An investigation of the comparability of teachers' assessments of student folios

Geofferey N. Masters
Bruce McBryde

Research Report Number 6

TEPA
Tertiary Entrance Procedures Authority
April 1994

An Investigation of the Comparability of Teachers' Assessments of Student Folios

Geofferey N. Masters
Australian Council for Educational Research

Bruce McBryde
Tertiary Entrance Procedures Authority

April 1994

Report of a study commissioned by the Tertiary Entrance Procedures Authority, Qld.

ISBN 0 646 18557 8

Other research publications available from TEPA are:

Research Series: Monographs and reprints

Monograph 1: Pathways to University in Queensland, TEPA Research Series Number 1, December 1992.

Admission of Students into Higher Education in Australia: A Collection of Recent Research Papers, TEPA Research Series Number 2, January 1993.

Research Reports

Year 11 subject selection: Issues and trends, (1991), Research Report Number 1, TEPA, Brisbane.

Student subject selection, (1993), Research Report Number 2, TEPA, Brisbane.

Preparing students for the Queensland Core Skills Test, (1993), Research Report Number 3, TEPA, Brisbane.

Factors associated with the rejection of TAFE offers through QTAC in the 1991–92 tertiary admissions period, (1993), Research Report Number 4, TEPA, Brisbane.

The Student Education Profile: Studies of community perceptions, (1993), Research Report Number 5, TEPA, Brisbane.

Surveys and Studies

Changes in QCS Test preparation 1992–93 (1993), Surveys and Studies No. 1, TEPA, Brisbane.

Rejecting offers of places in TAFE and universities – 1993 (1994), Surveys and Studies No. 2, TEPA, Brisbane.

1994 applications through QTAC Ltd (1994), Surveys and Studies No. 3, TEPA, Brisbane.

1994 interstate applications to QTAC (1994), Surveys and Studies No. 4, TEPA, Brisbane.

School principals' reports on Year 12 students: Possibilities for use in tertiary entrance selection (1994), Surveys and Studies No. 5, TEPA, Brisbane.

Tertiary Entrance Procedures Authority
PO Box 171
Brisbane Albert Street Q 4002
Australia

Telephone (07) 234 1498 or 1 800 804 991
Fax 61 7 234 1508

Contents

Foreword

In her 1990 review of tertiary entrance procedures in Queensland, Professor Nancy Viviani recommended that a newly-established Tertiary Entrance Procedures Authority (TEPA) institute major independent research into the comparability of assessments in Years 11 and 12 in Queensland schools. This recommendation followed Viviani's observation that there was a widespread lack of public confidence in the comparability of school assessments in that State.

Following its establishment, the Tertiary Entrance Procedures Authority developed an information statement summarising Viviani's analyses and listing types of research into the comparability of school assessments that TEPA might fund. Researchers were invited to submit expressions of interest to undertake investigations in one or more of the identified areas.

None of the proposed projects, however, addressed the central issue of the comparability of assessments across schools. To address this issue TEPA's research committee invited Professor Royce Sadler from Griffith University to prepare an overview paper on issues of comparability. Dr Geoff Masters, Associate Director (Measurement) at the Australian Council for Educational Research, was asked to liaise with Professor Sadler and to submit a research proposal to address more directly the question of the comparability of school assessments in the Queensland upper secondary school.

ACER's submission proposed the re-assessment of a sample of assessment folios submitted by Year 12 students in 1992. Each folio would be re-assessed by a number of assessors independently and without knowledge of that folio's final (exit) assessment at the end of 1992. An advantage of working with 1992 folios was that this would be the first cohort of students to be selected for higher education using the Student Education Profile (SEP) foreshadowed in the Viviani review. The details of the design for the re-assessment were further refined by Geoff Masters, Royce Sadler, and Bruce McBryde of the Tertiary Entrance Procedures Authority.

A sample of folios in English, Chemistry, Mathematics I and Modern History was drawn at the end of 1992 and re-assessed during 1993 by assessors nominated by chairs of the Board of Senior Secondary School Studies' District Review Panels in the four subjects. The collection and re-assessment of folios was organised by Bruce McBryde. Data collected through the study were then forwarded to ACER for analysis and interpretation.

The authors wish to acknowledge the valuable assistance of staff of the Board of Senior Secondary School Studies, the professional assistance of the 62 assessors involved in the study, the role of Mrs Lyris Mitchell in coordinating and supervising the assessment process, and the assistance of Mr Peter Congdon of the Australian Council for Educational Research.

Geoff N. Masters
Bruce McBryde

Executive Summary

Year 12 results in Board-accredited subjects in Queensland are reported on the Senior Certificate using five *Levels of Achievement*: Very High (VHA); High (HA); Sound (SA); Limited (LA); and Very Limited (VLA) Achievement. Levels of Achievement are awarded by teachers on the basis of folios of students' assessment work. These assessments are 'criterion based' in the sense that each Level of Achievement is an attained standard of performance in a subject, rather than an achievement level allocated to some pre-determined percentage of students. The Board of Senior Secondary School Studies verifies schools' assessments through a district review process in which each school submits student folios representative of its allocated Levels of Achievement together with the school's accredited work program and details of its assessment procedures.

An issue of ongoing debate is the question of the comparability of Levels of Achievement within a subject across schools. Does a VHA in chemistry from one school represent the same level of chemistry achievement as a VHA in that subject from any other school? In her 1990 review of tertiary entrance procedures in Queensland, Professor Viviani concluded that there was a general lack of public confidence in the comparability of school-based Levels of Achievement and recommended 'independent research' to provide evidence on how comparable assessments are across schools. The present study was commissioned by the Tertiary Entrance Procedures Authority as part of an ongoing program of research to gather such evidence.

A sample of 546 student assessment folios in Chemistry, English, Mathematics I, and Modern History was drawn from 62 schools throughout Queensland. These folios were exit folios at the end of 1992. Six independent assessments of each folio were then made by assessors familiar with the Board's review processes and nominated by chairs of District Review Panels in each subject. A total of 62 assessors took part in the study. Their assessments of each folio were compared with each other and also with the exit Level of Achievement allocated to that folio. All assessments were made on a 50-point scale obtained by dividing each of the five Levels of Achievement into ten 'rungs'.

A question of interest in the study was whether assessors would provide more reliable assessments when given access to the relevant school work programs, or whether Levels of Achievement would be allocated just as reliably without access to work programs. A second question was whether assessors would be able to judge folios more reliably when considering work from the same school (and thus based on the same set of assessment tasks) or whether they would display a similar level of agreement on folios sampled randomly from a number of schools and assessment programs. To explore these questions, each folio was assessed under three conditions: without the associated work program but with other folios from the same school; with the work program and with other folios from the same school; and without the work program and as part of a random selection of folios from different schools. Two independent assessments were made under each condition.

The results of the study reveal an exceptionally high level of agreement (correlation = 0.94) between assessors' scores on the 50-point scale. Approximately 98 per cent of paired assessments differed by less than one Level of Achievement (10 rungs); 90 per cent differed by less than half a Level of Achievement (5 rungs). A similarly high level of agreement (correlation 0.92) was obtained between the assessments made in this study and the original exit Levels of Achievement allocated to these folios. These levels of agreement are significantly higher than the levels of inter-marker reliability typically

reported for independent assessments of student work—including independent assessments of examination performances.

When assessments for each school subject were considered separately, some differences between subjects emerged. Assessors displayed lower levels of agreement with each other in English and Modern History (correlations of .89 and .92) than in Chemistry and Mathematics (.96 and .97).

Although it was anticipated that access to schools' work programs would increase assessors' understandings of the assessment context and so might increase levels of inter-marker agreement when assessing student work, there was no evidence of this occurring in practice. Assessors displayed the same levels of agreement with and without school work programs.

It was similarly anticipated that two assessors considering folios from the same school (and thus based on the same assessment program) might display higher levels of agreement in their allocation of Levels of Achievement than two assessors considering folios from a variety of schools and assessment programs. Again, this was not the case. Assessors displayed the same levels of agreement whether considering folios from the same school or from a number of different schools.

Given the unusually high levels of inter-marker agreement obtained in this study, it is important to note the differences between this study and most other reported inter-rater reliability studies. In most studies, each assessment of a student's work is made without knowledge of any other assessor's judgement of that work. Most studies of this kind are based on holistic judgements of written essays, folios of artwork, oral presentations, and performances of various types (e.g., dance, music performance, gymnastics). In the present study, assessments were made independently of each other, but each assessor had access to teachers' prior assessments of folios. Many folios consisted of collections of marked tests and assignments. Numerical scores were available to assessors, as were written comments made by teachers, although care was taken to remove any indication of teachers' allocated Levels of Achievement. In view of the marks and comments available to assessors of each folio, it is probably not surprising that the levels of inter-marker agreement were higher than usual.

Nevertheless, the results of this study indicate an exceptional level of agreement among assessors undertaking tasks very similar to those involved in the Board's review process. On the basis of this study there is reason for confidence that Board assessors are applying very similar standards in their review of student folios from different schools and in their recommendations concerning Levels of Achievement.

As well as displaying a high level of agreement with other assessors of the same folios, the assessors in this study also displayed a remarkable level of confidence in their own abilities to assign Levels of Achievement to folios. When asked to indicate an upper and lower bound (on the 50-point scale) for their assessment of each folio, assessors expressed a confidence of ±2 rungs for more than 80 per cent of their assessments. This level of marker confidence, coupled with the high level of inter-rater agreement, raises a question about the current practice of reporting assessments on only a 5-point scale.

1. The Comparability Issue

Comparability of assessments across schools has been a contentious issue in Queensland since the adoption of the Radford Report in 1970.

Radford review

The Radford Report was the result of a review of the system of public examinations in Queensland secondary schools commissioned in July 1969 by the Queensland government. The seven-member review committee chaired by W.C. Radford recommended the abolition of the Junior and Senior examinations taken at the end of Years 10 and 12 and replacing them with a system of moderated teacher assessments. The implementation of the recommendations of the Radford review resulted in scores on external examinations being replaced by teacher assessments expressed on a scale of 1 (lowest) to 7 (highest) in each subject. The statewide distribution of ratings in each subject was constrained to predetermined percentages (3%, 7%, 15%, 40%, 20%, 10%, 5%), and was therefore clearly norm-referenced in principle. In an attempt to establish comparability of teacher assessments across the State, a system of consensus moderation was introduced with teachers meeting to reach agreement on the award of the seven Levels of Achievement to student work.

Review of school based assessment (ROSBA)

The Queensland Board of Secondary School Studies commissioned a review of the implementation of the Radford recommendations in February 1976. That review, chaired by E. Scott and known as the Review of School-Based Assessment (ROSBA), recommended a number of changes to the system of teacher assessments, including the introduction of five Levels of Achievement to replace the earlier seven point scale:

Very High Achievement	(VHA)
High Achievement	(HA)
Sound Achievement	(SA)
Limited Achievement	(LA)
Very Limited Achievement	(VLA)

Under the proposals of ROSBA, these Levels of Achievement were to be based not on fixed percentages of ratings, but on specified criteria for each level. The ROSBA committee described teacher assessments made with reference to criteria as 'competency-based'. Criteria were to be provided in general terms in Board syllabus statements and interpreted in detail by schools in their work programs:

> The performance criteria for each level of competence in each subject should be stated in broad terms by the respective syllabus advisory committees and in specific terms by the schools in their accredited programs. In designing their work programs schools would be required to state: (i) their specific cognitive content, practical skill and affective objectives for a particular unit of work in a syllabus; (ii) the assessment techniques and exercises which would be used to determine achievement of those objectives; and (iii) the criteria to be used to determine the student's level of competency. (Scott, *et al.*, 1978, 38)

Viviani review

In December 1989 the Queensland Government abolished the tertiary entrance score as the principal basis for tertiary selection in that State. Prior to this decision, school assessments had been statistically rescaled against an external test (the Australian Scholastic Aptitude Test, together with, in more recent years, a writing task) and the

resulting scores combined to provide a single tertiary entrance (TE) score in terms of which all applicants to higher education institutions were ranked. Following this decision, Professor Nancy Viviani was commissioned to review tertiary entrance procedures and to make recommendations for a new system to replace the TE score.

The Viviani Report resulted in the establishment of the Tertiary Entrance Procedures Authority (TEPA) and the introduction of a Student Education Profile (SEP) to replace the earlier TE score.

The SEP is made up of: (i) results reported by the Board on the Senior Certificate (i.e., subject results in the form of Levels of Achievement, and results on the Queensland Core Skills Test) and; and (ii) results reported by TEPA on the Tertiary Entrance Statement (i.e., aggregates in the form of Overall Positions and Field Positions).

Under the new arrangements, tertiary selection is based on students' Overall Positions (OPs) and Field Positions (FPs). To calculate OPs and FPs, the Board asks schools to provide independently of the Levels of Achievement, a 'Subject Achievement Indicator' for each student on a scale of 200 (lowest achiever in the subject in that school) to 400 (highest achiever in the subject in that school). These are then scaled against students' performances on the Queensland Core Skills Test and aggregated in various combinations to provide OPs and FPs.

The Levels of Achievement in each subject reported on the Senior Certificate differ from the Subject Achievement Indicators and the FPs that they are used to produce in that they are not rank orders. Rather, they provide information about a student's achievement in each subject and are obtained by comparing that student's achievement with *criteria* specified in the relevant Board syllabus.

The question thus arises of how comparable these Levels of Achievement are from teacher to teacher and from school to school. This question was asked by Viviani who concluded that, at the time of her review, there was widespread concern about the comparability of school assessments. She concluded that these concerns were undermining public confidence in the validity of assessment practices in schools and selection procedures to universities.

The Viviani Report identified a number of concerns:

- Students are concerned when there is an apparent inconsistency between the Levels of Achievement (VHA, HA, etc) in Board subjects and their TE scores, particularly when cases arise of students with lower Levels of Achievement obtaining higher TE scores.
- Teachers (and students) are concerned about the extent to which their assessments are comparable across schools.
- Universities are concerned about the comparability of teachers' assessments because these could be used as a basis for entry into some courses.
- Other groups, particularly employers, are concerned about the comparability of teachers' assessments because they use this information in choosing among job applicants.

Viviani was concerned that the lack of confidence in the comparability of Levels of Achievement may have been contributing to the overuse of the TE score and the finer-grained information from which it was constructed. She also concluded that there were difficulties in relying on statistical scaling alone to solve the comparability problem and called for a three-pronged approach to the issue based on processes of accreditation, moderation and statistical scaling used in combination.

Viviani proposed that the issue of comparability of Levels of Achievement should be tackled directly to achieve two objectives:

- to identify a process to improve the comparability of teachers' assessments; and

- to move to a situation where improved comparability of Levels of Achievement within subjects across schools would give universities the option of making more use of this information in their selection processes.

