OFFICE OF THE BOARD OF STUDIES

MODERATION ISSUES IN A STANDARDS FRAMEWORK

THE INTRODUCTION OF THE STANDARDS-BASED HIGHER SCHOOL CERTIFICATE IN NEW SOUTH WALES

ACACA CONFERENCE

SYDNEY

JULY, 2001

DR JOHN BENNETT
DIRECTOR, INFORMATION SERVICES

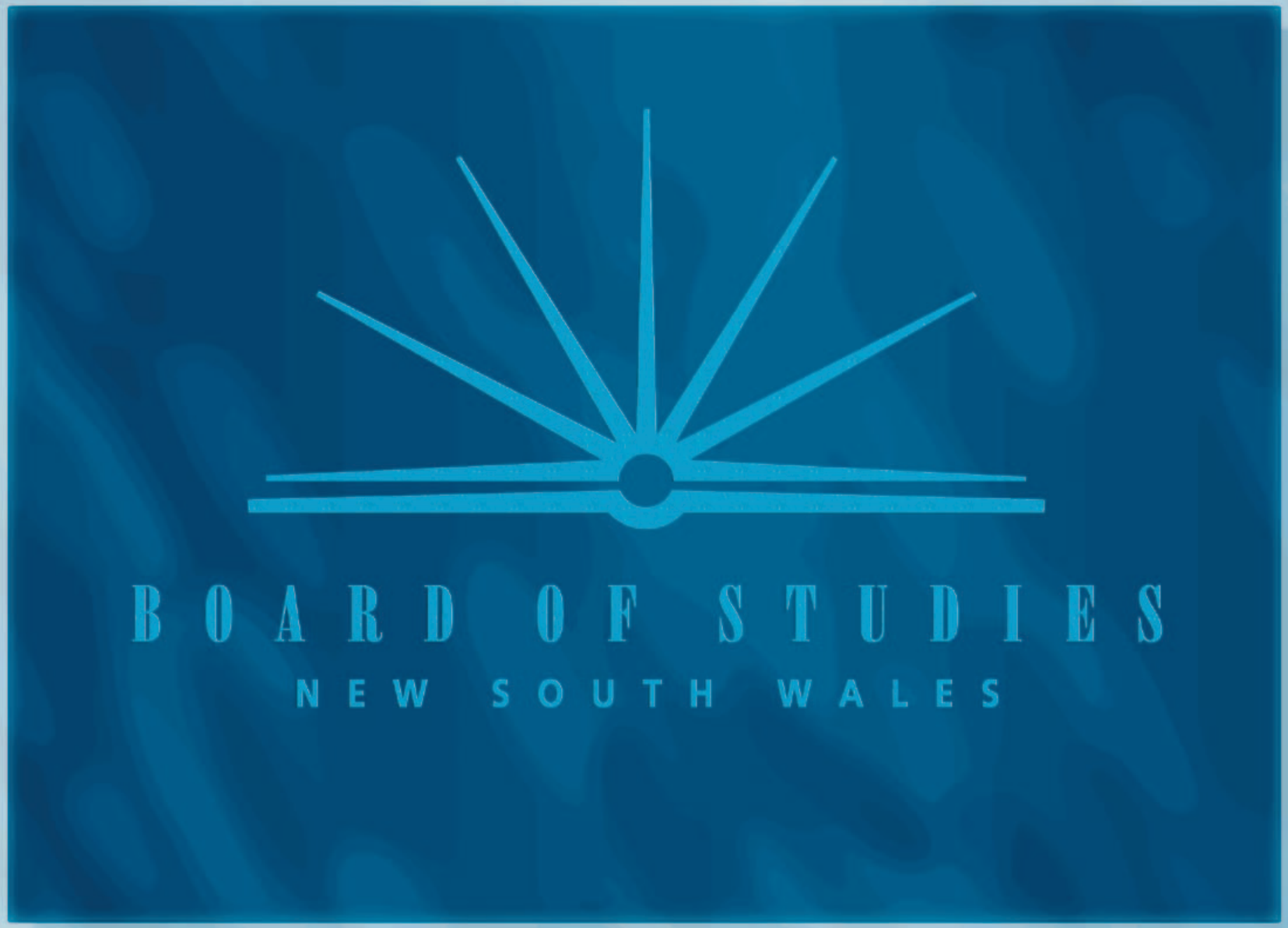
CONTENTS

1. **Copy of the slides used in the presentation**

2. **The Procedure for Aligning Student Achievement in the 2001 HSC Examinations to the Performance Scales**
   (A brief explanation of the standards-setting procedure that will be used)

3. **HSC Standards Packages**
   (An explanation of the nature, purpose and content of the standards packages to be produced following the 2001 HSC examinations)

4. **Setting Standards and Applying Them across Different Administrations of Large-scale, High-stakes, Curriculum-based Public Examinations**
   (A paper summarising key issues, research and practices associated with setting performance standards in examinations and their implications for the setting of standards in the NSW Higher School Certificate program)


BOARD OF STUDIES
NEW SOUTH WALES

MODERATION ISSUES IN A STANDARDS FRAMEWORK

THE INTRODUCTION OF THE STANDARDS-BASED HIGHER SCHOOL CERTIFICATE

THE TASK

In 2001 NSW will report student achievement in the HSC using a standards-based approach

- **the reports should clearly state what students know and can do**
- **the approach must enable comparisons between student performances from year to year**
- **the process should be transparent and not involve complex statistical techniques**

THE GOAL

HIGHER SCHOOL CERTIFICATE

2001 Course Report

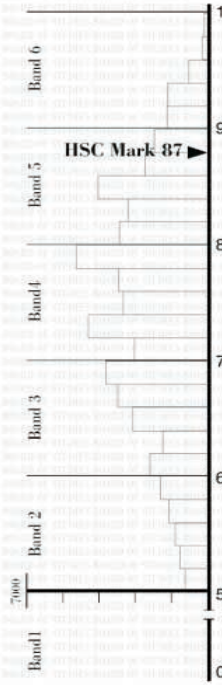
BOARD OF STUDIES
NEW SOUTH WALES

2 Unit Personal Development, Health and Physical Education

Sample Student

Assessment Mark 86 — Examination Mark 88

State Distribution



A typical performance in this band is demonstrated when a student:

Band 6 — Demonstrates extensive knowledge and understanding of the range of concepts related to health and physical performance. Comprehensively applies theoretical principles to design and evaluates specific strategies for improving health. Demonstrates a superior understanding of the interrelated roles and responsibilities of individuals, groups and governments in the management and promotion of health. Critically analyses movement and the range of factors that affect physical performance and participation. Provides relevant and accurate examples to justify complex arguments.

