1 - MATHEMATICS

STANDARD	DESCRIPTOR STATEMENTS
EXCELLENT	Calculates, approximates and estimates competently and considers feasibility of answers. Translates written problems into equivalent mathematical language. Demonstrates a thorough understanding of all concepts. Applies concepts, principles, and techniques to new situations. Develops arguments and proofs, and presents them clearly, logically and concisely. Develops solutions to complex problems independently, demonstrating a thorough understanding of inter-topic relationships.
VERY GOOD	Calculates, approximates and estimates competently, and considers feasibility of answers. Translates written problems into equivalent mathematical language. Demonstrates an understanding of a wide range of concepts. Analyses a given situation, and competently uses an appropriate formula. Organises and presents information clearly and accurately. Develops solutions to a variety of problems.
GOOD	Calculates and approximates competently. Understands and correctly uses mathematical language. Demonstrates an understanding of a variety of concepts. Recalls and demonstrates competent use of appropriate formulae in familiar mathematical situations. Organises and presents information clearly. Provides solutions to familiar problems.
SATISFACTORY	Calculates and approximates successfully in most situations. Understands a variety of mathematical terms and symbols. Demonstrates an understanding of basic concepts. Recalls the appropriate formula in familiar mathematical situations. Presents information clearly. Provides, with direction, solutions to familiar problems.
UNSATISFACTORY	Performs basic calculations successfully in some situations Understands the most basic mathematical terms and symbols. Demonstrates alimited understanding of some basic concepts. Recalls some basic formulae.

G2 - ENGLISH

STANDARD	DESCRIPTOR STATEMENTS
EXCELLENT	Demonstrates a through knowledge of texts and displays insight. Develops and sustains a well structured, integrated discussion/argument of text/material. Perceptive discussion of linguistic elements pertaining to literary and non-literary text type. Appropriate choice and constructive use of quotations. Demonstrates sophisticated style and usually has flair. Demonstrates a high level of writing originality and an appreciation of audience, purpose and situation. Can perceptively analyse and interpret features and characteristics of written language and visual text (including graphics).
VERY GOOD	Demonstrates a thorough knowledge of texts. Develops a structured argument with some complexity. Proficient discussion of linguistic elements pertinent to literary and non-literary text type. Appropriate choice and constructive use of quotations. Demonstrates sophisticated style.Demonstrates originality and appreciation of audience, purpose and situation. Can effectively analyse and interpret features and characteristics of written language and visual text (including graphics).
GOOD	Demonstrates a sound knowledge of texts. Communicates argument which is mostly sustained. Competent discussion of linguistic elements pertinent to literary and non-literary text type. Appropriate choice and some constructive use of quotations. Communicates clearly and competently. Demonstrates some originality and awareness of audience, purpose and situation. Can effectively recognise and interpret features and characteristics of written languages and visual text (including graphics).
SATISFACTORY	Demonstrates a fair knowledge of texts. Recounts story, possibly frames argument. Pedestrian discussion of linguistic elements pertinent to literary and non-literary text type. Some use of quotes to support argument lacking integration. Fluent but pedestrian in style. Lacks originality and limited awareness of audience, purpose and situation. Can recognise and interpret in a limited way features and characteristics of written language and visual text (including graphics).
UNSATISFACTORY	Limited knowledge of texts. Recounts literally. Limited or no discussion of linguistic elements. Limited or no use of relevant quotes. May be disjointed or poorly expressed. Lacks originality and little or no awareness of audience, purpose and situation. Recogruses with limited or no interpretation, features and characteristics of written language and visual text (including graphics).