The Viviani review concluded that there was insufficient information available on how comparable Levels of Achievement are across schools and that research should be undertaken to provide this information. The report made recommendations for two areas of research, including research into the comparability of current school assessments:

> TEPA should institute immediately major independent research into the comparability of assessment in Years 11 and 12 in schools. This research should provide an answer to the question of how comparable assessment outcomes are across schools, and should provide a benchmark for future research and policy action by TEPA. In addition, the Board should be funded to carry out research on assessment practices now and for the future. (Viviani, 1990, 52-53)

The present study has been commissioned by the Tertiary Entrance Procedures Authority as part of the 'major independent research' into the comparability of teachers' assessment in Queensland schools called for by Viviani.

2. Current Procedures

Before outlining the research questions addressed in this study it is useful to describe briefly the procedures used by the Board of Senior Secondary School Studies to develop syllabuses, to monitor their use in schools, and to certify Levels of Achievement for students. Central to these activities of the Board are the processes of *accreditation* and *certification*.

The Board has established Subject Advisory Committees to cover subjects in a diverse range of disciplines appropriate to students in Years 11 and 12. These committees are broadly representative of key stakeholder groups, including universities, schools, and the business community in Queensland. Each committee is responsible for the development and/or updating of Board syllabus statements in one or more subjects within a general discipline area. The Mathematics Subject Advisory Committee, for example, is responsible for the syllabuses in Mathematics A, Mathematics B and Mathematics C. At present there are over 50 syllabuses available to Year 11 and Year 12 students in Queensland.

Each syllabus is developed and revised through a cycle of development, trialing in a few schools, pilot testing in a larger number of schools, and general implementation in all schools. After a syllabus has been in operation for a number of years, it is reviewed, and if necessary the process of development of a new syllabus is undertaken.

Schools wishing to offer a subject for which there is a Board approved syllabus must develop a school *work program*. The work program describes in detail how the school intends to implement the relevant Board syllabus. It must outline the program to be used to assess student achievement and specify the criteria to be used in determining the Levels of Achievement to be awarded to students. The school's *assessment program* describes in some detail the number and type of assessments (examinations, projects, assignments, etc), the conditions under which they are to be administered (timed examination, take home, in class) and the weightings to be given to each assessment element. Work programs must be accredited by the Board. Although specific details in work programs vary from school to school, the work program of each school must be consistent with the Board syllabus. The process by which a school's work program in each Board subject is approved by the Board is known as *accreditation*.

In October each year, schools are required to submit the distribution of their estimates for Levels of Achievement of all students studying each Board subject, based on student achievement on the assessment program to that time. These estimates are made on a 50-point scale, 10 rungs for each of the five Levels of Achievement and are recorded on the Board's Form R6. Schools also are required to assemble the assessment folios of nine students in each subject for review by a panel composed of experienced teachers from schools in that district. The nine students are selected on the basis of the school's proposed distribution and must include the top student in the school, the student just above each boundary between adjacent Levels of Achievement and a student in the middle of each Level.

The review panel scrutinises each folio to establish whether the proposed school assessment is consistent with the standards specified in the school's work program. The panel also checks to ensure that important features of the student folios--such as the range and difficulty of assessment tasks, the marking schemes and the weighting of elements--are consistent with the school's work program. During this process, members of district panels may collaborate with other panel members if they are unsure about their judgements. Thus atypical cases and apparent anomalies usually are considered by several members of a review panel as well as by the district panel chair before a recommendation for change is made to the school concerned.

If the review panel does not agree with a school's proposed Levels of Achievement, then it provides advice to the school on the R6, outlining the reasons for the decision and specifying the levels which the panel considers appropriate. If the school accepts the recommendations of the district panel, no further action is required and the school must base its final distribution of exit achievement on that recommended by the panel. If the school does not accept the advice of the district panel, it must continue negotiations with the Board. If these negotiations do not lead to resolution of the differences in proposed Levels of Achievement, then the matter is forwarded to the State review panel for resolution. Typically, the school is required to provide the assessment folios of a wider sample of students. The process by which school decision-making on standards of student achievement is verified by the Board by matching student performance to stated criteria and standards in the accredited work program is known as *certification*.

The exit Level of Achievement awarded by a school to a student is based on the student's assessment folio in that subject at the end of the year and is expressed on the 5-point scale (VLA, LA, SA, HA, and VHA).

3. Research Questions

The central research question in this study concerns the comparability of assessments made by teachers in different schools: To what extent do teachers in Queensland schools apply the same standards in assessing student folios?

This question is addressed by asking a group of teachers with experience and a detailed knowledge of the assessment process in Queensland to assess independently a sample of student folios. Three broad questions are then asked:

- to what extent do these assessors apply the same standards when assessing student folios?
- how consistent are these assessors' judgements with the exit Levels of Achievement assigned to this sample of folios at the end of 1992?
- to what extent can differences between assessments made in this study and exit Levels of Achievement be attributed to differences in standards across schools?

3.1 Consistency of assessors

The central question in this study is the question of inter-rater reliability. If teachers have common understandings of the criteria for each of the five Levels of Achievement, and they apply those criteria consistently to student work, then assessments of the same folio by different teachers should be in close agreement.

Even in situations—such as the marking of external examinations—where markers have worked together for a number of years as members of a marking team, have developed shared understandings of criteria, and have had opportunities to exchange opinions about the quality of pieces of student work, perfect agreement among markers is an ideal that can only ever be approximated in practice. The usual approach to describing inter-marker agreement is to calculate a product moment correlation between two markers' assessments of different pieces of student work. The crucial research question is not whether there is variability across assessors, which there always is, but whether the variability is within acceptable bounds.

What are acceptable bounds?

Given that some level of disagreement among assessors is inevitable, the question of what constitutes an acceptable level of disagreement becomes a central issue in this study. There are at least two different approaches to addressing this question. The first is to establish an *a priori* level of acceptable disagreement, based perhaps on a consideration of the consequences of such disagreement. If assessors are asked to rate student folios on a 50-point scale (10 rungs within each of 5 Levels of Achievement), for example, then it might be considered that a tolerable level of disagreement would see ratings of the same folio differing by more than a full Level of Achievement (10 rungs) less than five per cent of the time. Such a 'rule' might be based on a decision that it is not unreasonable to accept a difference of ten rungs for one in every twenty students, but not more frequently than this. A more demanding rule would be less accepting of a disagreement of 10 rungs, allowing it to occur less often--once in every 50 students, perhaps.

The usual way to report levels of agreement is in terms of the correlation between the ratings of two assessors. However, it is more difficult to interpret and set acceptable levels of correlation *a priori*.

The second approach to deciding on a tolerable level of disagreement among raters is to study the levels of inter-marker agreement typically achieved in other, similar settings. In all attempts at standard setting there is inevitably a normative element (i.e., a need to refer to the levels typically achieved in practice) to judge whether the standard being set is reasonable. Under this approach, the decision about an acceptable level of inter-marker agreement would be informed by considering the levels of inter-marker agreement typically achieved in other Year 12 assessment systems, including levels of agreement among markers of external examination papers.

In this study, we have not begun by specifying an *a priori* level of inter-marker agreement that we consider acceptable. Nevertheless, we are aware that levels of inter-marker reliability for many open-ended assessments tasks which use holistic scoring vary from about 0.6 (but sometimes less) to about 0.9. Rather, we have sought other recent evidence of levels of inter-marker reliability in Year 12 assessment in Australia which might provide a frame of reference for interpreting the results of this study.

A difficulty in using findings from other Australian states is that assessment procedures vary markedly from state to state. Most recent Australian studies of inter-rater reliability have considered levels of agreement among markers assessing performances on the same task. Typically, these tasks are essays or other open-ended questions on external examination papers or common tasks (e.g., research projects) undertaken by all students across a system. Even assessments of folios of student writing do not correspond to the assessment of folios of student work in Queensland. Nevertheless, other studies of inter-marker agreement provide some indication of what might be reasonable to expect of assessors in this study.

One set of recent discussions of acceptable levels of inter-rater reliability have focused on current procedures in the state of Victoria where teachers assess student work on each of a small number of common assessment tasks on a 10-point grade scale. In a study of inter-marker agreement in Victoria, Brown and Ball (1992) proposed an *a priori* level of acceptable disagreement of ±1 grade in less than five per cent of cases. Eyers (1993), in a review of Brown and Ball's report, reported levels of inter-rater agreement achieved in other Year 12 assessment systems, including an external English trial, external HSC over a number of years and a study in another (unnamed) Australian state. In these studies, discrepancies of more than ±1 grade on a 10-point scale occurred between 25 and 46 per cent of the time, levels well outside the *a priori* limit proposed by Brown and Ball. Eyers proposed that an acceptable level of agreement might be better specified in terms of assessments within ±2 grades on a 10-point scale rather than assessments within ±1 grade.

McGaw (1993) summarised the findings of the studies reported by Brown and Ball and Eyers. All studies reviewed reported the percent agreement to ±2 grades by independent markers. These are shown in Table 1.

Table 1. Percent of markers agreeing to ±2 grades

Source	% within +/- 2 grades
Brown & Ball (Victoria)	86
External test (Victoria)	81
HSC exams (Victoria 1989)	86
Unnamed State 1	90
Unnamed State 2	91
Unnamed State 3	79

Influence of school work programs

Schools in Queensland base their decisions on Levels of Achievement on their interpretation of the standards of work demonstrated in relation to the criteria specified in the school's work program. For this reason there is likely to be some variation among schools in their interpretation of standards associated with each Level of Achievement. These differences are likely to be small, however, because each school's work program must be broadly consistent with the Board syllabus.

When Levels of Achievement are reported on the Senior Certificate, however, they are unaccompanied by details of individual schools' work programs. There is thus an intention that grades reported on the Certificate will convey levels of school achievement which can be compared from student to student without reference to work programs. This raises the question of whether Levels of Achievement have a meaning independent of work programs. Can different assessors agree of the allocation of an 'HA' rather than a 'VHA' to a student folio without considering the details of the work programs in the schools from which those folios come?

To address this question, some assessors were asked to rate folios without the relevant school work program and were later given access to the work program and an opportunity to make a different assessment in the light of this additional information. Other assessors were given schools' work programs from the outset.

Influence of collaboration

Another question of interest, given the way the assessment system operates in Queensland, concerns the influence of collaborating with a colleague. It seems likely that the Level of Achievement allocated by a teacher to a folio after discussing that folio with a colleague will sometimes be different from the assessment the teacher would have made had they been working alone.

In this study, the extent to which teachers are influenced in their assessments by the opinions of colleagues was investigated by asking each assessor to first work independently in assessing a folio and then to discuss that folio with a colleague (who had also made an independent assessments of that folio). Following this discussion each assessor then provided a new assessment. The extent to which assessors' ratings are influenced by discussions with a colleague was investigated by comparing their assessments before and after collaboration.

What confidence do assessors have in their ratings?

In making subjective ratings of student folios, teachers are expected to assign the Level of Achievement most appropriate to each folio. For many folios, particularly those around the cut-off between two Levels of Achievement, it will not always be clear which Level is most appropriate.

In this study the confidence with which assessors rate student work was investigated by first asking for a rating of each folio on a 50-point scale (10 rungs within each of the five Levels of Achievement), and then asking the assessor to provide an 'upper bound' and a 'lower bound' on this rating indicating the range within which they felt confident the folio belonged. The difference between the upper and lower bounds specified by an assessor provides some indication of the confidence with which their initial rating was made.

Are levels of inter-rater reliability different for different school subjects?

Levels of inter-marker reliability can be studied for Year 12 assessment procedures as a whole and for individual subjects separately. It is conceivable that greater inter-marker reliability will be obtained in some subjects than in others.

In this study, levels of inter-marker reliability are investigated in each of four Board subjects separately: Chemistry, English, Mathematics I and Modern History.

3.2 Consistency with exit Levels of Achievement

Each of the sampled folios in this study was assessed independently six times, allowing an investigation of levels of agreement among assessors as well as an investigation of possible influences on assessors' ratings. But there is also the question of the extent to which these six independent assessments of a folio were in agreement with that folio's exit Level of Achievement at the end of 1992.

The six independent assessments also could be compared with the October assessments of each folio (i.e., school assessment and panel recommendation), but because the October assessments were not based on the complete (exit) folio, these comparisons are likely to be less useful than those based on assessments of the folio at exit.

3.3 Differences across schools

When assessments made in this study differ from exit Levels of Achievement certified by the Board, the question raised is: What is the source of these differences? Are differences distributed randomly across the sample of folios reviewed, or are they due to differences in standards across schools? If there are differences in standards across schools, what is the magnitude of these differences?

4. Design of the Study

4.1 Sample of folios

A sample of student assessment folios was drawn for the 1992 cohort of Year 12 students. The sampled folios were collected at the completion of Year 12 (i.e., on exit) in four Board subjects:

Chemistry
English
Mathematics I
Modern History

Rather than drawing a random sample of folios, it was decided to select exit folios that had earlier been submitted as part of the Board's October review process. In this way, the selected folios represented a range of Levels of Achievement in each school. The procedures specified by the Board for selecting folios for the October review are described in Section 2 above and are the same for all schools, meaning that the folios from each school should represent the same mix of achievement levels, at least for large schools.

Typically, nine folios were drawn from approximately 15 schools for each subject. Schools were selected at random on the condition that at least four students in the school studied that subject. No school was selected for more than one subject. This process resulted in the collection of about 125 to 150 folios in each subject to provide a total of 546 assessment folios for re-assessment. Folios were drawn from 62 schools.

Instructions to schools requested that all assessments which contributed to the exit Level of Achievement be included in the folios submitted for re-assessment. Although the proposals submitted by the school and the recommendations of district review panels were based on an incomplete assessment program (at October), the exit Levels of Achievement and the assessments made in this study were based on folios containing assessments from Term 4 as well.

Folios usually were accompanied by schools' marking schemes and the weightings applied to individual assessments. Many schools provided aggregate weighted scores. All schools provided scripts with marks assigned by teachers and, in some cases, written teacher feedback to students. Thus, the role of the assessors in this project was not to mark the scripts but to estimate the Level of Achievement demonstrated by the assessments in each folio.

4.2 Appointing assessors

It was considered important to this study that independent judgements of Levels of Achievement were obtained from assessors familiar with the relevant Board syllabus and with demonstrated expertise in assessing student folios against syllabus criteria and school work programs as part of the October review process. A statement of the requirements for the appointment of assessors is provided in the Appendix. Assessors were identified by asking the chairs of the Board's eleven District Review Panels in each subject to nominate suitable people for this work. In this way, the study was able to ensure that all assessors had experience in and knowledge of the relevant syllabus, including the criteria for judging Levels of Achievement.