Band 5 — Clearly expresses ideas that demonstrate a thorough understanding of health and physical performance concepts. Identifies strategies for improving health and discusses the links between individual health behaviour, social issues and community health status. Demonstrates a detailed understanding of the interrelated roles of individuals, groups and governments in the management and promotion of health. Demonstrates an understanding of the interrelationships between the various factors that impact on physical performance. Supports arguments thoroughly by using relevant examples and current information.

Band 4 — Demonstrates a clear understanding of the broad concepts that impact on personal health and physical performance. Relates strategies for managing the major causes of sickness and death to the contributing risk factors. Demonstrates a sound understanding of the roles of individuals, groups and governments in promoting health. Describes a range of factors that affect the quality of physical performance. Communicates information in a clear and logical way, providing some examples.

Band 3 — Uses basic definitions and facts when explaining health and physical performance concepts. Identifies the major causes of sickness and death and establishes that a healthy lifestyle is a desirable goal. Demonstrates an understanding of the need for government and community action in relation to promoting health. Identifies some relevant factors that influence physical performance. Provides basic support for the arguments presented.

Band 2 — Recalls some simple facts and writes brief descriptions. Demonstrates an understanding of elementary terms and recognises simple cause and effect relationships as they apply to health and movement. Outlines some factors affecting health and identifies relevant illness prevention measures. Demonstrates an understanding of general movement principles. Provides limited support for the arguments presented.

Band 1

Student Number: 65487965

2660 180

Issued by the Board of Studies without alteration or erasure.

President

THE JOURNEY

- A major revision of the curriculum was undertaken with every course written to include clear expressions of (curriculum) standards in the form of outcomes

- Band Descriptions were developed - statements that summarise six different levels of performance in each course

THE JOURNEY (cont.)

- Examinations and marking guidelines have been prepared that will readily enable measurement of the extent to which students have achieved course outcomes
- An Angoff-based judgmental-empirical standards-setting procedure has been developed to establish the (performance) standards by aligning student marks to the performance scale

THE JOURNEY (cont.)

- Standards Packages will be created that encapsulate the performance standards and presents them in a way that makes them accessible and clear to all parties

THE IMPORTANCE OF 2001

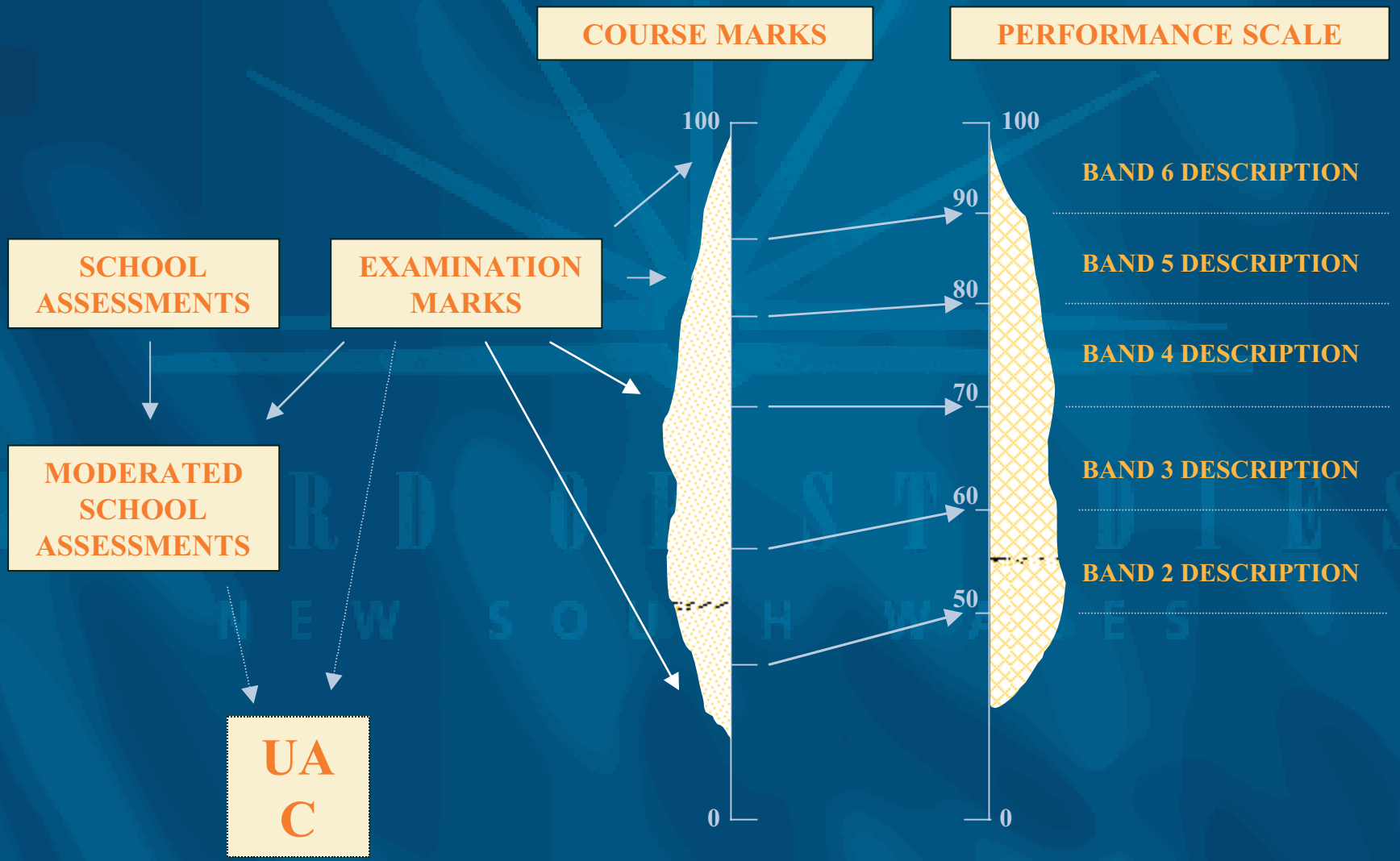
- **In 2001 the standards will be established against which student performance will be reported in 2001, 2002, 2003, . . .**