G3 - BIOLOGY

STANDARD	DESCRIPTOR STATEMENTS
EXCELLENT	Few, if any, gaps in knowledge of factual content. Highly developed skills in interpretation, analysis and manipulation of data presented in graphs, tables, charts etc. Has a thorough understanding of, and is able to apply, Scientific Method. (This includes the proper use of controls and the need for replication. Communicates effectively using a wide variety methods (eg. tables, written text, keys etc) incorporating an extensive use of scientific terminology Ability to apply general concepts to new situations and uses critical thinking to solve more complex problems. Can explain clearly and accurately complex biological processes eg. homeostasis.
VERY GOOD	Can explain clearly and accurately complex biological processes eg. homeostasis. Well developed skills of interpretation, analysis and manipulation of data. Well developed skills of interpretation, analysis and manipulation of data. Communicates effectively using a variety of methods incorporating use of scientific terminology. Ability to apply general concepts to many new situations and uses critical thinking to solve many problems. Can explain clearly and accurately most biological processes eg. osmosis.
GOOD	Sound knowledge of facts, perhaps with some significant gaps or misconceptions. Proficient in interpreting, analysing and manipulating data presented in straight-forward contexts. Can apply Scientific Method in simple contexts, but may not use controls or replications. Communicates clearly and adequately. The use of scientific terms maybe limited. Communicates clearly and adequately. The use of scientific terms maybe limited. Limited ability to apply general concepts to new situations and uses critical thinking to solve simple problems. Can recall definitions and explain simple biological processes eg. diffusion.
SATISFACTORY	Has a basic knowledge of Biology reliant on rote learning and recall of facts. Ability to interpret, analyse and manipulate data is limited to simple contexts. Can describe the stages of Scientific Method, but can only apply in a limited way. Communicates adequately given appropriate guidelines with a limited or inappropriate use of scientific terms. Has difficulty in applying general concepts to new situations. Can recall definitions, but generally unable to explain Biological processes.

Appendix H: The Item Cut-off Scores Initially Agreed by the Teams in 1995 and the Corresponding Values Estimated by the Extended Logistic Model

Table H1 shows the item cut-off scores (I) initially proposed by each team in Mathematics and the corresponding scores estimated (E) by the ELM.

Team 1	Excellent/ Very Good		Very Good/ Good		Good/ Satisfactory		Satisfactory/ Unsatisfactory	
Item	I	E	I	E	I	E	I	E
1	12	12	12	12	9	12	8	8
2	12	12	11	12	9	12	7	8
3	12	12	11	12	9	11	7	8
4	12	12	11	12	9	11	6	6
5	11	12	10	11	8	9	5	5
6	11	12	9	12	7	9	4	2
7	12	12	11	10	8	6	5	4
8	11	12	10	11	8	7	5	3
9	10	12	9	9	5	4	1	0
10	9	7	7	5	4	3	2	1
Total	**112**		**101**		**76**		**50**	

Team2

Item	I	E	I	E	I	E	I	E
1	12	12	12	12	11	12	8	8
2	12	12	11	12	9	12	8	8
3	12	12	11	12	9	11	7	8
4	12	12	11	12	8	12	6	6
5	11	12	10	11	9	9	6	5
6	12	12	10	12	8	9	4	3
7	12	12	9	9	7	6	4	4
8	10	12	9	10	6	7	4	3
9	11	12	8	8	6	4	3	0
10	9	8	6	5	4	3	2	1
Total	**113**		**97**		**77**		**53**	

Team3

Item	I	E	I	E	I	E	I	E
1	12	12	12	12	10	10	8	8
2	12	12	11	12	10	12	8	8
3	12	12	11	12	10	11	8	8
4	12	12	11	12	9	10	5	6
5	11	12	10	11	8	9	6	5
6	12	12	10	11	6	9	4	3
7	12	11	11	9	9	6	5	4
8	10	12	8	9	6	7	3	3
9	9	12	6	7	4	4	2	0
10	8	7	5	5	3	3	2	1
Total	**111**		**95**		**75**		**51**	

Table H2 shows the item cut-off scores (I) initially proposed by each team in English and the corresponding scores estimated (E) by the ELM

Item	Excellent/ Very Good		Very Good/ Good		Good/ Satisfactory		Satisfactory/ Unsatisfactory	
Team 1	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**
P1Q1a	3	4	3	3	2	3	1	1
b	4	3	3	3	3	2	2	2
c(i)	2	2	2	2	1	1	1	1
c(ii)	3	2	3	2	2	2	1	2
c(iii)	5	4	4	3	3	3	2	2
P1Q2	9	8	8	7	7.5	6	5	5
P1 Q3	8	8	7	7	6	6	5	4
P2 Q1	8.5	9	7.5	8	6.5	6	4.5	4
P2Q2	8.5	9	7.5	7	6.5	6	4.5	4
P2Q3	8.5	9	7.5	8	6.5	6	4.5	4
Total[1]	**102**		**90**		**77**		**54**	

Team 2

Item	I	E	I	E	I	E	I	E
P1Q1a	4	4	3	3	2.5	3	2	1
b	3	3	3	3	2	2	1	2
c(i)	2	2	1.5	2	1	1	1	1
c(ii)	3.5	2	3	2	2.5	2	2	2
c(iii)	5	4	4	3	3	3	2	2
P1Q2	8.5	8	7.5	7	6.5	6	5	5
P1 Q3	8.5	8	7.5	7	6.5	6	5	4
P2 Q1	8.5	9	7.5	8	6	6	4.5	4
P2Q2	8.5	9	7.5	7	6.5	6	5	4
P2Q3	8	9	7	8	6	6	4	4
Total[3]	**101.5**		**88.5**		**74**		**55**	

[3] These totals are obtained after items P1Q2 to P2Q3 are each multiplied by two to convert them from a maximum possible score of 10 to a maximum possible score of 20.