Assessors were invited to specify days and times at which they would be available for a six-hour session. To minimise disruption to schools, all sessions were conducted out of school time (evenings, Saturdays, and during school holidays) even though assessors were initially invited to nominate for possible sessions both in and out of school hours.

Although the selection process resulted in approximately 120 assessors being identified as suitable for the project, only 62 were actually used. Because some assessors nominated more sessions than others, the number of folios assessed varied from assessor to assessor. Some assessed as few as a dozen or so folios; others assessed more than one hundred. Most assessed around forty. Each person re-assessed folios in one subject only.

4.3 Designing assessment conditions

To investigate factors that might influence assessors' judgements of student folios (having access to the school's work program; seeing folios in school groups; and discussing student work with a colleague), the research design provided for the re-assessment of each folio under a number of different conditions. Each folio was independently assessed on six occasions, usually by six different assessors.

Two independent assessments were made under each of three conditions:

- without the school's work program, but in school groups (*Model 1*);
- with the school's work program and in school groups (*Model 2*);
- without access to work programs and not in school groups (*Model 3*).

Under Model 1, each of the two assessors first made an assessment without the school work program and was then given an opportunity to make a second assessment on the basis of the provided work program. In practice, so few assessors changed their initial assessment after seeing the work program, that this step was discontinued during the study, and hence a detailed analysis of these changes was not pursued. The Model 1 assessments analysed in this study are initial assessments made without the work program.

What assessors were asked to do

At the beginning of each session, assessors were given:

- instructions
- a master sheet
- assessors' comment sheet
- diagrammatic representation of rungs within levels of achievement
- student folios

In addition to the printed material, assessors were briefed on the tasks to be completed. Differences between the requirements of this study and those of the October review process were highlighted.

For the October review process, panel members are provided with as much information as possible about the assessment practices in the school and the processes used to arrive at estimated Levels of Achievement. This information includes the identity of the school, the school's work program, and the completed Form R6 showing the school's estimated Levels of Achievement for the folios to be reviewed and the distribution proposed for all Year 12 students studying the subject in the school. Most panellists have been involved in the moderation process for a number of years and are familiar with past problems in the accreditation of school work programs in the district and with the standards of material submitted by schools for review in previous years.

These procedures ensure that panel members and staff in schools have shared understandings of acceptable types of assessment programs and the interpretation of the syllabus, particularly the standards associated with criteria for awarding Levels of Achievement. The purpose of the October review panel meeting is to confirm that the school estimates are within acceptable limits rather than to make fresh, independent estimates of Levels of Achievement for the materials reviewed.

In contrast, the purpose of this study was to investigate the consistency of assessors' allocated Levels of Achievement under a range of conditions and without access to much of this information. The Levels of Achievement proposed by the school were not available

and, in two of the three research models, the school's work program was not available. Thus, assessors had to rely on their familiarity with the Board syllabus and on their previous experiences with a range of school work programs.

Influence of collaboration

To investigate the possible influence of collaboration on assessors' ratings, under each model, the two assessors first worked alone to provide a rating for each folio and were then given an opportunity to discuss that folio. Each assessor then had an opportunity to provide a *second* rating based on this discussion. This feature of the design provided an opportunity to compare each assessor's ratings before and after collaboration.

Assessor confidence

Another question of interest relates to the confidence with which assessors made their ratings of each folio. In an attempt to collect information on assessors' confidence in their assessments, each time a rating was made, assessors were asked not only to provide an assessment but also to specify an *upper and lower bound* to define a range within which they were confident the folio was located.

4.4 Recording assessments

All assessors recorded their ratings on a 50-point scale obtained by defining 10 'rungs' within each of the five Levels of Achievement (see Figure 1). All assessments were recorded as numbers between 10 and 59. In dividing each Level of Achievement into ten equal rungs, there is a clear intention that these rungs represent equal intervals within Levels. The assumption is made in this report that the 50-point scale approximates an interval scale[1] .

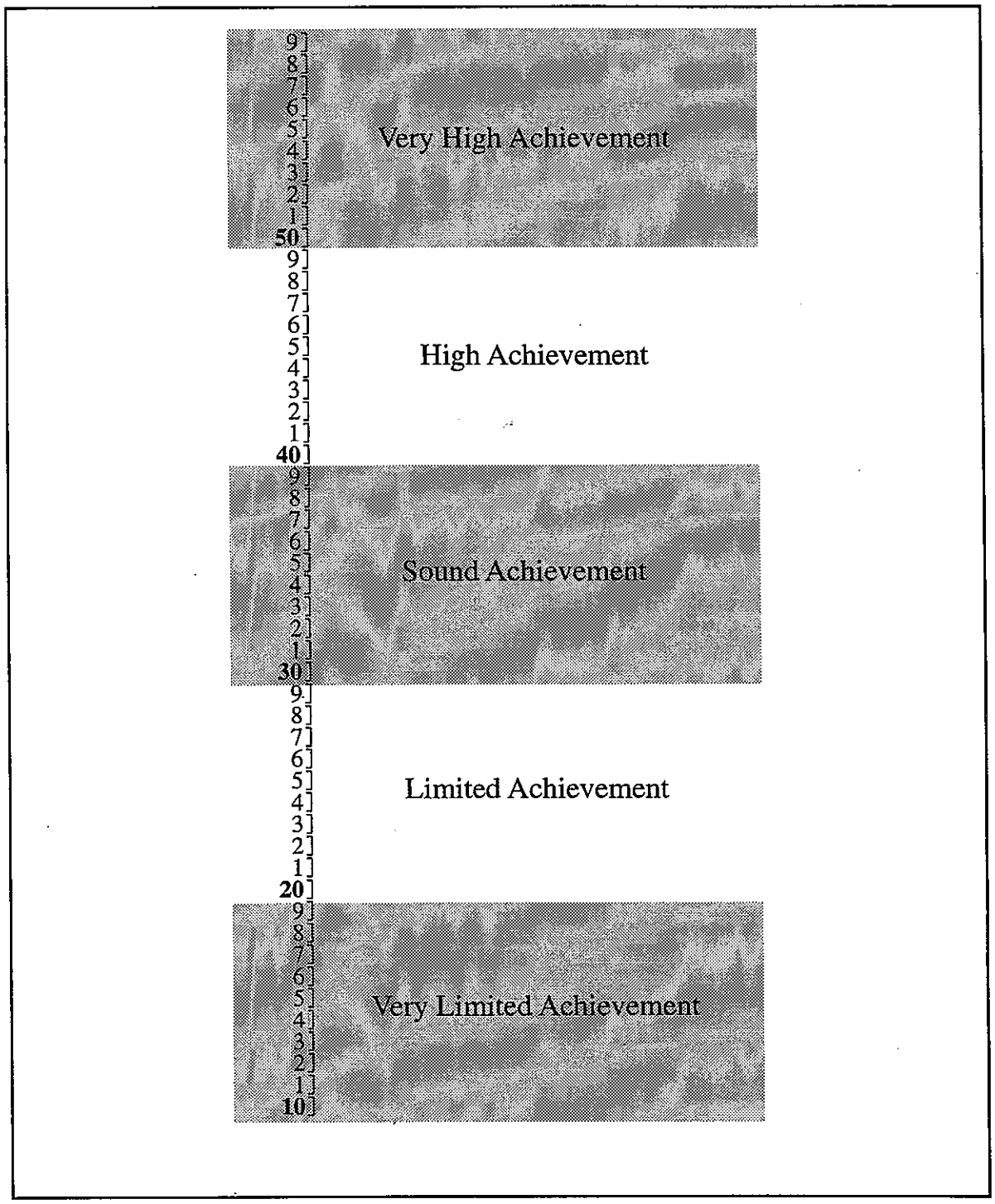
4.5 Comparing results with exit levels of achievement

Because each sampled folio had been through the October review process, a number of separate assessments were available for each folio. These included:

- A school assessment on a 50-point scale (Level of Achievement + rung) taken from the Board's Form R6 (October 1992);
- A panel recommendation, also on a 50-point scale (Level of Achievement + rung) taken from the Board's Form R6 (October 1992).
- An exit result (Level of Achievement on a 5-point scale) based on the folio at the completion of the year.

[1] This is likely to be a reasonable assumption and was supported by an Item Response analysis through which the 50 rungs were calibrated on a continuum based on assessors' ratings.

Figure 1. 50-point achievement scale (10 rungs within each of five levels)



Because assessments made in October were not based on the full student assessment folios, the most appropriate comparisons are between the six independent assessments made of each folio in this study and the exit Level of Achievement at the end of 1992. These comparisons allow the level of agreement between assessors' judgements and teachers' exit Levels of Achievement to be investigated.

5. Findings

5.1 Results of re-assessing folios

5.11 Inter-marker agreement

Each of the 546 folios was assessed independently by two assessors under each of the three models described above. These models are now considered one at a time, and the two independent ratings of each folio are considered.

Under Model 1, assessments were made in school groups and without access to the relevant work program. In Figure 2 the two assessments of each folio under Model 1 are plotted against each other. One of these assessments is referred to as the 'first' assessment, the other as the 'second' assessment, reflecting only the order in which these two ratings were entered in the project data file.

The diagonal line through the middle of Figure 2 is the line of perfect agreement between the two independent assessments. Other lines have been drawn to indicate differences of plus and minus half a Level of Achievement (5 rungs) and plus and minus a full Level of Achievement (10 rungs). In Figure 2, 98 per cent of the two independent assessments of each folio are within one Level of Achievement; 89 per cent are within half a Level of Achievement. The product moment correlation (i.e., the usual index of inter-marker reliability) is .94.

Under Model 2, assessments were made in school groups with the relevant work program. The two independent assessments of each folio under Model 2 are plotted against each other in Figure 3. In this case, 98 per cent of the two assessments are within one Level of Achievement, and 90 per cent are within half a Level of Achievement. The correlation between the two independent assessments is .94.

Under Model 3, assessments were made by presenting folios randomly (i.e., not in school groups) and without access to schools' work programs. The two independent assessments of each folio under Model 3 are plotted against each other in Figure 4. Under this model, ratings are even more consistent with each other, with some 99 per cent of assessments being within one Level of Achievement, and 91 per cent being within half a Level of Achievement. The correlation between the two assessments is again .94.

The inter-rater reliabilities achieved under these three models are summarised in Table 2. The inter-rater reliability has also been calculated by pooling assessments under all three models.

Table 2. Inter-marker agreement by model

Model	Description	N	r	% within +/- 5 rungs	% within +/- 10 rungs
1	in schl gps; no work prgm	546	.94	89	98
2	in schl gps; work prgm	546	.94	90	98
3	no schl gps; no work prgm	546	.94	91	99
all		1638	.94	90	98

Figure 2. First and second ratings of folios plotted against each other (Model 1, all subjects pooled)

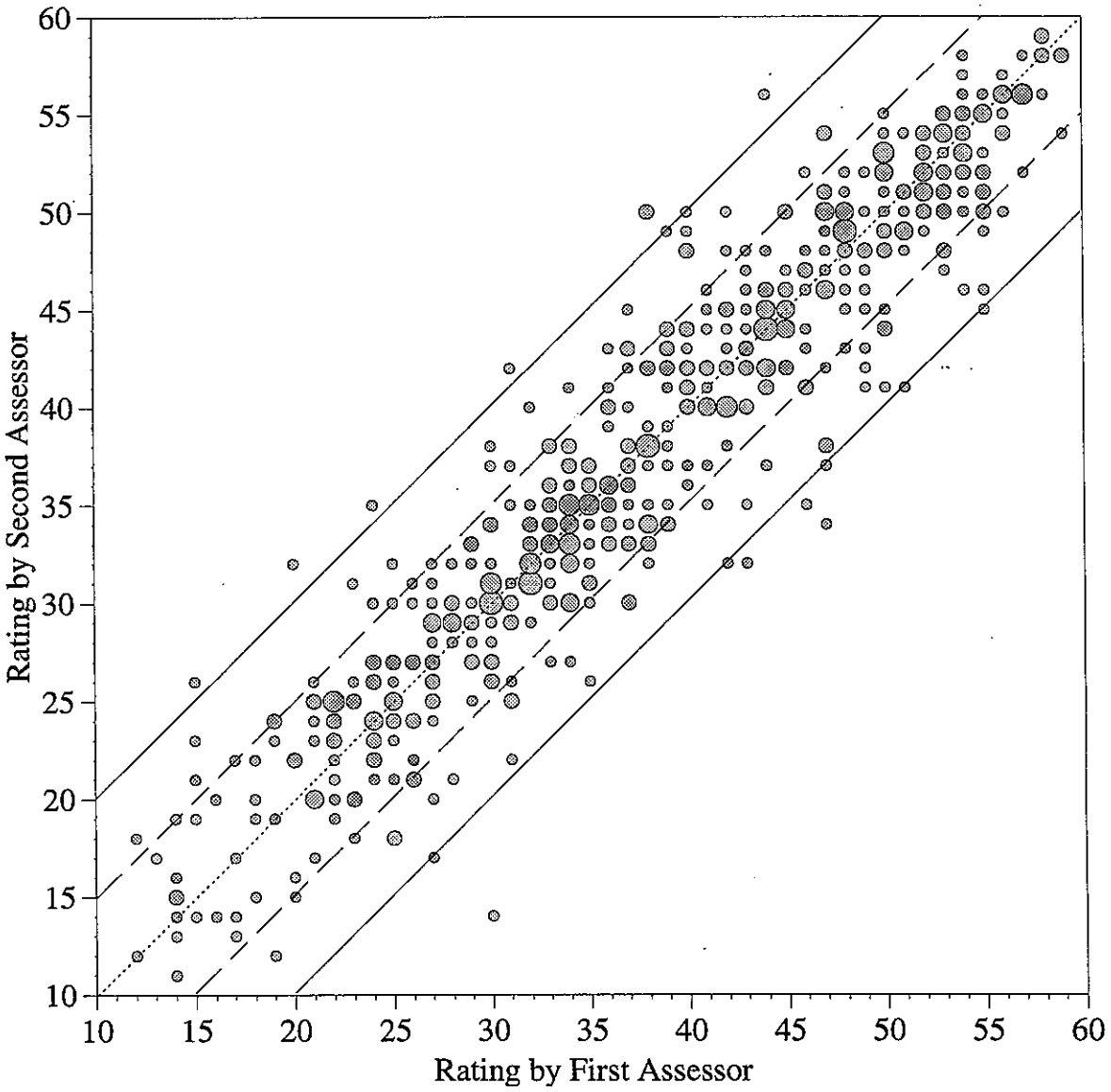


Figure 3. First and second ratings of folios plotted against each other (Model 2, all subjects pooled)