  That is

  – **In 2001 the standards are established**

  – **In 2002, 2003, . . . the SAME standards are used to report student performance**

ALIGNING MARKS TO THE PERFORMANCE SCALES



THE STANDARDS-SETTING OPERATION

- Appoint a team of judges
  - A team of judges will be formed - experienced teachers/markers who do not have a managerial role in the marking operation
- Train the judges
  - The judges will attend a training session in September (4 hours)

THE STANDARDS-SETTING OPERATION (cont.)

- **Brief the judges at marking (2 hours/course)**
  - **early in the marking operation (soon after final marking commences) judges will meet and be given course-specific materials**

THE STANDARDS-SETTING OPERATION (cont.)

- **Stage 1** **(8 hours/course)**
  - **(independently) judges review the performance descriptions and marking schemes and record what mark they think a student at the *borderline* of each band would score for each question**

THE STANDARDS-SETTING OPERATION (cont.)

- **Stage 2** (8 hours/course)
  - the judges come together as a team, they are
    - » briefed on special statistical reports and
    - » discuss and review the decisions they made individually during Stage 1
- **Stage 3** (8 hours/course)
  - the judges meet a further time
    - to consider selected student scripts and
    - further discuss and review their decisions

THE STANDARDS-SETTING OPERATION (cont.)

- **The HSC Consultative Committee**
  - **Finally, the team meets with the HSC Consultative Committee to discuss the operation and recommend the band cut-off marks for their course(s) (early December)**

CAPTURING THE STANDARDS

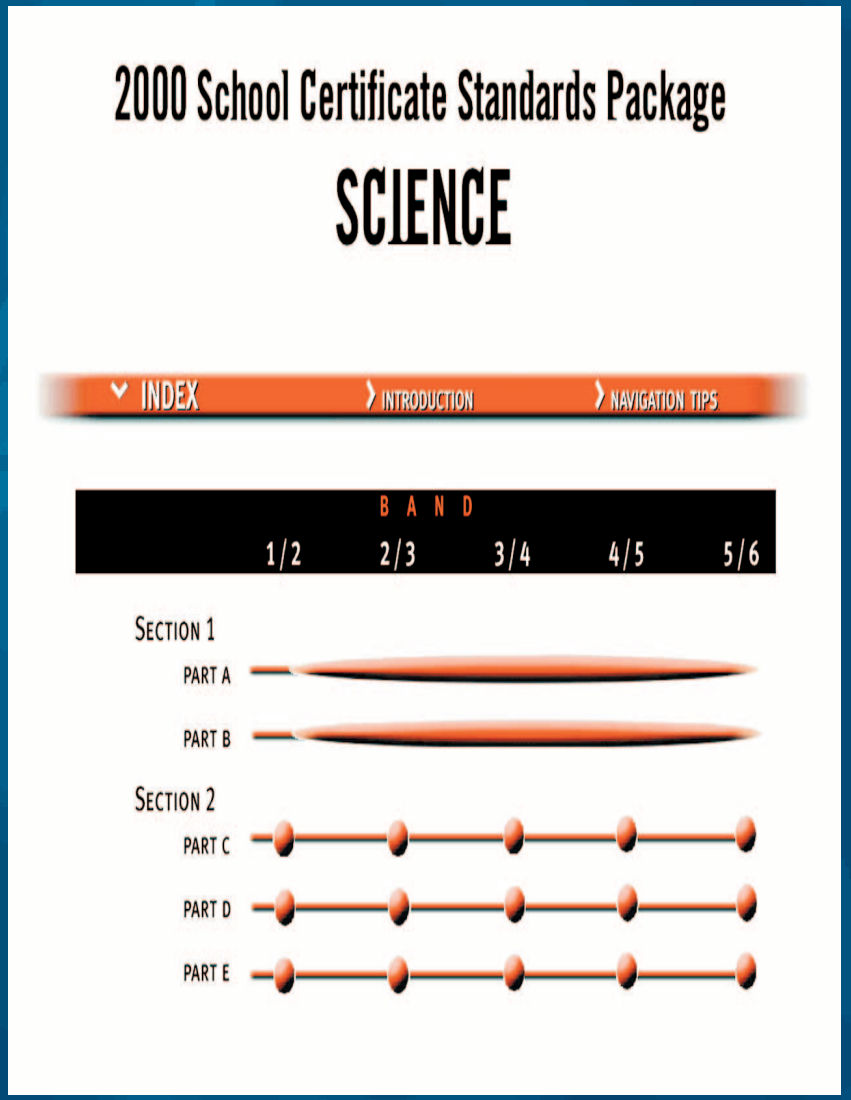
- Standards Packages will contain
  - the examination paper and marking guidelines
  - the Band Descriptions
  - sample responses of students who obtained the band cut-off marks for extended response questions, projects, performances
  - statistical information on the response patterns of students who obtained the band cut-off marks for objective questions

CAPTURING THE STANDARDS (cont.)