Table H3 shows the item cut-off scores (I) initially proposed by each team for the core sections of the Biology examination the corresponding scores estimated (E) by the ELM

Item	Excellent/ Very Good		Very Good/ Good		Good/ Satisfactory		Satisfactory/ Unsatisfactory	
Team 1	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**
mc	13	15	11.4	13	9.7	11	7.5	9
Q16	2.7	2.5	2.4	2.5	2.0	1.5	1.5	1.5
Q17	2.7	3	2.4	3	2.0	2	1.5	1.5
Q18	2.5	2	2.1	1.5	1.8	1.5	1.4	1.5
Q19	2.7	2.5	2.3	2	1.8	2	1.4	1.5
Q20	2.4	1.5	2.0	1.5	1.6	1	1.2	1
Q21	2.7	2.5	2.4	2.5	2.0	2	1.5	1.5
Q22	2.5	3	2.3	2	1.9	1	1.4	1
Q23	2.7	2.5	2.4	2	2.1	2	1.7	2
Q24	2.6	2	2.3	2	1.8	1.5	1.5	1.5
Q25	2.7	2.5	2.4	2	2.1	1.5	1	1
Q26	4.0	3.5	3.2	3	2.5	3	3	3
Q27	4.5	4	3.9	3	3.1	3	2	2
Q28	4.4	5	3.8	4.5	3.2	4	3.5	3.5
Q29	4.0	3	3.5	2	3.0	1	0.5	0.5
Q30	4.2	4	3.7	3	3.2	2.5	2	2
Q31	4.0	4	3.5	3	3.0	3	2.5	2.5
Total	**64.3**		**56**		**46.8**		**35.6**	

Team 2	**I**	**E**	**I**	**E**	**I**	**E**	**I**	**E**
mc	13.5	15	11.2	12.5	8.5	10	5.9	8
Q16	2.7	2.5	2.1	2.5	1.5	1.5	0.9	1
Q17	2.7	3	2.1	3	1.5	1.5	0.9	1.5
Q18	2.7	2	2.4	1.5	1.8	1.5	1.5	1.5
Q19	2.4	2.5	2.1	2	1.5	1.5	1.2	1.5
Q20	2.4	1.5	2.4	1.5	1.2	1	0.6	1
Q21	2.4	2.5	1.8	2	1.2	1.5	0.6	1.5
Q22	2.7	3	2.4	2	1.8	1	0.9	0.5
Q23	3.0	2.5	2.4	2	2.1	2	1.5	2
Q24	2.7	2	2.4	2	1.8	1.5	1.2	1
Q25	2.7	2.5	2.4	2	2.1	1	1.5	0.5
Q26	4.0	3.5	3.5	3	2.5	3	1.5	2.5
Q27	4.5	4.5	4.0	3	3.0	2.5	2.0	2
Q28	4.5	5	3.5	4.5	2.5	4	2.0	3
Q29	4.0	4	3.0	2	2.0	1	1.0	0.5
Q30	4.5	4	3.5	3	2.5	2.5	1.5	2
Q31	4.0	4	3.5	3	2.5	3	1.5	2.5
Total	**65.4**		**54.7**		**40.0**		**26.2**	

Appendix I: Questionnaire Given to Judges Involved in the Study in Both 1994 and 1995

(Those judges only involved in the study in 1995 received a questionnaire containing the questions relating to the Second Stage)

NAME:

You have participated in a lengthy and involved process to develop and test a model for setting standards in examinations such as the HSC. I am very grateful for your support, cooperation and the professional manner with which you have approached the task. I hope that you have found your involvement to be rewarding.

During our meetings you provided some very useful feedback which will be useful I improving the model. I would now like you to reflect on the whole process and provide any comments you would like to make. I have posed some questions of interest to me but invite you to make any other comments you wish. Please attach other pages if you need more space.

Thank you again for your support.

The first stage - The Process of Setting and Refining the Standards

1. Did the process adopted enable the team to arrive at a satisfactory consensus?

2. How important in the process was:

   the individual work prior to meeting?

   the group discussions?

   the statistical date (that is the graph you were shown)?

   the sample scripts?