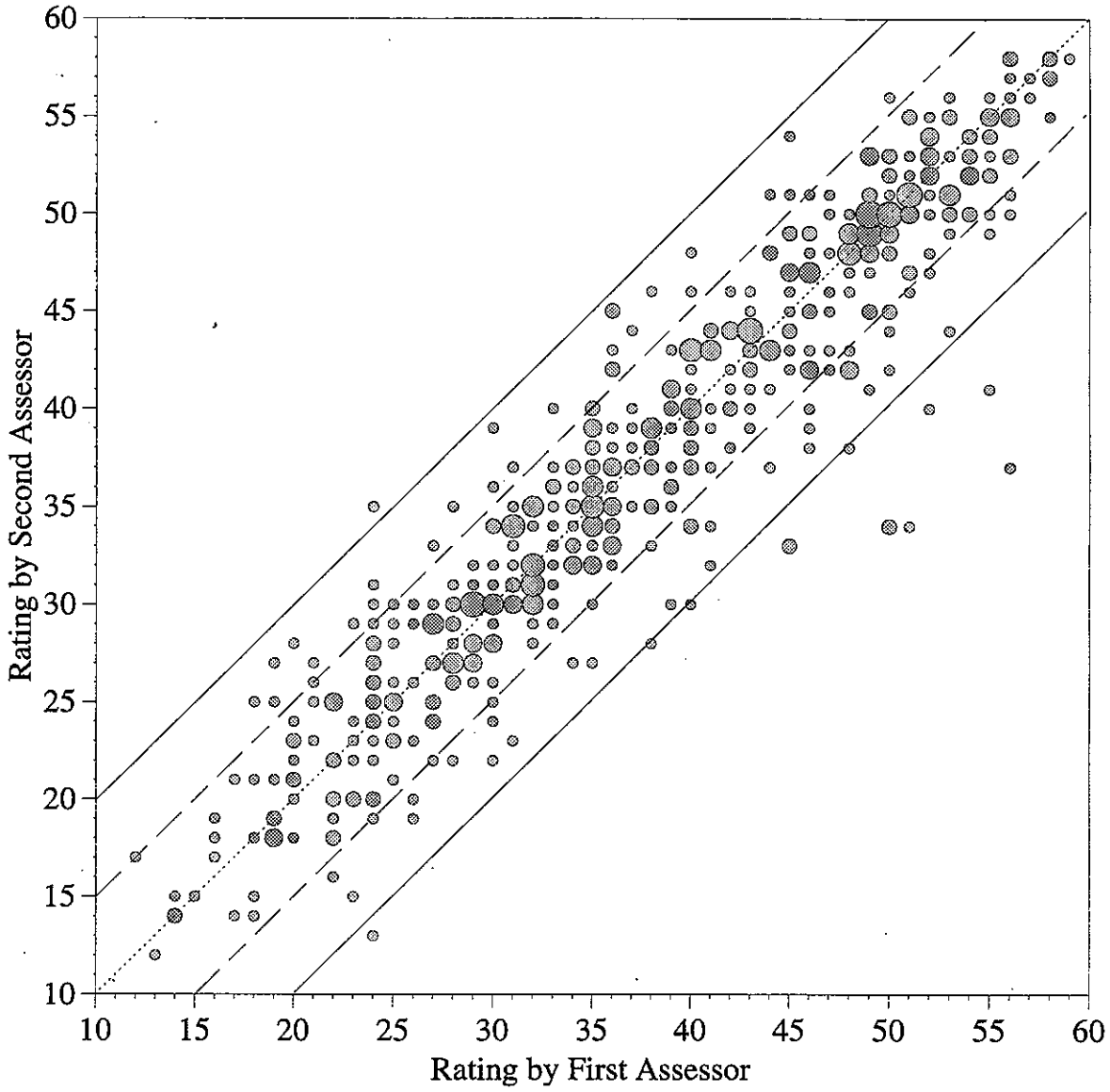
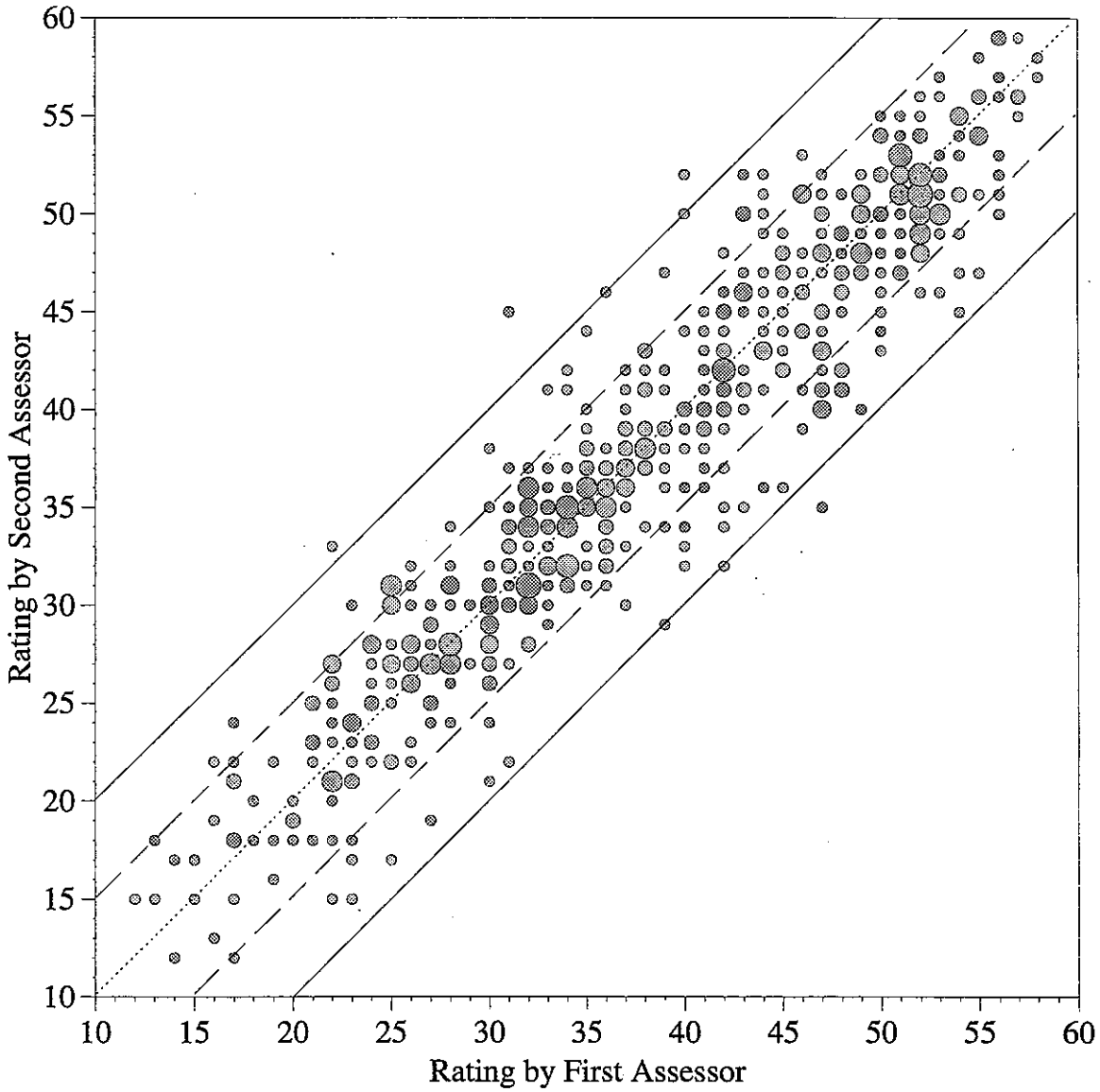


Figure 4. First and second ratings of folios plotted against each other (Model 3, all subjects pooled)



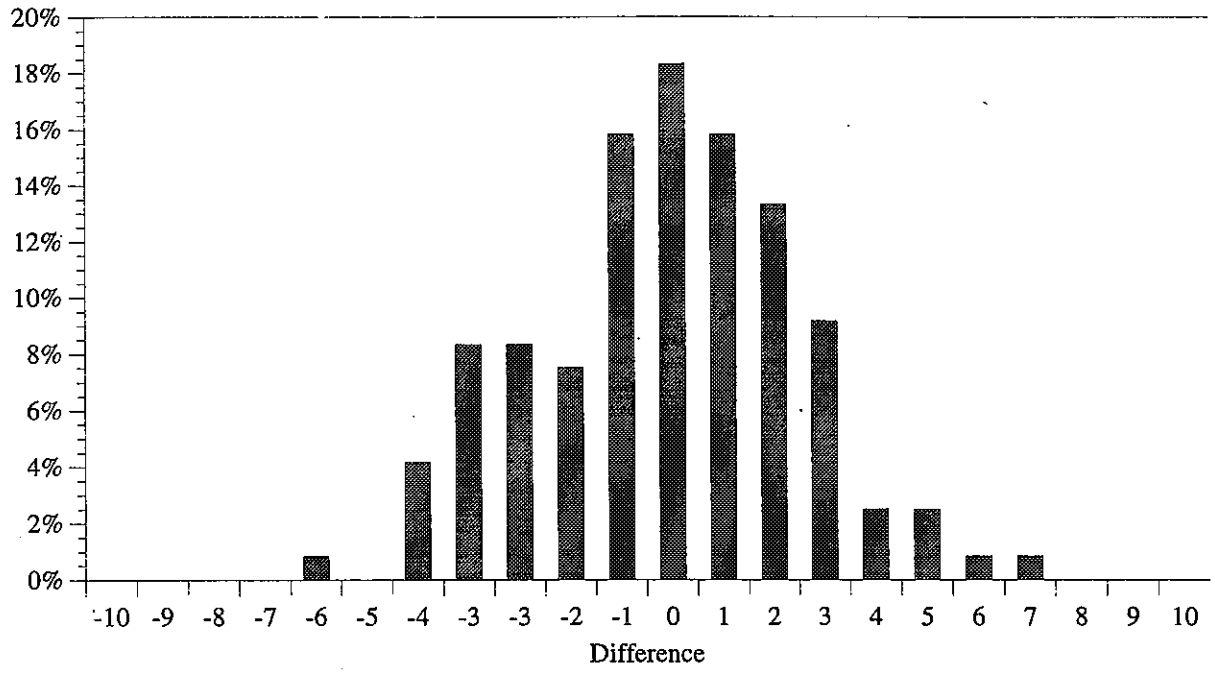
It is instructive to compare the inter-rater reliabilities in Table 2 with inter-rater reliabilities reported in other studies, particularly studies of Year 12 assessment arrangements in some other Australian states. While no other system uses the procedures applied in Queensland, some recent work in Victoria provides a point of reference for interpreting the results in Table 2.

Brown & Ball (1992)

Brown and Ball investigated levels of inter-rater reliability in relation to assessments for the Victorian Certificate of Education. They drew samples of 120 students in English and 120 students in mathematics (reasoning & data) by asking four schools to provide the work of 30 students for re-marking in English and another four schools to provide the work of 30 students in mathematics. Re-marking was carried out by three chairpersons of the relevant subject panels. Each piece of student work was re-marked by only one chairperson.

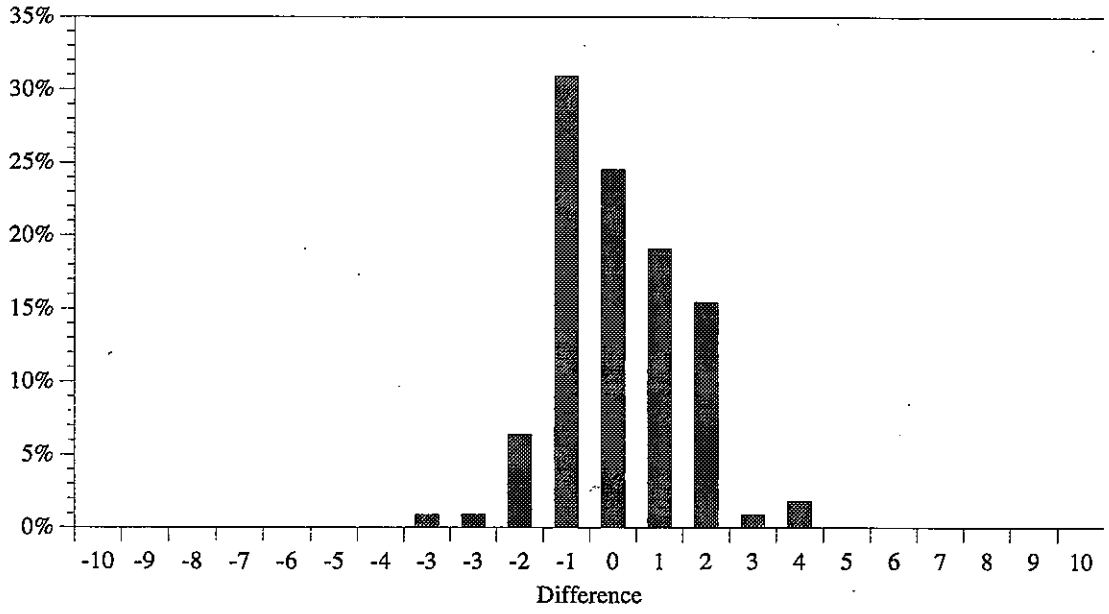
Brown and Ball then compared the original assessment of each student's work (expressed on a 10-point scale from E to A+) with the re-assessment on the same scale. They did not calculate an inter-rater reliability coefficient, but did provide a histogram of differences between original assessments and re-assessments for each of English and mathematics (Figures 5 and 6).

Figure 5. Differences between original assessments and re-assessments, VCE English task



Brown and Ball concluded that the discrepancies in Figures 5 and 6 reflected an unacceptable level of discrepancy between initial assessments and re-assessments. An acceptable level of difference between the initial assessment and re-assessment, they argued, would be a difference of no more than one grade (on this 10-point scale) in less than 5 per cent of cases. This level of inter-rater agreement was clearly not achieved for either the English or mathematics assessment task in the Brown and Ball experiment.

Figure 6. Differences between original assessments and re-assessments, VCE math task



Hill (1993) and Eyers (1993)

In a commentary on the Ball and Brown study, Hill (1993) notes that the criterion for acceptability introduced by Ball and Brown implies an inter-rater reliability of about .88.

To provide a frame of reference for studying the findings of Brown and Ball, Hill (1993) reported the results of a study in which a random selection of 22 student scripts on an externally-marked VCE English test were re-marked blind by five volunteer assessors. The 22 scripts were drawn to provide two scripts from each grade level (ungraded, E,..., A+). The 22 scripts were re-marked by each of the five assessors who were unaware that the full range of grades had been sampled or that there had been two scripts drawn from each grade.

Differences between the grades originally assigned to scripts and the grades resulting from the re-marking of each script are plotted as a histogram in Figure 7. Inter-marker reliabilities are shown in Table 3.