- These Standards Packages will be used by
  - the teams of judges in 2002, 2003, . . . to ensure they apply the SAME standards to measure and report student achievement every year
  - teachers and students to gain a clear understanding of the standards of performance applied in the HSC

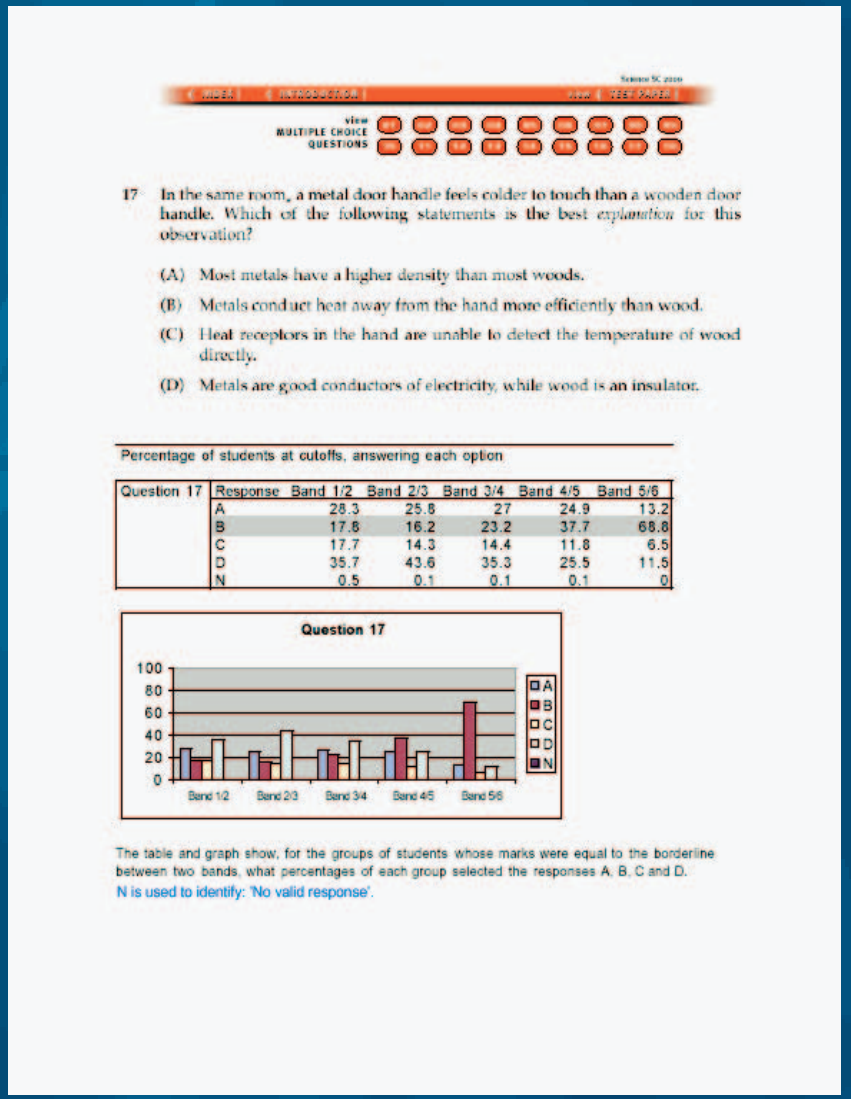
STANDARDS PACKAGES

- Some sample pages from the SC standards packages



STANDARDS PACKAGES

- **Some sample pages from the SC standards packages**



MULTIPLE CHOICE QUESTIONS

17 In the same room, a metal door handle feels colder to touch than a wooden door handle. Which of the following statements is the best *explanation* for this observation?

(A) Most metals have a higher density than most woods.

(B) Metals conduct heat away from the hand more efficiently than wood.

(C) Heat receptors in the hand are unable to detect the temperature of wood directly.

(D) Metals are good conductors of electricity, while wood is an insulator.

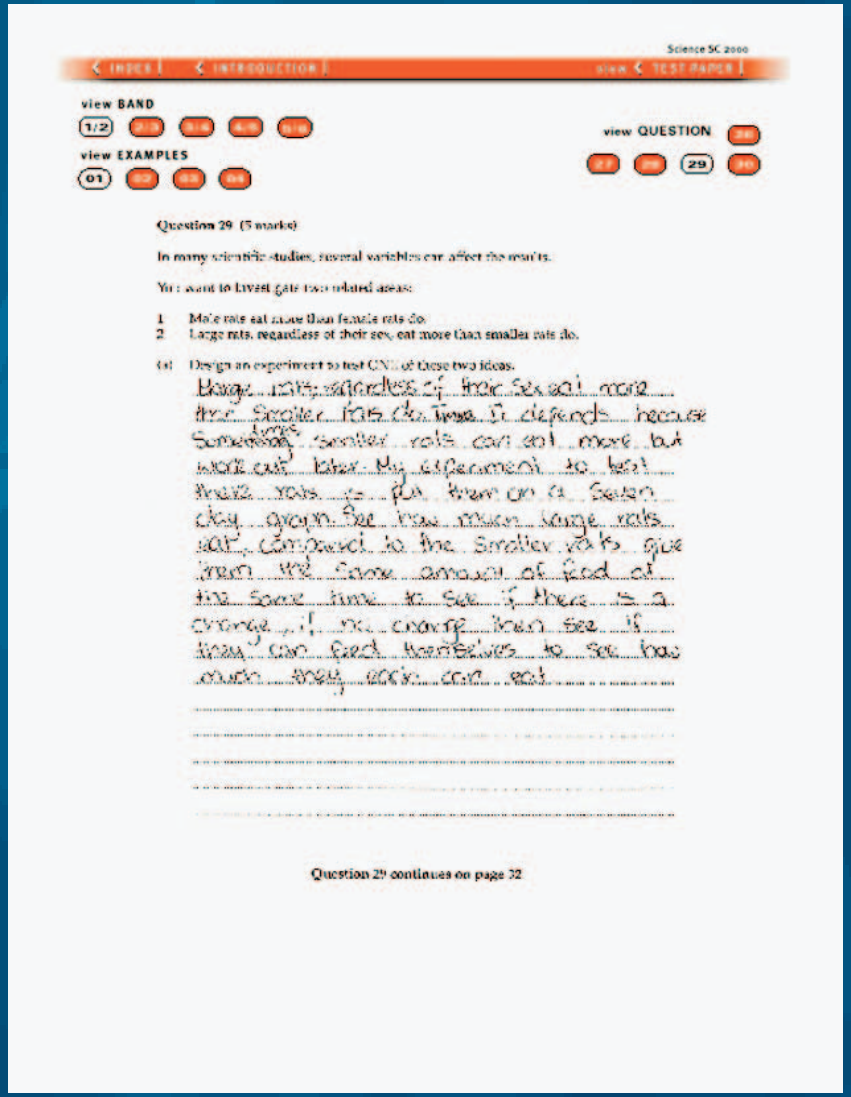
Percentage of students at cutoffs, answering each option