3. When coming up with a cut-off value during group discussions were you happy with the process for gaining consensus or would you have preferred that we simply averaged the opinions of the various team members? Would you have preferred some other approach?

4. In defining the standards levels you created how important is the?

   examination paper?

   descriptor statements?

   student scripts?

5. Were you satisfied with the standards agreed upon by the group?

6. How confident were you that the standards you created and defined would be clearly understood by other teachers of your subject?

7. Do you think what you created could be put in a form which would be simple for non-teachers to understand and interpret?

8. What do you see to be the strengths of this models?

9. What do you see to be the weaknesses of this model?

10. Any other comments

The Second Stage - The Process of Applying the Standards Set in the Initial Year to an Examination in a Subsequent Year

1. How easy was it to pick up the standards from the materials which defined the standards? (That is the *Descriptors*, the *Examination Paper* and the *Student Scripts*)

2. Was any component of the standards materials more helpful to you than the rest? If so, which?

3. Describe how you went about the process of becoming familiar with the standards.

4. How did you go about imposing the standards based around the 1994 examination (as you interpreted them from the material) onto the 1995 examination paper?

5. How important in the process of matching the standards from 1994 to 1995 was:

the individual work prior to meeting?

the group discussions?

the statistical data (the graphs showing the difficulty of the questions and the ability of the students)?

the sample scripts?

6. What, if any, other information would have proved useful?

7. What, if any, modifications would you make to the process?

8. Do you think the Descriptor Statements are appropriate and accurate? Would you like to see any changes made to this aspect?

9. Could you apply this process in a real situation? What, if any, changes would be necessary to make it work?

10. What do you regard as the strengths of this model for matching standards across different years?

11. What do you see to be the weaknesses of this model for matching standards across different years?

12. Any other comments

Appendix J: Calculation of the Tertiary Entrance Score (TES)

The Tertiary Entrance Score (TES) is calculated by the authority responsible for coordinating the selection of students for places in university courses. It is an aggregate with a maximum possible value of 500, and consists of the sum of the course marks achieved by each student, after the examination marks awarded in every course have been scaled to take into account the academic ability of the candidature of each course.

In its report on the 1997 Higher School Certificate, the Technical Committee on Scaling, a standing committee of the NSW Vice-Chancellors Conference, explains the process, and the need perform this operation in the following terms:

> Different courses have different quality candidatures with some courses being attempted by students of higher than average ability. When the marks (*scores*) in all subjects (*courses*) are standardised to the same median as occurs with the Board marks (*course scores reported by the examining authority administering the NSW Higher School Certificate examinations*) students taking subjects with high quality candidatures have lower marks than if they were competing with students of a lower academic calibre. Scaling attempts to remove this disadvantage by adjusting the marks so that the average per-unit mark (*course score*) for a course reflects the average academic quality of the course candidature.
>
> The scaling process then determines weightings for (2 unit) courses according to the quality of their candidatures, and the (*examination*) scores for each course are adjusted accordingly. The quality of a candidature is defined as the *average academic performance of the candidature* where the academic performance of a student is the average performance in all courses attempted.
>
> While the order of merit within each course is not affected by the scaling process, the scaled scores will in most cases be different from the original scores. The maximum effect of the scaling occurs for the middle students because marks of 0 and 50 are unaltered.
>
> An aggregate (out of 500) is formed by summing the scaled marks from the best unit of English the best unit from each of the two Key Learning Area groups and the best seven units chosen from the remaining units (*ie according to certain rules for the calculation of the TES*). Because scaled scores differ from unscaled scores, the order of merit by the scaled aggregate will in most eases differ from that obtained by adding the Board's marks (*examination scores*) for the best ten units.
>
> From the *NSW Vice-Chancellors Conference Technical Committee on Scaling, Report on the Scaling of the 1997 NSW Higher School Certificate, p. 3.*

REFERENCES

Andrich, D. (1978) A Rating Formulation for Ordered Response Categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1988) *Rasch Models for Measurement.* Sage Publications, Newbury Park, California.

Angoff, W. (1971) Scales, Norms and Equivalent Scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600) Washington, DC: American Council on Education.

Becker, D. and Forsyth, R. (1992) An Empirical Investigation of Thurstone and IRT Methods of Scaling Achievement Tests. *Journal of Educational Measurement, 29,* 341-354.

Berk, R. (1986) A Consumer's Guide to Setting Standards on Criterion Referenced Tests. *Review of Educational Research, 56,* 137-172.