The results in Table 3 show a level of inter-rater reliability in marking the external English test similar to the level obtained in this study for re-assessed folios. However, as can be seen from the percentages of discrepancies of two or more grades, the Brown and Ball criterion of a difference of one grade in less than five per cent of cases was not met even in the double marking of this external test.

Eyers (1993), in his review of the Brown and Ball study, compared their findings with results from some other Australian Year 12 assessment systems and concluded that 'on a 10-interval scale, differences of *two* (not one) are generally regarded as acceptable for inclusion'. Eyers reported results from other states of 90%, 91% and 79% of re-markings within ±2 grades and concluded that the discrepancy rate in the Brown and Ball study was 'well within the range of those found and accepted in other situations where a blind double marking is legitimately applied' and met 'normal professional tolerances including those achieved with external examinations' (Eyers, 1993, 13).

Figure 7. Differences between original assessments and re-assessments, VCE English test

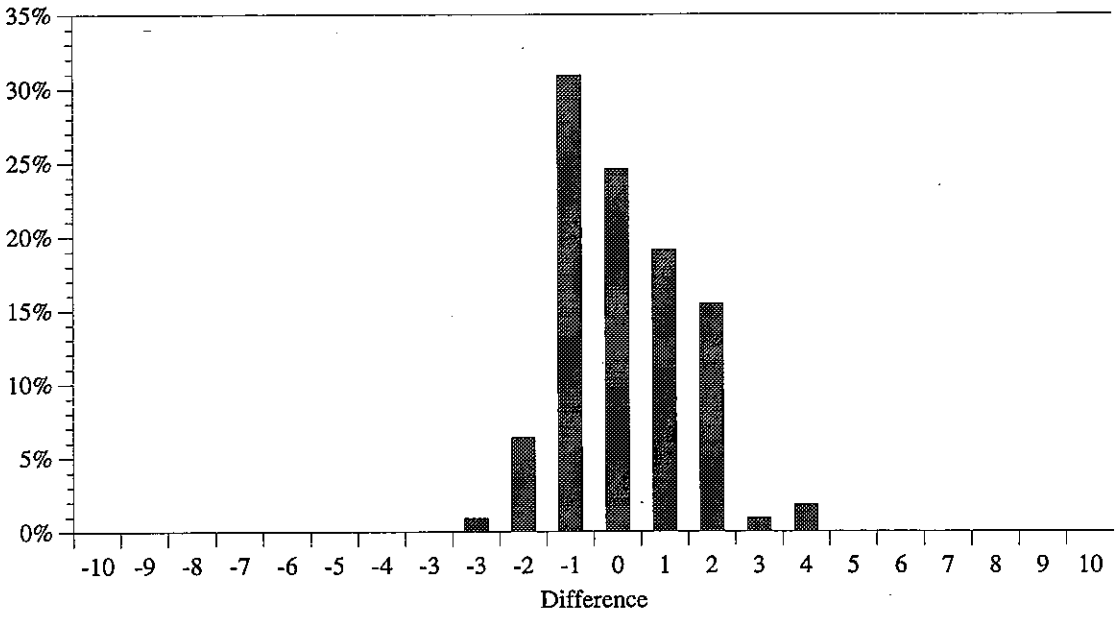


Table 3. Inter-rater reliability in marking the VCE English test

Marker	r	% discrepancies of 2 or more grades
1	.94	23
2	.94	14
3	.93	23
4	.94	27
5	.89	41
all	.93	25

If a difference of ±2 grades on a 10-point grade scale is considered equivalent to one Level of Achievement on the 5-point scale used in Queensland, then it is clear that the inter-rater reliabilities achieved in this study (Table 2) are significantly greater than those reported by Brown and Ball (1992), Hill (1993) and Eyers (1993).

While other studies of inter-rater reliability provide some basis for interpreting the levels of inter-marker agreement achieved in the present study, there are also important differences in the assessment procedures on which these studies are based. In other studies of inter-rater reliability, double blind marking is used with no assessor having access to the marks or comments of any other assessor. Students attempt the same general assessment task and markers use an agreed set of criteria to assess performances on that task.

In the present study, student folios were not always based on the same assessment tasks. Under Models 1 and 2, assessors saw folios in school sets, and within each school set students completed a common assessment program and set of tasks. But this was not the case under Model 3.

Assessors in this study had a significant advantage over markers in other studies in that student folios included the marks and comments of the teachers who initially assessed them. Some folios included aggregated marks. Where aggregate scores were available, it is possible that some assessors based their ratings of folios on numerical scores, perhaps associating Levels of Achievement with particular score ranges.

Because of the rather different task undertaken by assessors in this study it is difficult to know what a reasonable level of inter-rater reliability might be, or to decide whether the criterion of 95 per cent of markers being within a range ±2 grades on a 10-point scale is a reasonable expectation.

5.12 Influence of work programs and school groupings

A conclusion that can be draw with a little more confidence concerns the possible influence of providing assessors with access to school work programs and providing folios in school groups. It is sometimes argued that the allocation and interpretation of Levels of Achievement requires access to the relevant school work program.

In this study there is an opportunity to study the influence of work programs through a comparison of Models 1 and 2. Under these models, assessors saw folios in school groups, but only under Model 2 did they also have access to the work program. If the work program makes it easier for assessors to allocate and interpret Levels of Achievement, then it seems reasonable to expect that inter-rater agreement will be higher under Model 2 than under Model 1.

In fact, inter-rater reliability is not higher under Model 2 than Model 1. This suggests that the presence or absence of the work program has little if any influence on assessors' abilities to agree on appropriate ratings of student work.

In the case of school groupings, it might be anticipated that higher levels of inter-marker agreement would be obtained when folios are presented in school groups that when they are presented randomly from a range of schools. The reason for this is that, within school groups, students' folios contain performances on some or all of the same school assessment tasks. Where a folio consists of a set of marked school tests, it might be expected that an assessor would be able to reproduce the rank order of students within a school almost perfectly on the basis of students' test scores.

The hypothesis that presenting folios in school groups should make it easier for assessors to compare students and thus lead to higher levels of agreement among markers can be tested in this study by comparing Models 1 (in school groups) and 3 (not in school groups).

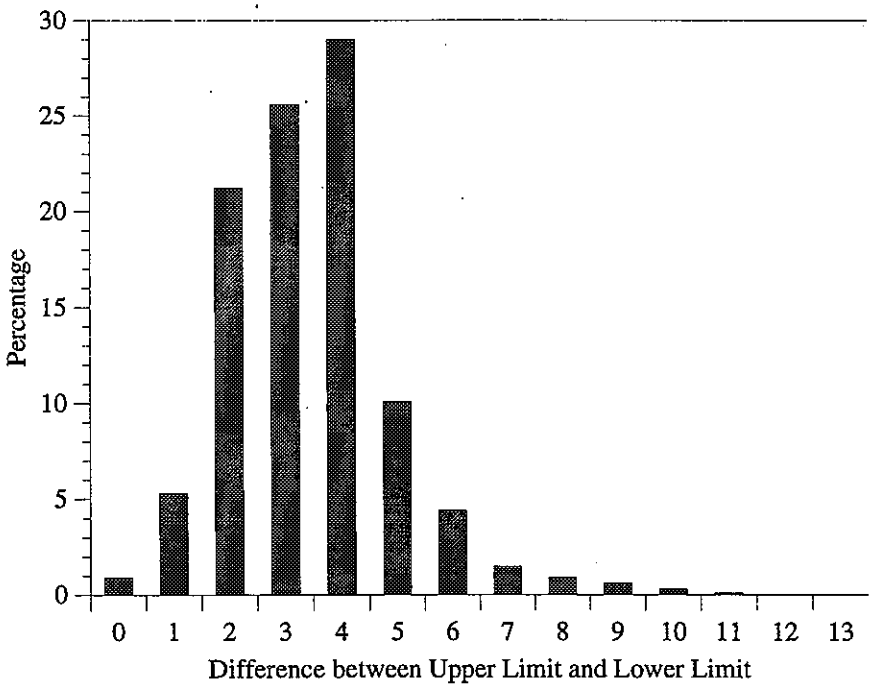
The evidence of this study is that assessors have been able to reach agreement on student folios at least as well when folios are presented randomly as when they are presented in school groups. A comparison of Figures 2, 3 and 4 and of the data in Table 2 suggests that having access to work programs and having folios presented in school groups has not significantly improved levels of inter-rater reliability. At least in these data, there is little evidence to support arguments for the necessity of seeing folios in school groups and in conjunction with school work programs.

5.13 Levels of assessor confidence

As noted above, information was collected about the confidence with which assessors made their ratings of student folios by asking them to provide not only a 'best estimate' for each folio but also an upper bound and lower bound on this rating. The difference between the upper and lower bounds provides some indication of the confidence with which assessors made their ratings.

These differences have been calculated for all six independent ratings made of each folio in all four school subjects. The distribution of differences is shown in Figure 8.

Figure 8. Measures of assessor confidence



From Figure 8 it can be seen that the vast majority of assessments were made with a high degree of confidence. The difference between the upper and lower limits was less than four score points in more than 50 per cent of cases and less than five score points in more than 80 per cent of cases. In other words, on a 50-point scale, the vast majority of ratings were made with a level of confidence corresponding to ±2 score points.

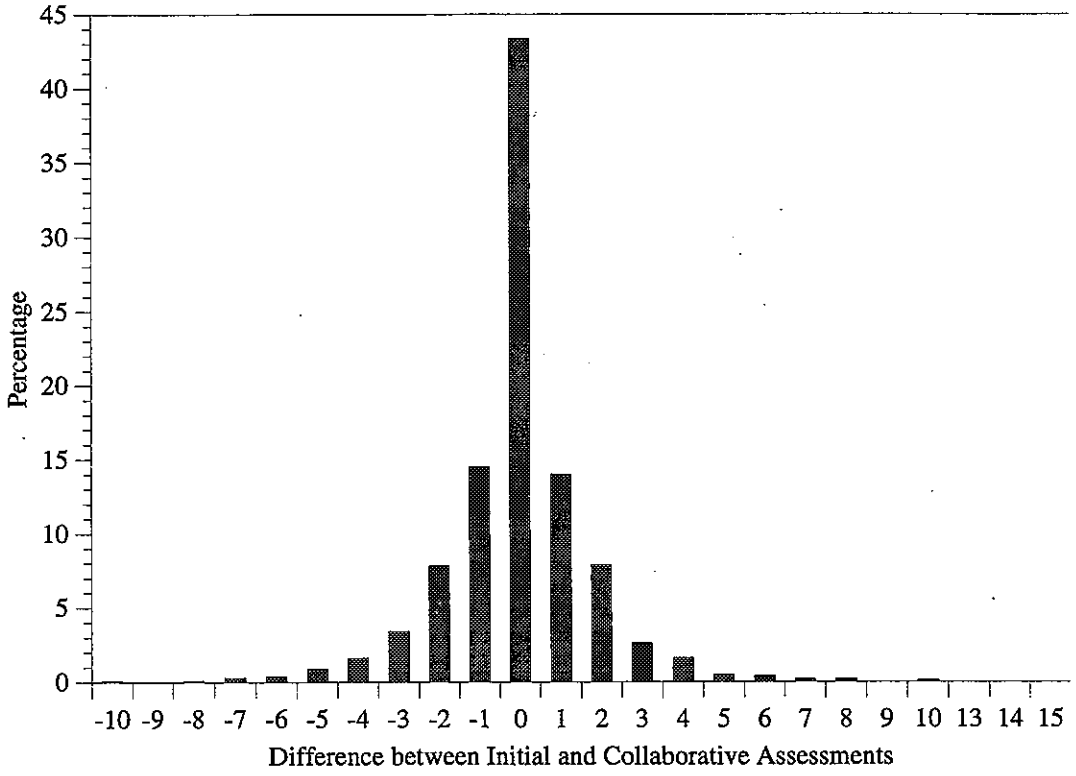
The high degree of confidence some assessors expressed in their ratings is surprising. In about one per cent of cases, assessors claimed perfect confidence in their rating (i.e., *no* uncertainty about an allocated rating on a 50-point scale). This may suggest that some assessors did not understand the task. On the other hand, the fact than more than 25 per cent of ratings were assigned a confidence of ±1 score point on a 50-point scale suggests that assessors have a high degree of confidence in their ability to make such assessments reliably.

5.14 Influence of collaboration

Under each of the three models in this study, two assessors worked independently to assess a number of folios in a subject. The two assessors were then asked to discuss their independent assessments of each folio and, following this discussion, to each provide a second assessment of the folio. In this way, it was hoped to obtain some feel for the extent to which assessors are influenced by collaboration with a colleague in their assessment of student work.

The difference between each assessor's second assessment (after discussion with a colleague) and first (independent) assessment has been calculated and the distribution of differences (across all three models and all four subjects) is shown in Figure 9.