Question 17	Response	Band 1/2	Band 2/3	Band 3/4	Band 4/5	Band 5/6
	A	28.3	25.8	27	24.9	13.2
	B	17.8	16.2	23.2	37.7	68.8
	C	17.7	14.3	14.4	11.8	6.5
	D	35.7	43.6	35.3	25.5	11.5
	N	0.5	0.1	0.1	0.1	0

The table and graph show, for the groups of students whose marks were equal to the borderline between two bands, what percentages of each group selected the responses A, B, C and D.
N is used to identify: 'No valid response'.

STANDARDS PACKAGES

- Some sample pages from the SC standards packages



Science SC 2000

INDEX | INTRODUCTION | TEST PAPER

view BAND

1/2

view EXAMPLES

01

view QUESTION

29

Question 29 (3 marks)

In many scientific studies, several variables can affect the results.

You want to investigate two related areas:

1 Male rats eat more than female rats do.
2 Large rats, regardless of their sex, eat more than smaller rats do.

(a) Design an experiment to test ONE of these two ideas.

Question 29 continues on page 32

THE SCOPE OF THE TASK

- The standards-setting operation will involve
  - 114 separate standards-setting tasks
  - 88 teams of judges
  - 454 judges
  - 30 support staff

OFFICE OF THE BOARD OF STUDIES NSW

The Procedure for Aligning Student Achievement in the 2001 HSC Examinations to the Performance Scales

This section provides details of the procedures to be used in establishing the cut-off marks between the bands that are used in reporting student achievement in the course. The action of determining these cut-off marks will result in students' examination marks being aligned with the Performance Scale and so indicate the performance standard the student has reached. This procedure is shown in Figure 1.

The procedure involves using teams of judges (highly experienced teachers/markers) who determine the examination marks that correspond to the borderlines between one band and another. The procedure is a multi-staged one that gives the judges several opportunities to review and refine their earlier decisions. To inform their decisions the judges review statistical data and samples of student responses.

All judges will attend a Training Session prior to the commencement of the HSC written examinations. This training will focus on the process and what the judges need to do. At these training sessions certain course-specific issues will be investigated and resolved. For some courses that have practical or performance components that are marked before the commencement of the written examinations an additional session will be held earlier.

At the marking centre, once the marking schemes are finalised and final marking has commenced a Briefing Session will be held for each team at the marking centre. At this session the materials to be used including recording sheets, band descriptions and marking guidelines/schemes will not be used.

Stage 1

Statements (draft Band Descriptions) have been prepared that seek to describe the standards of performance in each course it is expected will be typically demonstrated by students who achieve the performance bands from 2 to 6.

In this first stage of the procedure each judge independently considers these Band Descriptions and develops "an image" of the type of student described. That is, each judge develops an understanding of the knowledge and skills typically possessed by students in each band. Once they have done this they refine those images so that they match students they believe would be on the borderline between two bands.

The judges then consider each examination question in turn. For questions that are scored dichotomously (ie, right or wrong) a judge records the probability that a borderline

student will get the question right. (Alternatively, they can answer the question – of ten such borderline students, how many are likely to get it right?) For questions which are scored polytomously (eg. short response items marked out of 5, essays marked out of 20) a judge records the mark he/she believes students at each borderline will receive.

Each judge then manually sums their question cut-off marks for each borderline and looks at the total cut-off marks he/she created to check that they are satisfied with the outcome. The judge might make minor adjustments to some question values at this point. (The cut-off mark between Band 5 & Band 6 is considered to be the lowest possible mark in Band 6. Similarly for other borderlines.)

The sheets on which the judges have recorded their decisions are collected. Each judge's question cut-off marks recorded by a judge are added by the computer to give the cut-off marks between each of Band 5 & Band 6, Band 4 & Band 5, Band 3 & Band 4, Band 2 & Band 3 and Band 1 & Band 2. Then, for the borderline between Band 5 & Band 6 the cut-off marks proposed by the judges for a question are averaged. The averages for all questions are added to obtain the initial test cut-off mark for Band 6. Any issues concerned with question weightings that must be applied and optional questions are taken into account prior to this addition. This process is repeated for all other bands.

Stage 2

The judges come together and compare the band cut-off marks they each have proposed for each question and for the total test.

At this stage the judges are also given statistical information on the performance of students on the test. The information is presented in a form that makes it easy for the judges to see how students at various ability levels have performed on the questions.

The judges review and discuss the statistical data and the decisions they have made individually.

During these discussions the judges have the opportunity to modify the question (and hence, total paper) cut-off values they recorded.

After all changes are made, the recording sheets are again collected and beginning with the Band 5/Band 6 borderline, the cut-off marks now proposed by the judges for a question are averaged. The averages for all questions are added to obtain a new test cut-off mark for Band 6. This process is performed for the other borderlines.

Stage 3

At the next step the judges consider a sample of student scripts that have been awarded marks equal to, above and below the cut-off marks established in Stage 2.