Berk, R. (1995) Something Old, Something New, Something Borrowed, a Lot to Do. *Applied Measurement in Education, 8,* 99-109.

Berk, R. (1996) Standard Setting: The Next Generation. *Applied Measurement in Education, 9,* 215-235.

Beuk, C. (1984) A Method for Reaching a Compromise Between Absolute and Relative Standards in Examinations. *Journal of Educational Measurement, 21,* 147-152.

Block, J. (1978) Standards and Criteria: A Response. *Journal of Educational Measurement, 15,* 291-295.

Brennan, R.L. and Lockwood, R.E. (1980) A Comparison of the Nedelsky and Angoff Cutting Score Procedures Using Gereralizability Theory. *Applied Psychological Measurement, 4,* 219-240.

Breyer, F.J. and Lewis, C. (1994) Pass-Fail Reliability for Tests with Cut Scores: A Simplified Method. *(RR 94-39). Princeton, NJ: Educational Testing Service.*

Busch, J. and Jaeger, R. (1990) Influence of Type of Judge, Normative Information, and Discussion on Standards Recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145-163.

Chang, L., Dziuban, C., Hynes, M. and Olson, A. (1996) Does a Standard Represent Minimal Competency of Examinees or Judge Competency? *Applied Measurement in Education, 9,* 161-173.

Choppin, B. (1983) *The Rasch Model for Item Analysis.* Center for the Study of Evaluation. Report No. 219. University of California.

CITO (1990) Overview of the Activities of the Department of Examinations in Secondary Education. *Unpublished collection of papers by CITO staff.*

Cizek, G. (1993) Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30,* 93-106.

Cizek, G. (1996) Standard-Setting Guidelines. *Educational Measurement: Issues and Practice* Spring: 13-21.

Cross, L., Impara, J., Frary, R. and Jaeger, R. (1984) A Comparison of Three Methods of Obtaining Minimum Standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113-129.

Cross, L., Frary, R., Kelly, P., Small, R. and Impara, J (1985) Establishing Minimum Standards for Essays: Blind Versus Informed Reviews. *Journal of Educational Measurement, 22,* 137-146.

De Gruijter, D. (1985) Compromise Models for Establishing Examination Standards. *Journal of Educational Measurement, 22,* 263-269.

DeMauro, G. and Powers, D. (1993) Logical Consistency of the Angoff Method of Standard Setting. *(RR 93-26). Princeton, NJ: Educational Testing Service.*

Divgi, D. (1986) Does the Rasch Model Really Work for Multiple Choice Items? Not If You Look Closely. *Journal of Educational Measurement 23,* 283-298.

Ebel, R. (1972) *Essentials of Educational Measurement* (2nd Edition). Englewood Cliffs, N.J. Prentice-Hall. (Reprinted in Ebel, R. and Frisbie, D. *Essentials of Educational Measurement* (4th Edition) 1986 (pp 279-283).

Engelhard, G. and Anderson, D. (1996) A Binomial Trials Model for Examining the Ratings of Standard-setting Judges. *Applied Measurement in Education* (In Press)

Engelhard, G. and Cramer, S. (1997) Using Rasch Measurement to Evaluate the Ratings of Standard-setting Judges. In M. Wilson, G. Engelhard & K. Draney (Eds.) *Objective Measurement: Theory into Practice, Vol. 4* (pp 97-112) Norwood, NJ : Ablex.

Engelhard, G. and Gordon, B. (1997) Setting and Evaluating Performance Standards for High Stakes Writing Assessments. In M. Wilson and G. Engelhard (Eds.) *Objective Measurement: Theory into Practice, Vol. 5.* (In Press)

Engelhard, G and Stone, G. (1997) Evaluating the Quality of Ratings Obtained from Standard-setting Judges. *Educational and Psychological Measurement* (In Press).

Faggen, J., Melican, G. and Powers, D. (1995) Effects of Mode of Item Presentation on Standard Setting. *(RR 95-26). Princeton, NJ: Educational Testing Service.*

Fehrmann, M., Woehr, D. and Arthur, W. (1991) The Angoff Cutoff Score Method: The Impact of Frame-of-Reference Rater Training. *Educational and Psychological Measurement, 51,* 857-872.

Geisinger, K. (1991) Using Standard-Setting Data to Establish Cutoff Scores. *Educational Measurement: Issues and Practice, 10,* 17-22.

Glass, G. (1978) Standards and Criteria. *Journal of Educational Measurement, 15,* 237-271.