Figure 9. Changes after collaboration

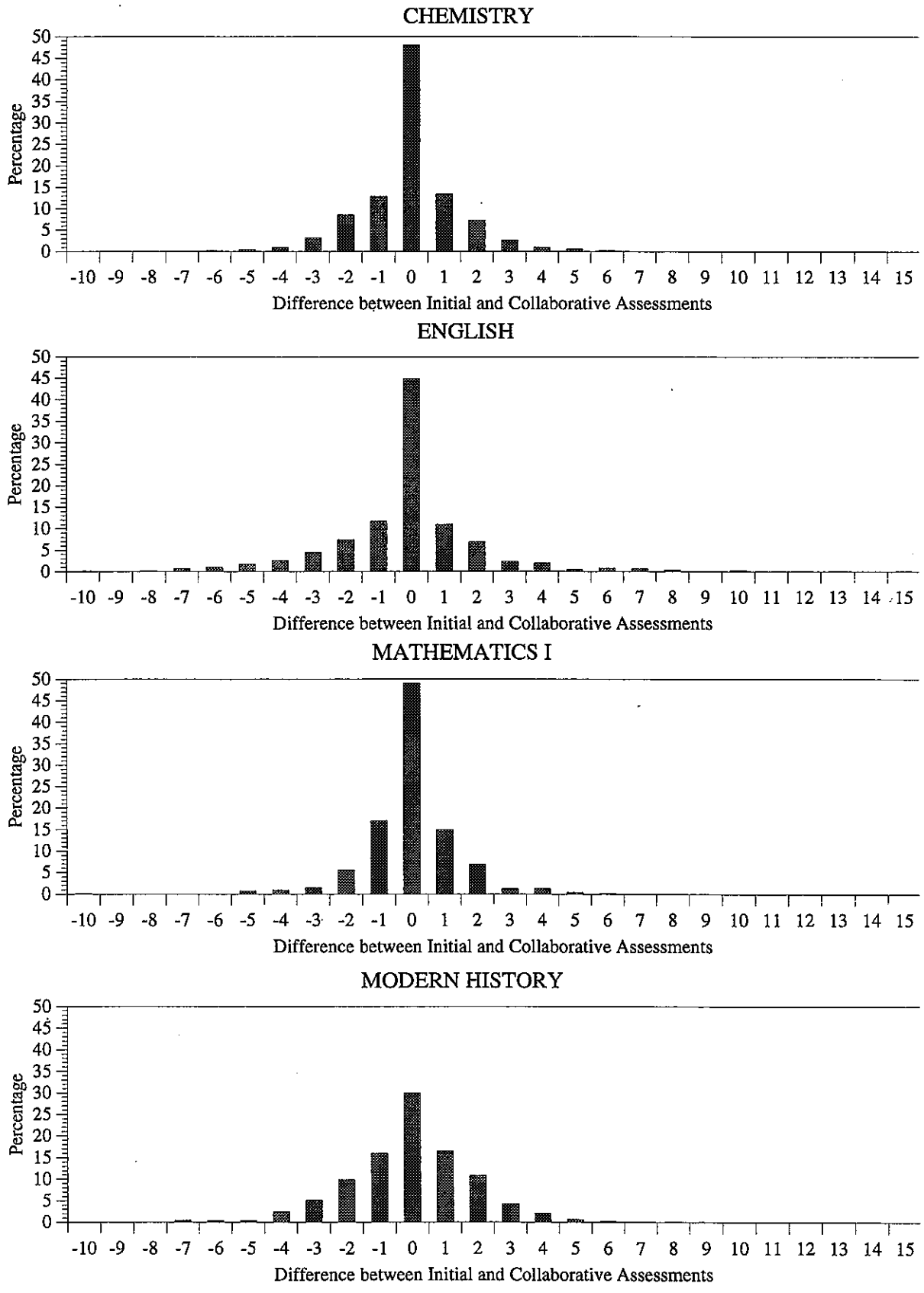


It can be seen from Figure 9 that 44 per cent of second assessments were identical to the first assessment made (i.e., difference = 0). The vast majority of revised assessments were within ±2 score points of the initial assessment. A very small number of larger revisions were made, a handful representing 10 score points or a full Level of Achievement. As might be expected, the distribution is almost symmetrical, reflecting the fact that as many assessors revised their initial assessments upwards as downwards.

The general conclusion from Figure 9 is that assessors either felt no need to revise their initial assessment after discussing each folio with a second, independent assessor or made a small revision of the order of one or two score points on the 50-point scale.

Figure 10 shows the difference between assessors' initial ratings and their assessment after collaboration for each of the four subjects separately. In each of these subjects, assessors made only small changes to their initial ratings after collaboration. Almost 50 per cent of assessments were unchanged in Chemistry, English and Mathematics. Assessors of folios in History more often changed their folio ratings than assessors in the other three subjects, but in all subjects the changes tended to be of the order of a rung or two only.

Figure 10. Changes after collaboration (all subjects)



5.15 Differences among school subjects

In the preceding analyses, levels of inter-rater reliability have been investigated for each model separately by combining assessments made for all four school subjects. As noted above, these analyses suggest that similar levels of inter-marker reliability are achieved with and without access to the school work program and with and without folios being presented in school groups. With this established, it is now interesting to explore possible differences among the four school subjects. Are higher levels of inter-rater reliability obtained for some subjects that for others?

To investigate levels of inter-rater reliability for folios in each subject separately, the first and second assessment of each folio have been plotted against each other in Figures 11, 12, 13 and 14 for the four Board subjects in this study. As above, a product moment correlation has been calculated, and the percentage of differences greater than ±5 rungs and greater than ±10 rungs have been calculated. These are summarised in Table 4.

Table 4. Inter-marker agreement by subject

Subject	N	r	% within +/- 5 rungs	% within +/- 10 rungs
Chemistry	417	.96	95	100
English	372	.89	82	95
Mathematics I	468	.97	94	99
Modern History	381	.92	88	99
all	1638	.94	90	98

From Figures 11, 12, 13 and 14 and Table 4 it is seen that there *are* differences in inter-rater reliability from subject to subject. The highest levels of agreement between assessors occur in Mathematics and Chemistry; lower levels of agreement occur in History and English. This finding is broadly consistent with observations made in other studies, and may be explained in terms of the greater use of analytic rather than holistic scoring of assessment tasks in Mathematics and Chemistry.

5.2 Comparisons with exit levels of achievement

So far, analyses have focused on ratings made by the team of assessors used in this study, with particular emphasis on the level of inter-rater reliability. The central question asked has been: Can a group of assessors familiar with and experienced in the assessment of student folios provide consistent, and hence comparable, ratings of student work on the 50-point scale used in the Board's assessment and panel review process?

A second question that can be asked is: How consistent are ratings made by this group of assessors during 1993 with the original Levels of Achievement awarded to these folios at the end of 1992? Because the exit Levels of Achievement were expressed on the 5-point scale (VLA, LA, SA, HA, VHA), the comparison in this case is between the more detailed ratings requested of assessors in this study (5 Levels; 10 rungs within each Level) and the coarser 5-point scale on which exit achievements are reported.

Figure 11. First and second ratings of folios plotted against each other (Chemistry)

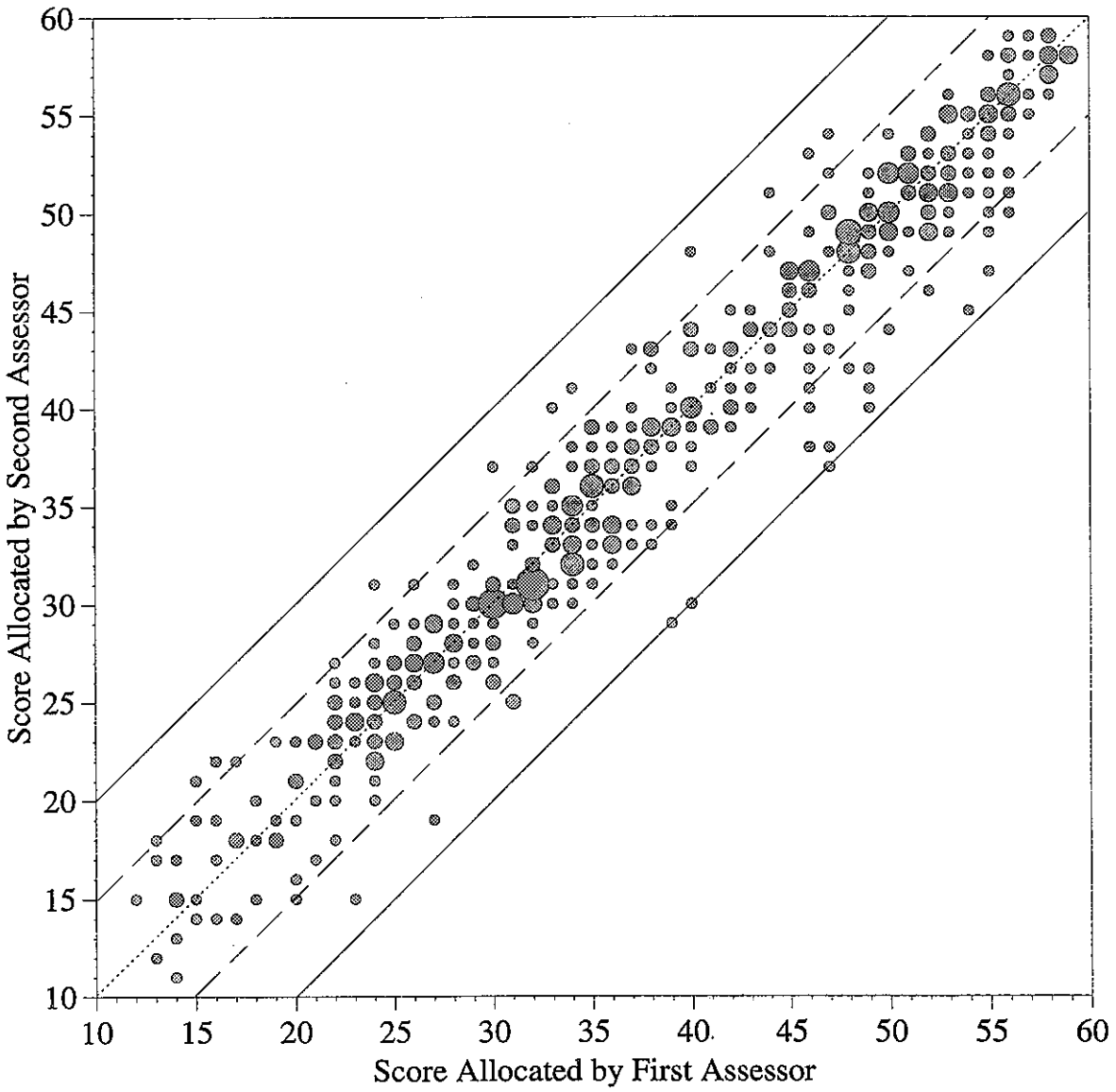


Figure 12. First and second ratings of folios plotted against each other (English)

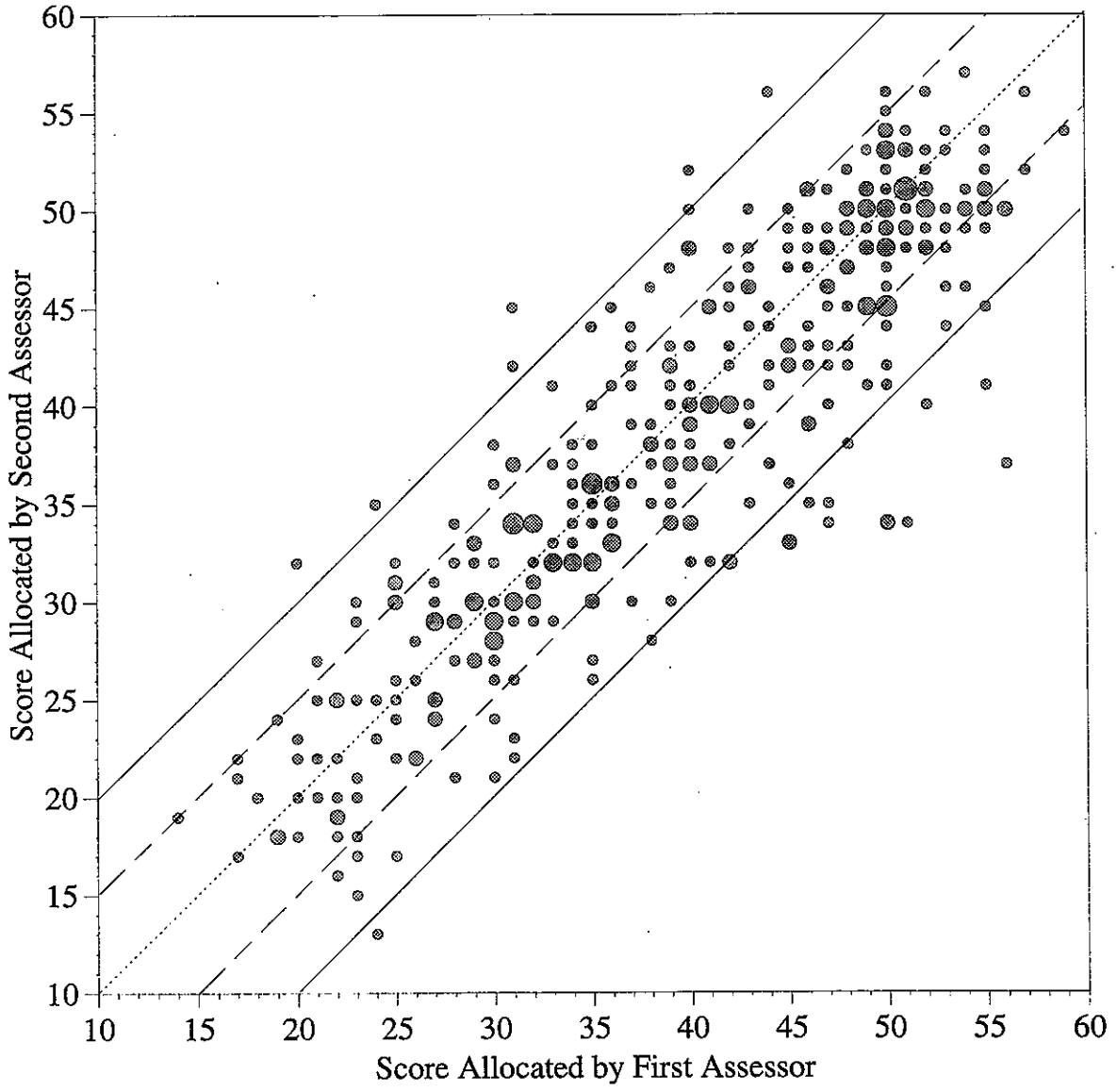


Figure 13. First and second ratings of folios plotted against each other (Mathematics)