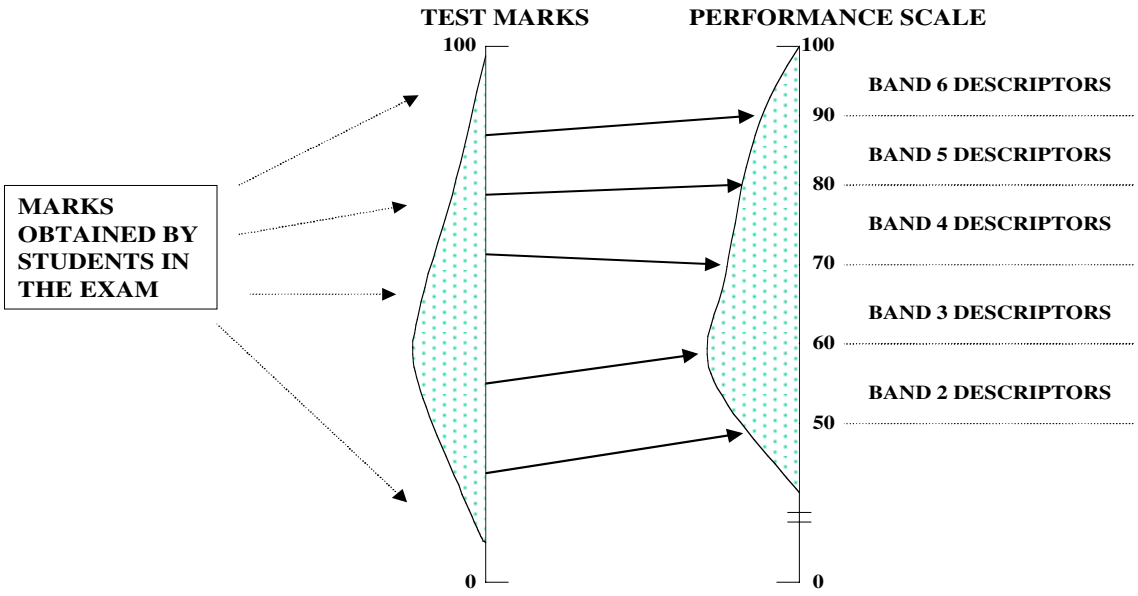
In reviewing these scripts the judges need either to confirm that the scripts awarded the cut-off marks they propose demonstrate levels of performance which are on the borderline between two bands, or to justify why a modification to a cut-off mark should be made. In essence, they need to explain why the scripts scoring the cut-off marks can be said to match, albeit only just, the descriptions corresponding to the bands they will be awarded.

Finally

The judges consider the band descriptions in light of the evidence relating to the standards of achievement demonstrated in the student scripts. They will identify any changes they believe are necessary and advise the HSC Consultative Committee accordingly.

At the conclusion of this process, when the judges have finalised the cut-off marks they will discuss the application of the procedure and the outcomes with the HSC Consultative Committee. The Consultative Committee, on behalf of the Board, will have the responsibility for approving the band cut-off marks to be used. These will then be applied to the examination mark distribution and bands will be allocated to students accordingly.

THE STANDARD-SETTING TASK



OFFICE OF THE BOARD OF STUDIES

HSC Standards Packages

The essential feature of a standards-referenced system of assessing and reporting student achievement being introduced by the NSW in 2001is that standards of performance corresponding to different levels of achievement in a course are established and student achievement is reported in terms of these same standards every year. It is thus possible to compare the performances of students who have studied a course in different years, even though they will have sat for entirely different examinations. Using a standards-referenced approach also allows for a very rich form of reporting where student achievement can be described in terms of statements summarising what students know and can do.

In order to introduce and operate a standards-referenced approach that can appropriately accommodate the nature of the curriculum and the nature of the examinations and tests used for the HSC and SC it is necessary to develop clear illustrations of the standards that will be used in each course. In this regard simple word descriptions of the knowledge, skills and understandings typically possessed by students who achieve each of the standard levels established for a course are not sufficient on their own to adequately and unambiguously define the performance standards. While such descriptions are an essential element of the standards, particularly in reporting student achievement, they need to be accompanied by two other elements. Firstly, there needs to be examples of the tasks students were required to do. In the case of the HSC and SC this means the examination questions and activities students were given. Second, there need to be samples of student responses and/or statistical data (as appropriate) that show how students at each possible standard level responded to those questions or activities.

Once the three elements - the descriptions, the tasks and the sample student responses and data - are brought together into an integrated package we will have a clear illustration of the performance standards associated with a course.

- Teachers and students then can develop a clear understanding of what is required of students in order to achieve each performance standard in a course.
- Those markers and others responsible for applying the Board's standards-setting procedure each year can use these standards packages to internalise the standards they are to apply when determining the band cut-off marks.

For each HSC course the standards package will consist of a CD-ROM containing the

- *Band Descriptions* (the summary of knowledge, skills and understandings typically demonstrated by students who achieve each band),

- *Examination Paper* (showing each question students were required to answer or task they were required to perform),

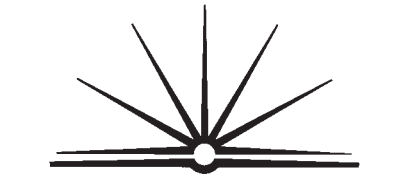
- *Samples of student responses for those parts of the examination requiring students to produce a response.* These responses illustrate the nature and quality of the responses typically produced by students whose marks in the examination placed them at the borderline between each pair of bands. This material is organised in such a way that if a particular section or question is selected, and then a particular borderline selected, it is then possible to view or hear the responses of a sample of students at that borderline.

  - For each section or question (as appropriate) on the examination requiring a written or verbal response or a performance the responses of several students, whose work was of the standard typical of students placed at the borderline between band 5 & band 6 will be provided. Similarly for the other borderlines.
  - In the case of those examinations that contain a musical, dance or dramatical performance videos of these performances along with comments from markers will be incorporated
  - In the case where students are required to produce a product such as a body of artwork, a design project or the like, images of these works will be provided along with comments from the markers.