Goldstein, H. (1979) Consequences of Using the Rasch Model for Educational Assessment. *British Educational Research Journal, 5,* 211-222.

Goldstein, H. (1980) Dimensionality, Bias, Independence and Measurement Scale Problems in Latent Trait Test Score Models. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246.

Hambleton, R. and Cook, L. (1977) Latent Trait Models and Their Use in the Analysis of Educational Test Data. *Journal of Educational Measurement, 14,* 75-96.

Hambleton, R. (1978) On the Use of Cutoff Scores with Criterion-Referenced Tests in Instructional Settings. *Journal of Educational Measurement, 15,* 277-290.

Hambleton, R. and Plake, B. (1995) Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Applied Measurement in Education, 8,* 41-55.

Harasym, P. (1981) A Comparison of the Nedelsky and Modified Angoff Standard-Setting Procedure on Evaluation Outcome. *Educational and Psychological Measurement, 41,* 725-734.

Harris, D. (1991) A Comparison of Angoff's Design I and Design II for Vertical Equating Using Traditional and IRT Methodology. *Journal of Educational Measurement, 28,* 221-235.

Henriksson, W. (1993) Effects of Repeated Test Taking on Swedish Scholastic Aptitude Test. *Paper presented to the Annual Meeting of the International Association for Educational Assessment, Mauritius, May 1993.*

Hills, J., Subhiyah, R. and Hirsch, T. (1988) Equating Minimum-Competency Tests: Comparison of Methods. *Journal of Educational Measurement, 25,* 221-231.

Hofstee, W. (1983) The Case for Compromise in Educational Selection and Grading. In S. B. Anderson & J. S. Helmick (Eds.), *On Educational Testing.* San Francisco: Jossey-Bass.

Huynh, H. and Ferrara, S. (1994) A Comparison of Equal Percentile and Partial Credit Equatings for Performance-Based Assessments Composed of Free-Response Items. *Journal of Educational Measurement, 31,* 125-141.

Impara, J. and Plake, B. (1996) Teacher's Ability to Estimate Item Difficulty: A Test of the Assumptions of the Angoff Standard Setting Method. *Paper presented to the Annual Meeting of the National Council on Measurement in Education, New York, April 1996.*

International Baccalaureate Organisation. (1996) Grade Award Support Document. *Unpublished Handbook for Judges Involved in Grade Setting.*

Jaeger, R. (1982) An Iterative Structured Judgment Process for Establishing Standards on Competency Tests of Theory and Application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. (1988) Use and Effect of Caution Indices in Detecting Aberrant Patterns of Standards-Setting Judgments. *Applied Measurement in Education, 1,* 17-31.

Jaeger, R. (1990) Establishing Standards for Teacher Certification Tests. *Educational Measurement: Issues and Practice, 9,* 15-20.

Jaeger, R. (1991) Selection of Judges for Standard-Setting. *Educational Measurement: Issues and Practice, 10,* 3-14.

Jaeger, R. (1995) Setting Performance Standards Through Two-Stage Judgmental Policy Capturing. *Applied Measurement in Education, 8,* 15-40.

Kahl, S., Crockett, T., DePascale, C. and Rindfleisch, S. (1994) Using Actual Student Work to Determine Cut Scores for Proficiency Levels. *Paper presented at the National Conference on Large Scale Assessment, Albuquerque, June 1994.*

Kane, M. (1986) The Interpretability of Passing Scores. *American College Testing Program Technical Bulletin No. 52. Iowa.*

Kane, M. (1987) On the Use of IRT Models With Judgemental Standard Setting Procedures. *Journal of Educational Measurement, 24,* 333-345.

Kane, M. (1994) Validating the Performance Standards Associated with Passing Scores. *Review of Educational Research, 64,* 425-461.

Kolen, M. (1981) Comparison of Traditional and Item Response Theory Methods for Equating Tests. *Journal of Educational Measurement, 18,* 1-11.

Linn, R. (1978) Demands, Cautions and Suggestions for Setting Standards. *Journal of Educational Measurement, 15,* 301-308.

Livingston, S. and Zieky, M. (1982) *Passing Scores. A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: Educational Testing Service.

Livingston, S. and Lewis, C. (1995) Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Education Measurement, 32,* 179-197.

Lord, F. (1950) Notes on Comparable Scales for Test Scores. Educational Testing Service *Research Bulletin*, 48.