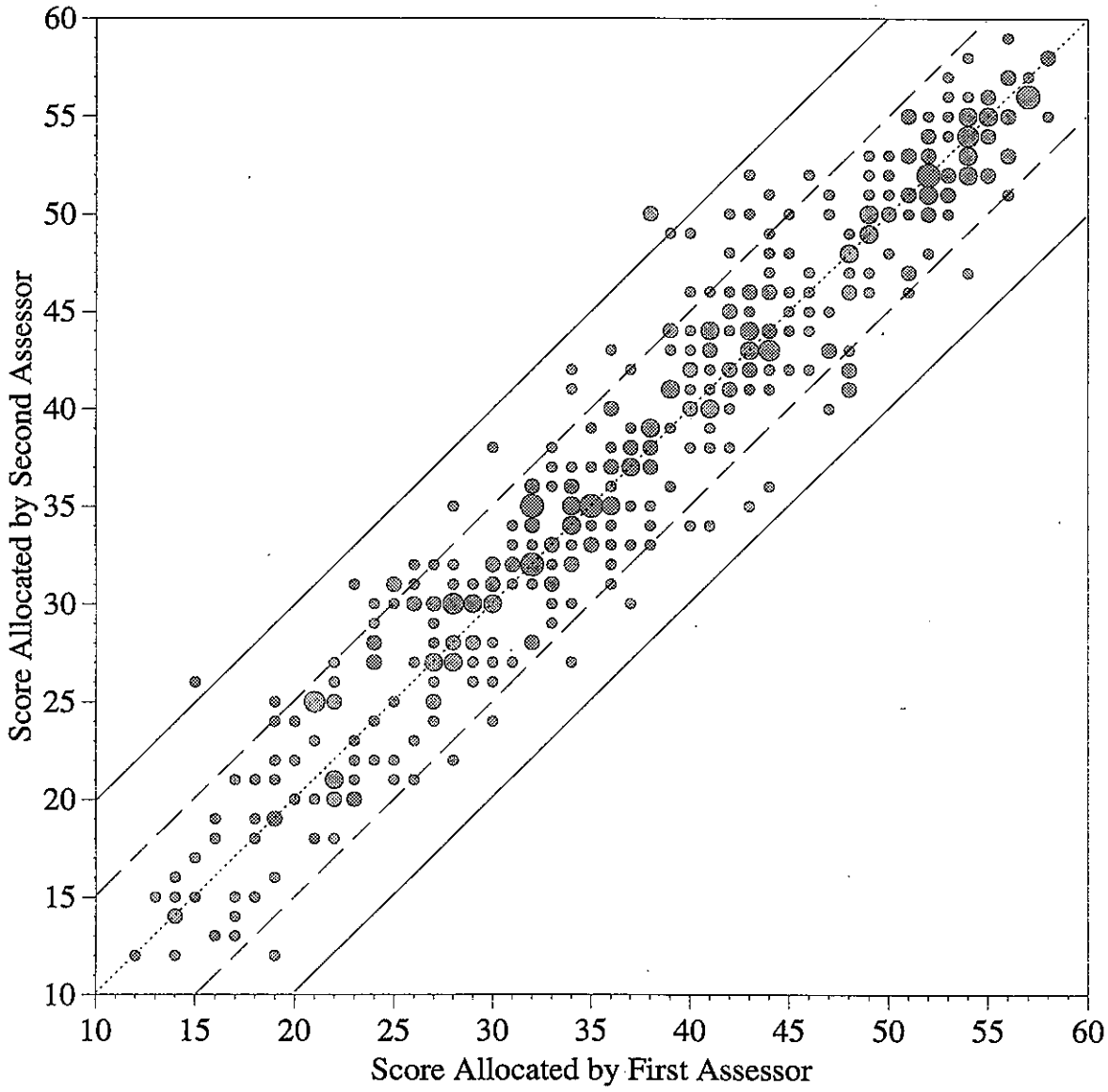


Figure 14. First and second ratings of folios plotted against each other (Modern History)

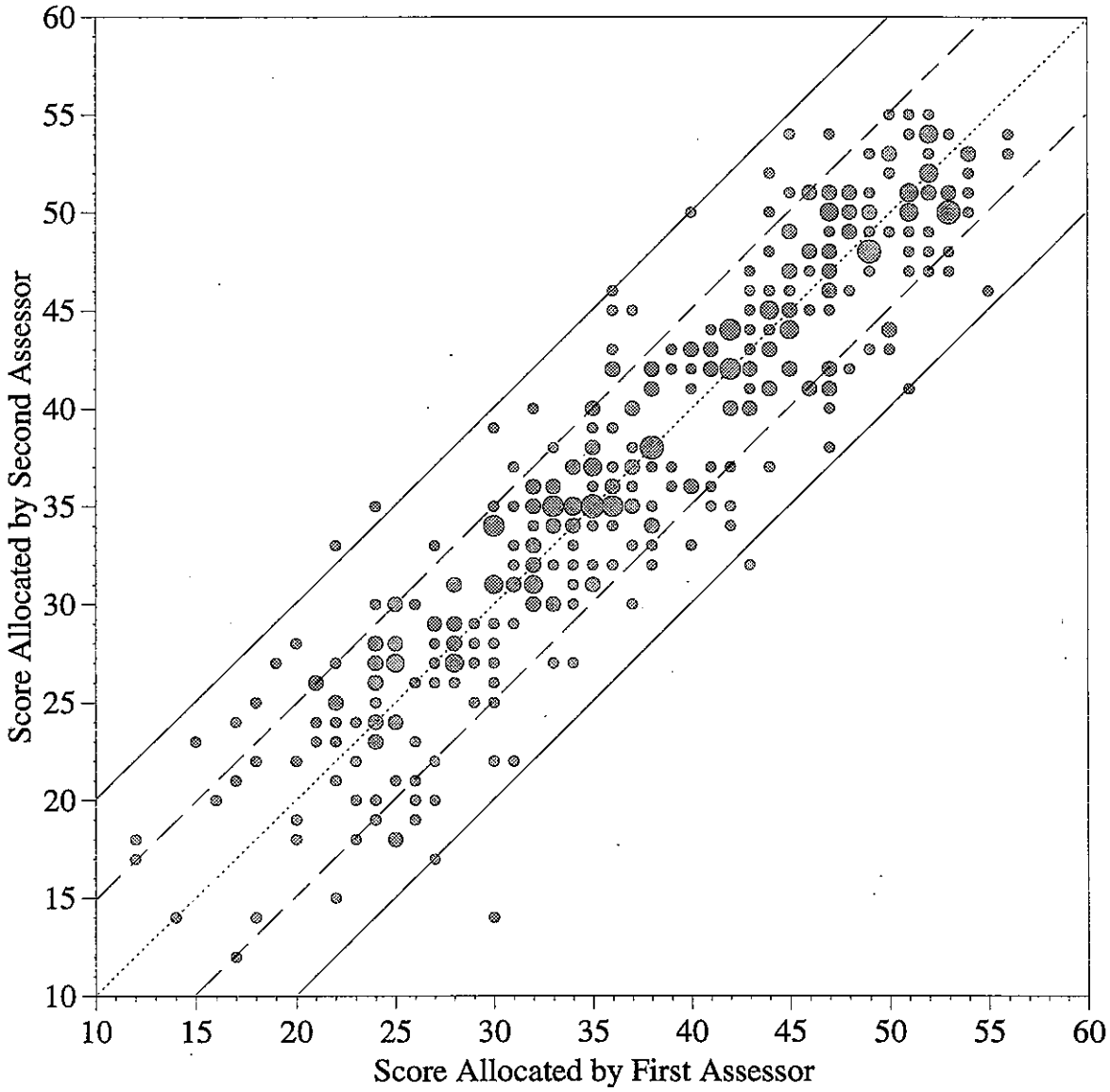


Figure 15 shows assessors' ratings plotted against exit Levels of Achievement in 1992. All six independent assessments of each folio in each subject are included in this picture. The shaded regions of Figure 15 indicate perfect agreement between assessors' ratings and exit Levels of Achievement in that assessors' ratings of 10 to 19 correspond to a VLA; ratings of 20 to 29 to an LA; ratings of 30 to 39 to an SA; and so on.

It can be seen from Figure 15 that the majority of assessors' ratings fall within the Level of Achievement actually awarded at the end of 1992. A proportion of assessors gave ratings above the awarded Level, and some gave ratings below the awarded Level. However, these discrepancies are almost all less than ten rungs (one Level of Achievement) in magnitude.

Table 5 shows the correlation between assessors' ratings and exit Levels of Achievement and the percentages of assessors' ratings in agreement with each Level of Achievement.

Figure 15. Assessors' ratings vs exit Levels of Achievement (all subjects combined)

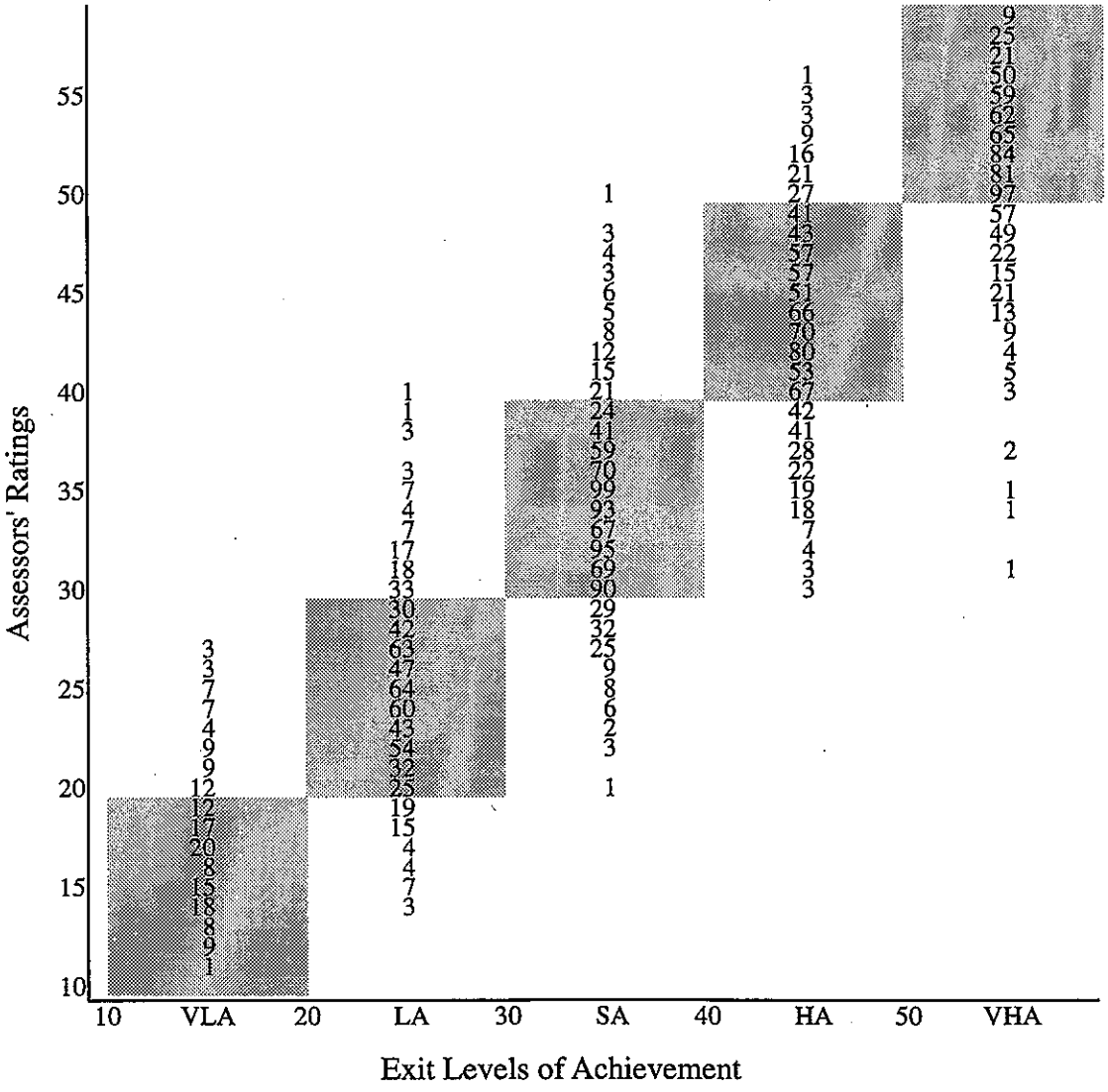


Table 5. Agreement between assessors' ratings and exit Levels of Achievement

	Percent Agreement					
	VLA	LA	SA	HA	VHA	Correlation
Independent assessment	67	76	79	68	73	.92
After collaboration	65	81	81	71	71	.93

As can be seen from Table 5, between 67 and 79 per cent of independent folio ratings by assessors in this study corresponded to the exit Level of Achievement awarded to those folios at the end of 1992. This represents a correlation between these two results of .92--a level of agreement similar to that obtained *between* assessors in this study. Given that a slightly lower correlation might be anticipated in this case because one set of results is expressed on a 5-point, rather than a 50-point scale, this is again an indication of an exceptionally high level of agreement between different assessors of the same work.

Another interesting question is whether assessors' ratings are more consistent with the 1992 exit Levels of Achievement after each pair of assessors has discussed a folio and given it a new rating. This question is addressed in Figure 16 and Table 5. It can be seen that after assessors collaborate in producing new ratings of folios their assessments are slightly more highly correlated with the original exit Levels of Achievement than were their initial, independent ratings. This difference is, however, very small.

5.3 Comparisons among schools

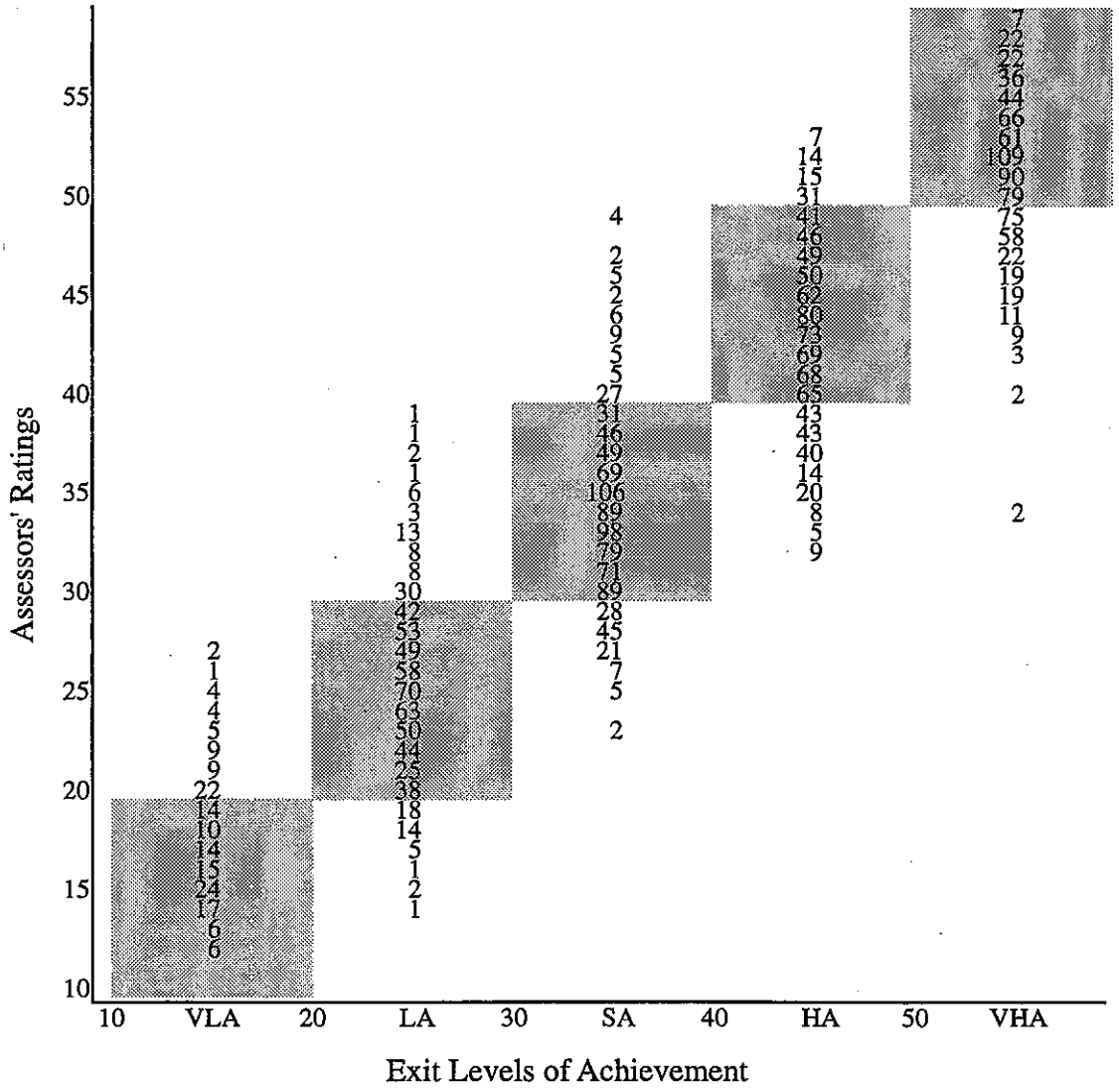
Although there is a generally high level of agreement between assessments of folios made in this study and exit Levels of Achievement, it is possible that agreement is lower in some schools than in others. To explore this possibility, the average difference between assessor rating and exit Level of Achievement has been calculated for each school in the study. These average differences are shown in Table 6 where they are expressed as proportions of a Level of Achievement.