- *Tables and graphs for questions where students are required to select an answer, such as multiple choice and true/false items*. These tables provide statistics on the response patterns of students. This material is organised in such a way that for the groups of students whose marks are equal to the borderline between two bands, the percentage of each group that selected each of the responses A, B, C and D is given. By reading the question and then looking at this student response data teachers will develop an understanding of how well students at each borderline answered each question, and importantly, the types of errors students tended to make. This analysis will help to provide a "picture" of the level of knowledge and skills typical of students at each level of performance.

Where there are less than 1000 HSC students studying the course the standards package will have a similar structure, but the samples of student responses will be fewer. For very small candidature courses only a few student responses, together with examiner comments will be provided.

It is planned that the standards packages will begin arriving in schools during Term 1 2002.



Occasional Paper

Setting Standards and Applying Them across Different Administrations of Large-scale, High-stakes, Curriculum-based Public Examinations

A paper summarising key issues, research and practices associated with setting performance standards in examinations and their implications for the setting of standards in the NSW Higher School Certificate program

John Bennett

November 1998

Source: J Bennett, PhD thesis, UNSW 1998

Published by Board of Studies NSW
GPO Box 5300
Sydney NSW 2001
Australia

November 1998

Tel: (02) 9367 8111
Fax: (02) 9367 8484
Internet: http://www.boardofstudies.nsw.edu.au

ISBN 0 7313 4163 5

Introduction

The NSW Government's HSC White Paper introduces a standards-based approach to assessment and reporting of student achievement for both the Higher School Certificate and the School Certificate. This is a significant departure from norm-referenced approaches that have been used in these programs in the past. The successful implementation of standards-based approaches requires that careful consideration be given to adapting procedures and strategies that have succeeded in other contexts to suit the particular needs of the NSW programs.

This paper identifies key issues associated with the setting of performance standards in public examinations using informed professional judgment, and the application of these standards across different administrations of the examination. It draws upon the research literature that identifies activities and features commonly accepted as essential ingredients of any creditable standard-setting procedure. The paper also considers procedures used in the application of standard-setting approaches in other major curriculum-based public examination programs conducted elsewhere in the world. Specifically, the GCE A-level examinations, the Standard and Higher Grade examinations conducted by the Scottish Examinations Board and the International Baccalaureate examinations are addressed. These examination programs are similar to the HSC in that an entirely new examination is used at each administration. Hence, the use of the same examination paper or standard statistical approaches to linking standards across time are unsuitable.

Included at the end of this paper are two flow charts showing a multistage standard-setting procedure used in a research study (Bennett, 1998). The first shows the procedure as it would be applied in an initial year when the standards were being developed. The second shows how the procedure would operate in a second and subsequent years to place the performances of students in those years on the same performance scale.

1. Standard-setting Using Informed Professional Judgment – Lessons from the Literature

The terms 'standards', 'performance standards', 'standards of performance' and 'achievement standards' are used interchangeably in this paper 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'. 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' 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).

Although Waltman refers to 'descriptions of knowledge, skills and abilities', a description is not sufficient on its own to clearly articulate a performance standard. A written description can provide a useful summary of a standard. To properly specify a standard, however, the description must be supported by examples of the tasks students need to perform and samples of student responses to those tasks which clarify and exemplify the performance standards summarised in the descriptors. This material can be referred to as the 'standards package' (Bennett, 1998).

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.

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.

1.1 The Use of a Structured Multistage Approach

While early standard-setting procedures (eg Nedelsky, 1954; Ebel, 1972) 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 during following stages.

Various researchers (eg Jaeger, 1982; Cross, Impara, Frary and Jaeger, 1984; Cizek, 1996; and Berk, 1996) advocate the use of a structured, multistage 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' (p 100).

1.2 The Selection of Judges

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

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' performances 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 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. Berk (1996) states that a broad-based panel of the most qualified and credible judges should be selected.

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 some 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. In addition, 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.

1.3 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.

Reid (1991) argues that judges must not only 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 suggests three criteria that 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. For example, in a situation where a 'pass/fail' cut-off score is to be determined, 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. Their view is that:

> '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 follow during the stages of the procedure when they are working individually.

1.4 The Initial Steps in Establishing Cut-off Scores

Numerous studies conducted over the past 20 years indicate that judgmental-empirical standard-setting procedures built upon the Angoff (1971) approach give the most acceptable outcomes. In fact, most of the current procedures that use teams of judges to set performance standards are refinements and extensions of the Angoff method. Judgmental-empirical methods are those which use professional judgment supported by empirical data.

Fehrmann, Woehr and Arthur (1991) found 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.

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.

1.5 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 standard-setting procedures used today, judges generally arrive at their decisions individually and then meet to discuss their decisions with their colleagues. During this discussion the judges are given the opportunity to vary their own decisions if they wish. It is usual then for the average of the decisions of the individual judges 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) 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).

1.6 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 that can be provided.

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. Bennett (1998) also identified this possibility. 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.

In a number of recent studies, researchers have analysed student performance data using latent trait models and then provided judges with this information in a variety of ways to inform the standard-setting process (eg McGaw, 1997; Englehard and Gordon, 1997; Bennett, 1998). Bennett has shown that student performance data analysed in this way can be presented in a manner which provides powerful support to judges involved in setting standards in examinations of the type used in the NSW Higher School Certificate program.

By examining a sample of student scripts that 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.

1.7 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 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. Standard-setting Using Informed Professional Judgment – Lessons from Other Examination Programs

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 that 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.

2.1 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 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).

2.3 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 that 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 student scripts that 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).