Lord, F. (1980) *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Luijten, A. (1988) Internal Versus External Assessment in the Dutch Examinations at 16+ and 18+. *Educational Psychology.*

McGaw, B. (1996) *Their Future, Options for Reform of the Higher School Certificate.* Department of Training and Education Co-ordination New South Wales.

McGaw, B. (1997) *Shaping Their Future, Recommendations for Reform of the Higher School Certificate.* Department of Training and Education Co-ordination New South Wales.

McKinley, D., Newman, L. and Wiser, R. (1996) Using the Rasch Model in the Standard Setting Process. *Paper presented at the annual meeting of the National Council of Measurement in Education New York. April 1996.*

McLean, L. and Ragsdale, R. (1983) The Rasch Model for Achievement Tests-Inappropriate in the Past, Inappropriate Today, Inappropriate Tomorrow. *Canadian Journal of Educational, 8,* 71-76.

Masters, G. (1982) A Rasch Model for Partial Credit Scoring. *Psychometrika, 49, 269-272.*

Meskauskas, J. (1976) Evaluation Models for Criterion-Referenced Testing: Views regarding Mastery and Standard-setting. *Review of Educational Research, 46,* 133-158.

Messick, S. (1994) Standards-based Score Interpretation: Establishing Valid Grounds for Valid Inferences. *(RR 94-57). Princeton, NJ: Educational Testing Service.*

Mills, C. (1983) A Comparison of Three Methods of Establishing Cut-off Scores on Criterion-Referenced Tests. *Journal of Educational Measurement, 20,* 283-292.

Mills, C. and Melican, G. (1988) Estimating and Adjusting Cutoff Scores: Features of Selected Methods. *Applied Measurement in Education, 1,* 261-275.

Mills, C., Melican, G. and Ahluwalia, N. (1991) Defining Minimal Competence. *Educational Measurement: Issues and Practice, 10,* 7-10.

Mills, C. (1995) Comments on Methods of Setting Standards for Complex Performance Tasks. *Applied Measurement in Education, 8,* 93-97.

Mislevy, R., Sheehan, K. and Wingersky, N. (1992) How to Equate Tests with Little or No Data. *(RR 92-20). Princeton, NJ: Educational Testing Service.*

Morgan, G. (1982) The Use of the Latent Trait Measurement Model in the Equating of Scholastic Aptitude Tests. In D. Spearitt (Ed) *The Improvement of Measurement in Education and Psychology* (pp. 189-208) ACER, 1982.

Morrison, H., Busch, J. and D'Arcy, J. (1994) Setting Reliable National Curriculum Standards: a Guide to the Angoff Procedure. *Assessment in Education, 1,* 181-199.

Nedelsky, L. (1954) Absolute Grading for Objective Tests. *Educational and Psychological Measurement, 14,* 3-19.

Norcini, J., Lipner, R., Langdon, L. and Strecker, C. (1987) A Comparison of Three Variations on a Standard-Setting Method. *Journal of Educational Measurement, 24,* 56-64.

Norcini, J., Shea, J. and Kanya, D. (1988) The Effect of Various Factors on Standard Setting. *Journal of Educational Measurement, 25,* 57-65.

Norcini, J., Shea, J. and Ping, J. (1988) A Note on the Application of Multiple Matrix Sampling to Standard Setting. *Journal of Educational Measurement, 25,* 159-164.

Norcini, J. (1990) Equivalent Pass/Fail Decisions. *Journal of Educational Measurement, 27,* 59-66.

Norcini, J., Shea, J. and Grosso, L. (1991) The Effect of Numbers of Experts and Common Items on Cutting Score Equivalents Based on Expert Judgment. *Applied Psychological Measurement, 15,* 241-246.

Norcini, J. and Shea, J. (1992) Equivalent Estimates of Borderline Group Performance in Standard Setting. *Journal of Educational Measurement, 29,* 19-24.

Norcini, J. and Shea, J. (1997) The Credibility and Comparability of Standards. *Applied Measurement in Education, 10,* 39-59.

New South Wales Vice-Chancellors Conference Technical Committee on Scaling, *Report on the Scaling of the 1997 NSW Higher School Certificate March, 1998.*

Plake, B. and Kane, M. (1991) Comparison of Methods for Combining the Minimum Passing Levels for Individual Items into a Passing Score for a Test. *Journal of Educational Measurement, 28,* 249-256.

Plake, B., Melican, G. & Mills, C. (1991) Factors Influencing Intrajudge Consistency During Standard-Setting. *Educational Measurement: Issues and Practice, 10,* 15-26.