From Table 6 it can be seen that assessments of English and History folios from two schools were, on average, between .3 and .4 of a Level of Achievement *above* those students' exit Levels of Achievement. Assessments of Chemistry folios from one school were more than .4 of a Level of Achievement *below* those students' exit Levels of Achievement. For most schools the mean difference was within the range -.2 to +.2 of a Level of Achievement.

Table 6. Number of schools with mean differences indicated

Mean Difference (Assessors-Exit)	English	Math 1	History	Chemistry
.3 to .4	1		1	
.2 to .3		1		
.1 to .2	1	1	1	1
-.1 to .1	4	6	9	7
-.2 to .1	4	4	1	5
-.3 to -.2	4	5	1	2
-.4 to -.3			2	
< -.4				1

Figure 16. Assessors' ratings vs exit Levels of Achievement after collaboration (all subjects combined)



5.4 Comparisons among assessors

In the same way it is possible to ask whether some assessors were more consistent with exit Levels of Achievement than others. Do some markers have higher standards than others, or do differences occur randomly across markers?

To address this question, differences between assessors' estimates and exit Levels of Achievement were analysed. Each assessor's average assessment was compared with the average Level of Achievement assigned to that set of folios on exit. Results are shown in Table 7.

It can be seen from Table 7 that most assessors' mean ratings were within .2 of a Level of Achievement from the exit assessments. Only one (English) assessor's mean rating differed by more than .3 of a Level of Achievement from the average exit result.

Table 7. Number of assessors with mean differences indicated

Mean Difference (Assessors-Exit)	English	Math 1	History	Chemistry
.3 to .4				
.2 to .3			2	
.1 to .2	1	1	1	
-.1 to .1	10	9	6	7
-.2 to .1	5	4	3	3
-.3 to -.2	3	4	1	1
-.4 to -.3	1			
< -.4				

6. Discussion

The central question addressed in this study is the question of the comparability of allocated Levels of Achievement across teachers and schools. Do different teachers have similar understandings of the criteria for each of the five Levels of Achievement in a subject and do they apply similar standards in allocating these Levels? This question is important because it determines the extent to which users of the Senior Certificate--in particular tertiary institutions and employers--can treat Levels of Achievement from different schools as directly comparable indicators of student performance.

As noted by Viviani in her review of tertiary entrance procedures in Queensland, there has been a general lack of confidence in the comparability of reported Levels of Achievement on the Senior Certificate. Part of the explanation for this lack of confidence may lie in the procedures by which details of curriculum and assessment are established in Queensland schools. Although the relevant Board syllabus provides a common framework for curriculum and assessment in a subject, each school's work program describes how that syllabus is to be implemented and the program to be used to assess student achievement and award Levels of Achievement. This raises a question about the comparability of assessments from school to school. Given that there are different assessment programs in different schools, can Levels of Achievement be compared meaningfully across these different programs? Or are Levels of Achievement meaningful only when considered in conjunction with each school's work program, and so comparable only within schools?

This study addressed the comparability issue by investigating the extent to which a group of teachers familiar with the relevant Board syllabus and experienced in the panel review process were able to produce equivalent assessments of a sample of student folios. The degree of agreement among markers can be expressed either in terms of differences in their assessments (e.g., how often do markers of the same folio disagree by a full level of achievement?) or, more usually, in terms of an inter-marker reliability coefficient. In this study, both approaches to studying inter-marker agreement were used.

An important question in a study of this kind is the question of an acceptable level of marker agreement (or, equivalently, a tolerable level of disagreement). Once again, there

are at least two different ways of setting an acceptable level of inter-marker agreement. The first is to set an *a priori* requirement in terms of absolute differences between assessors (e.g., a discrepancy of a full level of achievement in fewer than five per cent of cases). The second is to compare the achieved level of inter-marker reliability with levels of inter-marker reliability achieved in other relevant assessment programs. In this study, an *a priori* level was not set, but evidence was sought from other studies to provide a frame of reference for interpreting the levels of agreement obtained here.

By any measure, this study has revealed outstanding levels of inter-rater reliability. Between-marker agreement in allocating Levels of Achievement to student folios exceeds the level typically reported for systems of holistic scoring and is comparable with the highest levels reported for experienced markers of essays on external examinations. The fact that

- high levels of agreement were achieved across 62 assessors;
- the three assessment conditions imposed in this study consistently produced an inter-rater reliability of .94; and
- a similarly high level of agreement was obtained with the exit Levels of Achievement for these folios

suggests exceptional inter-marker consistency in the interpretation of standards and criteria.

A significant finding is that assessors were able to provide similarly high levels of inter-marker agreement with and without schools' work programs and whether considering folios in school groups or mixed randomly across schools. Lack of access to the school work program did not decrease assessors' abilities to produce highly consistent ratings of student work. Considering folios based on different school assessment programs also did not decrease the level of inter-marker consistency. These findings increase the likelihood that Levels of Achievement currently reported on the Senior Certificate are reasonably comparable from school to school and have a meaning independent of the details of the particular work programs from which they are derived.

The very high levels of inter-marker agreement obtained in this study are in part explained by the task the assessors were asked to perform. The folios considered by the assessors were collections of assessments made in each school. In many cases, particularly in some subjects, assessment folios were comprised mainly of tests that had been marked in the school. Many schools provided aggregate weighted scores for folios and, in some cases, written teacher feedback to students. Thus, the assessors were not marking student scripts but estimating a Level of Achievement for each folio on the basis of the marked tests and assignments it contained. Before giving folios to assessors, an attempt was made to remove any indication of the Level of Achievement considered appropriate by the teacher.

The task given to assessors is thus very different from the task of assessing an essay or a research project or a folio of student writing in the absence of any previous assessment or teacher comment on the piece of work. For this reason, the level of agreement among assessors might be expected to be significantly higher than the agreement obtained in a double blind assessment of the kind reported in other recent Australian studies.

Nevertheless, the important question in this study concerns the extent to which Levels of Achievement awarded by different teachers and schools represent equivalent performances regardless of school work programs and methods of assessment. The evidence of this study is that assessors are able to make highly consistent assessments of student work to provide Levels of Achievement that are largely comparable from school to school.

A final observation is provoked by the degree of confidence assessors expressed in their ratings of student folios. This is a remarkable feature of the data in this study. When asked to specify an upper and lower bound on their ratings, assessors expressed a confidence corresponding to ±2 rungs on a 50-rung scale for more than 80 per cent of their ratings. This level of marker confidence, coupled with the high level of inter-rater

agreement, raises a question about the current practice of reporting assessments on only a 5-point scale.

An advantage of reporting achievements on a coarse (e.g., 5-point) scale is that, for folios near the middle of any given Level of Achievement, that Level can be reported with a high degree of confidence. But this advantage is bought at a cost. The smaller the number of levels, the more significant the implications of misallocation for folios near boundaries between levels. The question of the appropriate number of levels is a policy issue that needs to take into account how Levels of Achievement are to be used. Certainly, the assessors in this study appear to believe that folio assessments could be made more precisely than they currently are.

7. Future Directions

The present study has focused primarily on the reliability with which assessors allocate Levels of Achievement to folios of student work. The study was undertaken as an exploratory study to establish benchmark data on levels of comparability under conditions similar to current Queensland review and certification processes and to inform future research directions concerning comparability.

Sadler (1993) suggests three broad directions for future research into comparability of school assessments in Queensland:

- undertake further reliability studies using other Board subjects and/or incorporating research strategies to test other hypotheses;
- undertake studies to examine threats to the validity of current assessment and review procedures and their impact on comparability;
- develop and administer reference tests to provide information about comparability of school-based assessments.

In the light of the findings of the present study, some thoughts on each of these possible research directions are offered.

Reliability Studies
In the present study, the four subjects included from the fifty-plus Board subjects were selected to be representative of subjects with large enrolments which figure prominently in university course requirements, and also to provide a balance between subjects which focus primarily on quantitative skills and processes and subjects which focus primarily on verbal skills and processes.

The three research models also were chosen to cover a range of conditions. Model 2 was designed to be reasonably similar to the Board's review and certification procedures which are known and understood by the assessors. Paralleling the Board's October review process, sets of nine folios of marked scripts from the same school were considered. These covered a range of levels of achievement and were considered together with the school's work program. The procedures differed from the normal Board process in that the name of the school was not provided, and in most cases the assessor was from a different review panel. Assessors were asked to work alone and did not have access to Levels of Achievement or other information usually provided on the Form R6 which accompanies folios at district review panels.

At the other extreme, Model 3 was designed to differ from the panel review process but to have sufficient similarities to allow assessors to estimate Levels of Achievement without

further training or additional time for review. Although the folios contained teacher marks and comments, each set contained five folios drawn from different schools and unaccompanied by the relevant work programs. Also, because they were sampled randomly, the five folios did not cover representative Levels of Achievement.

Future reliability studies could:

- select scripts which were not included in the Board's October Review Process;
- remove teachers' marks and comments from scripts;
- employ assessors not familiar with the review procedures used by the Board.

The potential disadvantage of using scripts from the Board's October Review Process is that assessors may have been alerted to the fact that they were dealing with the full range of achievement levels and so attempted to identify folios at mid-points and thresholds between levels. There is little evidence that this was occurring in practice, and the use of a random rather than systematic selection of scripts for assessment is unlikely to have an effect on the outcome of the study.

The disadvantage of providing marks and teacher comments is that assessors may be given strong guides to the levels assigned by teachers, meaning that the two sets of assessments are not truly independent assessments of the same work. This is a legitimate concern in relation to the present study. It is compounded by the suspicion that some teachers and assessors may associate levels of achievement with score ranges (e.g., percentages). Officers of the Board maintain that percentages needed to obtain various levels of achievement vary considerably across schools and even within schools as levels depend on trade-offs in achieving different criteria, not on percentages. However, the extent to which teachers and assessors make use of numerical scores as guides to the allocation of Levels of Achievement is uncertain.

It could be argued that the comparability of school-based assessments depends essentially on the extent to which teachers share a common understanding of criteria and standards for assessment. From this perspective, the comparability question is a question not so much about the reliability of the Board's review process as a question about the extent to which teachers throughout the system would allocate the same levels of achievement to samples of student work. A third possibility for future reliability studies would be to ask classroom teachers (rather than selected reviewers) to allocate levels of achievement to samples of student work—in the absence of marks, written comments, and work programs—to establish the extent to which criteria and standards are being applied consistently in schools.

Validity Studies

A second approach to the comparability issue would be to investigate assessment processes in more detail. To what extent do different teachers pay attention to the same features of student work, apply similar weightings to different criteria, use similar procedures for deciding on the final level of achievement? As Sadler (1993) points out, research of this kind would be intensive, qualitative, analytic, and to some extent philosophical. Although validity studies may provide valuable information about the extent to which assessment *processes* are comparable from assessor to assessor, they do not directly address the question of the comparability of *outcomes* in the way that reliability studies do.

Reference Tests

Reliability studies address the question of the extent to which assessments made by one assessor are consistent with assessments made by other assessors. Validity studies address the question of the extent to which different assessors apply similar processes and attend to relevant criteria in making their assessments. Studies based on reference tests also address the question of consistency. In this case, the question is: How consistent are teachers' assessments of student achievement with independent evidence based on test performances?

A reference test is a subject-specific test which examines core knowledge and skills in a subject area. Sadler (1993, 23) defines 'core' as the combined body of subject matter and specific skills which are characteristic of and in some cases peculiar to a subject. The use of reference tests was supported in both the Radford and ROSBA reviews.

Reference tests could be developed in a small number of subjects (e.g., English and mathematics) and used as a component of Year 12 assessment on an ongoing basis. Alternatively, reference tests of this kind might be developed merely for the purposes of research into levels of comparability of school-based assessments.

If a reference test is to be used on an ongoing basis, there are several ways in which such a test might be used. On one hand, a test in, say English, could be used as point of reference for the automatic rescaling of school assessments in English in an attempt to bring them to a common scale and to make them comparable. This use of a reference test is similar to the use that some Australian states make of an external subject examination as a point of reference for rescaling school assessments in a subject.

An alternative use of a reference test (Hill, Brown & Masters, 1993) is as a frame of reference not for the automatic scaling of teachers' assessments, but for identifying and following up schools whose school-based assessments appear unexpectedly high or unexpectedly low in relation to results on the test.

References

Board of Senior Secondary School Studies (1993). *Handbook of Procedures for Accreditation and Certification*, Brisbane.

Brown, T. & Ball, S. (1992). A Report on the VCE Verification Process. Victorian Curriculum and Assessment Board, Melbourne.

Eyers, V. (1993). Commentary on the Brown and Ball report (unpublished), February 1993.

Hill, P.W. (1993). Comments on 'A Report on the VCE Verification Process', Melbourne.

Hill, P.W., Brown, T., & Masters, G.N. (1993) *Fair and Authentic School Assessment*. Advice to the Victorian Board of Studies on verification, scaling, and reporting of results in the VCE. Melbourne.

McGaw, B. (1993). School based and external assessment: International, interstate and local issues. In IARTV *School based and external assessment: Uses and issues in the postcompulsory years*, IARTV Seminar Series, July 1993, No.26. Jolimont, Vic.: Incorporated Association of Registered Teachers of Victoria, pp.3-9.

Radford, W.C., *et al.*, (1970). *Public Examinations for Queensland Secondary School Students*. Queensland Department of Education, Brisbane.

Sadler, D. R. (1993). Comparability in school-based assessment in Queensland secondary schools. Tertiary Entrance Procedures Authority, Brisbane

Scott. E., et al., (1978). *A Review of School-Based Assessment in Queensland Secondary Schools*. Board of Secondary School Studies, Brisbane.

Viviani, N. *The Review of Tertiary Entrance in Queensland, 1990*. Report submitted to the Minister for Education, Brisbane.