3. Some Parameters and Specific Issues

For the type of examinations used for the HSC, standard-setting procedures need to take account of the following:

- *Profile of the Judges*. The judges engaged in the standard-setting task must be highly experienced in teaching the course and preparing students for the examination. Preferably, they have been involved

in marking student responses so that they have a good understanding of the range of responses typically produced by students. Others who will be associated with the students at the next stage, such as university lecturers, may also provide a useful service.

- *Size of the Teams*. As a general rule, the team should consist of approximately six judges. Teams of this size will enable the necessary detailed discussion and debate to occur, which contributes to the integrity of the process. With large teams, discussion of individualised decisions can become superficial, with judges 'jumping to a compromise position' simply to hurry the process along. It may be that a structured procedure is established so that many more people have input during the preliminary stages. The advice received from this large group can then be fed into the detailed discussion and debate required to ensure the integrity of the process undertaken by the smaller team. It is essential, however, that, between them, the judges have adequate understanding of and expertise in all aspects of the course being examined. If this is not possible, other strategies must be in place to overcome any lack of expertise or understanding on the part of the judges.

- *Use of Student Performance Data Derived from Latent Trait Analysis*. The provision of feedback to judges on student performance resulting from the use of Latent Trait models can be a powerful ingredient of the standard-setting process (Bennett, 1998). While this form of analysis and presentation of data provides insights into aspects of student performance that may not be evident from other sources, there is value in providing the judges with a variety of different data on student performance, including those produced by classical methods.

- *Use of Student Scripts*. Providing judges with samples of student scripts is an important element in the process. The judges can best use these to validate or refine earlier decisions they have made. Given the scope for variation in difficulty, emphasis and format of HSC examinations from year to year, it can be a challenging task to determine whether student scripts from different years represent the same standard of performance. It is important, however, that these judgments be made. Equally, it is important that the sample of scripts reviewed be judged as to whether they fit the descriptors for the bands in which the score awarded would place them.

- *Use of a Compensatory Approach.* For the type of examinations used in the HSC program, a compensatory approach is 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, using a conjunctive approach and 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 markers 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.

- *Use of a Compromise Approach*. In a standards-based system, once the performance scale has been established (calibrated), the scale remains the same from year to year. That is, the requirements to achieve a particular standard are the same across different administrations of the examination. More or fewer students may achieve that standard from year to year, but the standard itself remains fixed.

  In the initial year when the standards are being established and the scale created, however, it is quite appropriate to decide that the cut-off scores will be set so as to place certain proportions of that initial candidature into the various standard levels. Unguided, judges tend to set cut-off scores that reflect unreasonably high expectations of students (Bennett, 1998; Wiliam, 1996). In the initial year, setting the cut-off scores so that acceptable proportions of that candidature fall into each

standard ensures that the scale is established in such a way that each standard level on the scale can contribute to meaningful reporting of the range of student achievement.

It is essential, if this is done, that the standards then be summarised in the descriptor statements and clarified and exemplified by the examination tasks and the sample examination scripts at the cut-off points. After this initial year, cut-off scores are set using this material and not by seeking to place proportions of students in each standard level.

Conclusion

In undertaking the task of introducing a standards-based approach to the assessment and reporting of student achievement in the HSC program, there are important lessons to be learnt from the literature on standard-setting and equating, and from studying the methods used elsewhere in other large-scale, high-stakes, curriculum-based public examination programs.

The viability of using teams of suitably qualified judges to set standards and to link them over time to other forms of the examination is questioned by some, who see it as lacking the rigour and precision of more empirical approaches. The studies and practices referred to in this paper identify the key issues that must be considered in such an exercise. The paper then provides strategies and parameters that need to be built into a standard-setting procedure to deliver the required level of integrity and validity.

Berk (1996) encapsulates this position in noting that the validity of a judgmental standard-setting procedure is dependent upon the expertise and experience of the judges and the application of the procedure itself. The credibility of the group of judges and the fidelity of the procedure can result in our being prepared to 'accept the judges' decision'.

The Standard-setting Process and Linking Standards across Time

Source: J Bennett, PhD thesis, UNSW 1998

Source: J Bennett, PhD thesis, UNSW 1998

References

Angoff, W. (1971) 'Scales, Norms and Equivalent Scores'. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp 508-600), American Council on Education, Washington, DC.

Bennett, J. (1998) 'A Procedure for Equating Curriculum-based Public Examinations Using Professional Judgment Informed by the Psychometric Analysis of Response Data and Student Scripts'. Unpublished doctoral thesis, University of New South Wales.

Berk, R. (1996) 'Standard Setting: The Next Generation'. *Applied Measurement in Education*, 9, 215-235.

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.

Ebel, R. (1972) *Essentials of Educational Measurement* (2nd ed.). Englewood Cliffs, NJ.

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.

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.

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. (1991) 'Selection of Judges for Standard-setting'. *Educational Measurement: Issues and Practice*, 10, 3-14.

Kane, M. (1986) 'The Interpretability of Passing Scores'. *American College Testing Program Technical Bulletin No. 52*. Iowa.

Linn, R. (1978) 'Demands, Cautions and Suggestions for Setting Standards'. *Journal of Educational Measurement*, 15, 301-308.

McGaw, B. (1997) *Shaping Their Future: Recommendations for Reform of the Higher School Certificate*. Department of Training and Education Co-ordination, New South Wales.

Mills, C., Melican, G. and Ahluwalia, N. (1991) 'Defining Minimal Competence'. *Educational Measurement: Issues and Practice,* 10, 7-10.

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. 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.

Popham, W. (1978) 'As Always Provocative'. *Journal of Educational Measurement*, 15, 297-300.

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*. School Curriculum and Assessment Authority, London.

Scottish Examination Board. (1996) *Handbook for Examinations 1996*. Scottish Examination Board, Dalkeith.

Waltman, K. (1997) 'Using Performance Standards to Link Statewide Achievement Results to NAEP'. *Journal of Educational Measurement*, 34, 101-121

Wiliam, D., (1996). 'Meanings and Consequences in Standard Setting'. *Assessment in Education,* 3, 287-307.