Plake, B., Impara, J. and Potenza, M. (1994) Content Specificity of Expert Judgements in a Standard-Setting Study. *Journal of Educational Measurement, 31,* 339-347.

Plake, B. (1995) An Integration and Reprise: What We Think We Have Learned. *Applied Measurement in Education, 8,* 85-92.

Plake, B. and Impara, J. (1996) Intrajudge Consistency Using the Angoff Standard Setting Method. *Paper presented at the Annual Meeting of the National Council on Measurement in Education New York. April 1996.*

Plake, B., Hambleton, R. and Jaeger, R. (1997) A New Standard-setting Method for Performance Assessments: The Dominant Profile Judgment and Some Field-test Results. *Educational and Psychological Measurement, 57, 400-411.*

Plake, B. (1998) Setting Performance Standards for Professional Licensure and Certification. *Applied Measurement in Education, 11, 65-80.*

Poggio, J. and Glasnapp, D. (1994) A Method for Setting Multilevel Performance Standards on Objective or Constructed Response Tests. *Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April 1994.*

Popham, W. (1978) As Always Provocative. *Journal of Educational Measurement, 15,* 297-300.

Putham, S., Pence and Jaeger, R. (1995) A Multistage Dominant Profile Method for Complex Performance Assessment. *Applied Measurement in Education, 8,* 57-83.

Rasch, G. (1960/1980) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research Copenhagen (Reprinted, with Foreword and Afterword by B.D. Wright, by University of Chicago Press: Chicago, 1980)

Reid, J. (1991) Training Judges to Generate Standard-Setting Data. *Educational Measurement: Issues and Practice, 10,* 11-14.

School Curriculum and Assessment Authority. (1996) *Code of Practice for GCE A and AS Examinations,* March 1996. London.

Scottish Examination Board. (1996) *Handbook for Examinations 1996.* Scottish Examination Board, Dalkeith.

Scriven, M. (1978) How to Anchor Standards. *Journal of Educational Measurement, 15,* 273-275.

Shepard, L. (1980) Standard Setting Issues and Methods. *Applied Psychological Measurement, 4,* 447-467.

Siegel, S. (1956) *Nonparametric Statistics for the Behavioural Sciences.* McGraw-Hill, New York.

Smith, D. (1994) *Where Now? Destinations of Young People Who Miss Out On Higher Education.* Report to the National Youth Affairs Research Scheme. National Clearinghouse for Youth Studies, Hobart, Tasmania.

Smith, R. (1986) Person Fit in the Rasch Model. *Educational and Psychological Measurement, 46,* 359-372.

Smith R. and Smith J. (1988) Differential Use of Item Information by Judges Using Angoff and Nedelsky Procedures. *Journal of Educational Measurement, 25,* 259-274.

Taylor, N. (1979) *HSC Repeats.* Research Report, NSW Department of Technical and Further Education

Tognolini, J. and Andrich, D. (1995) Differential Subject Performance and the Problems of Selection. *Journal of Assessment and Evaluation in Higher Education, 20,* 161-173.

Tognolini, J. and Andrich, D. (1996) Profile Analysis of Students Applying for Entry to Tertiary Institutions. *Applied Measurement in Education, 9, 323-353.*

Van der Linden, W. (1982) A Latent Trait Method for Determining Intrajudge Inconsistency in the Angoff and Nedelsky Techniques of Standard Setting. *Journal of Educational Measurement, 19,* 295-308.

Waltman, K. (1997) Using Performance Standards to Link Statewide Achievement Results to NAEP. *Journal of Educational Measurement, 34,* 101-121

Webb, M. and Miller, E. (1995) A Comparison of the Paper Selection Method and the Contrasting Groups Method for Setting Standards on Constructed-Response Items. *Paper pesented at the Annual Meeting of the National Council on Measurement in Education. San Francisco April 1995.*

Wiliam, D., (1996). Meanings and Consequences in Standard Setting. *Assessment in Education, 3,* 287-307.

Wright, B. (1967) Sample-free Test Calibration and Person Measurement. *Paper presented at the Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service. October 1967.*

Wright, B. (1977) Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement, 14,* 97-115.

Wright, B. and Stone, M. (1979) *Best Test Design.* Mesa Press, Chicago.

Wright, B. and Masters, G. (1982) *Rating Scale Analysis.* Mesa Press, Chicago.

Zieky, M. (1996) A Historical Perspective on Setting Standards. *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessments of the National Assessment Governing Board and the National Center for Educational Statistics, October 1994.* Personal Correspondence.