Assessment strategies for directing further learning in mathematics

John F. Izard
ACER
9 Frederick Street
Hawthorn Vic 3122
Australia

Christopher R. Haines
Department of Mathematics
City University
London EC1V 0HB
United Kingdom

Paper presented at the 1992 Joint Conference of the Australian Association for Research in Education and the New Zealand Association for Research in Education, Deakin University, Geelong, 22-26 November 1992.

[Note for the reader of this paper in computer file format: The text is presented first; tables and figures are collected at the end of the file (after the references). Tables and figures may have lines longer than 80 characters.]

The context for learning

The structure within which the teaching and learning of mathematics takes place is complex and the methods employed both by the student and the teacher are many and various. Within the tertiary sector the traditional approach to motivate students has been to present a few specific examples in an exciting environment before moving on to deal with a topic in its generality within some global structure. For example, visualising a cyclist passing through traffic lights or the sighting of a kingfisher provides the fixative in a photographic sense for the central mathematical concepts. Examples such as these are often used by the teacher and are given to the student to enhance the learning environment. They are, however, not 'owned' by the student so that the impact on the learning process is not as great in all cases as it otherwise might be.

Investigational and project work has long been an established feature in primary and secondary education where the learning experience has a considerable emphasis on the pupil as an individual. Generally, in the UK, the tertiary sector has been reluctant to commit itself to these free expressions of student competence citing the strictures of the three-year honours degree course, the amount of material which must be covered and the attendant problems of assessment. It is recognised that projects and investigational work encourage the student to think independently and to show initiative in ways which lead to greater levels of achievement and course satisfaction. Further, the development of individual and group communication skills are important (Haines,1991a). Just as the UK national curriculum for schools identifies attainment targets concerned with using and applying mathematics, in the tertiary sector it is the use and application of mathematics in real situations which has been a major feature of modelling courses.

The variety and choice of projects which are open to individuals challenges

us to take account of added value achieved by the student. How has the student developed relative to the starting point and how have his or her capabilities been extended? The student with no previous experience in complex analysis, who is successful in a project on mappings and stereographic projections has gained considerably in the content area as well as in the process skills of the activity.

The place and nature of assessment in relation to projects at the tertiary level has been addressed by Berry and Haines (1991) in which a start has been made on the task of describing those individual attributes which constitute a project in all its generalities. It is important that dependable measures are developed for incorporation in rating and assessment schemes to reflect the dynamic nature of the varying inputs of student work. Further, without such measures one cannot assess the value-added component. Item response theory provides possible approach to developing such measures which has wide applicability and has been used within curriculum frameworks throughout the world (Izard, 1992).

This paper describes research in which comprehensive descriptors of oral and written achievement were developed to assess complex behaviour (as observed in mathematical modelling projects undertaken at British universities) to recognise high achievement and to reward curriculum-intended achievement. Analyses of non-dichotomous data from trials of the scheme are presented and interpreted in terms of the consistency and precision of assessing candidate achievement, the stringency of examiners, and the differing demands of particular course requirements.

The role of assessment

Advocates of the teaching of mathematics through modelling activities in the classroom consider that these activities are more authentic, and address such situations where there are multiple (correct) solutions, where there are a diversity of problem-solving approaches, or where the skills cannot be demonstrated easily in pencil-and-paper format (as in traditional examinations). The assessment practices that are adopted with such teaching of mathematics through modelling activities also play a crucial role. Assessment has the function of providing valid evidence of learning achievement to inform students, to facilitate provision of further learning, or to certify that a required level has been reached. Such information is of particular relevance to individual students and their teachers. Teachers are able to develop and improve the educational process if they have identified the strengths of their pupils and know which areas of study require attention. "The information provided by assessment should do more than portray a learner's level of performance. It should inform the actions of all participants in the learning situation... Links must be forged between the assessment, the instruction it reviews, and the instruction it anticipates" (Clarke, 1989, p.4).

Gasking (1948) argues that examinations exert considerable pressure on the subjects taught, on the topics within those subjects, and on the teaching strategies which are used. He draws two conclusions (p.10). "First, a reformer, a teacher or an administrator, who wishes to introduce certain changes into school education, will, if the examinations remain the same, find his efforts defeated in the long run, unless the changes he introduces increase the chance of examination success." "But conversely, if a reformer succeeds in introducing changes in the type of examination, he will automatically tend to bring about such changes in the content and method of education as are likely to make for success in the changed type of examination". Gasking also discusses the effects of indirect measurement as compared with the direct measurement of relevant skills and knowledge. "The easiest way ... for teachers to get good results is to concentrate on imparting those capacities which are directly measured in the examination, even if these are not the real objectives of the course, but are merely capacities which the examiners are trying to use as indirect measures. Thus education is perverted away from its real objectives, on to that which the examiners use as indirect measures" (p. 14). Since the easiest things to measure relatively objectively are knowledge of facts (particularly isolated facts rather than interrelationships) and knowledge of the verbal expression of theories, generalisations and definitions, examinations tend to have high proportions of such questions, rather than questions which test understanding, or applications in novel circumstances. Initially, using factual recall provides an indirect measure of the higher order skills based on the assumption of a high correlation between the intended skills and those actually measured. However intensive coaching in the indirect criterion tends to reduce this correlation, and the indirect measure ceases to be a valid indicator of the true objectives (Gasking, 1948, p. 15). Madaus (1988) also emphasizes the influence of assessment practices on curriculum.

Gasking points out that when examiners perceive that students without the intended skills are passing, they may raise the standard required in the indirect criterion. This diverts schools even more from the intended skills and makes the examination an even less valid measure of those skills. The only solution is to devise direct measures. In the teaching of mathematics through modelling activities, it is important that the assessment strategy matches the curriculum intention.

To provide a more comprehensive picture of students' learning achievements, assessment strategies are being extended to cover knowledge, understandings, skills and personal qualities not assessed by traditional tests, such as the development of initiative, taking responsibility for learning, and applying problem-solving strategies. Extending assessment strategies to include more realistic tasks relies on the reasonable presumption that the assessment tasks should accurately reflect the curriculum intentions. Use of mathematical projects, in whatever form, as

assessment tasks has the potential for ensuring that the tasks mirror the desired skills, thus contributing to valid assessment. However the very wide range of possible responses makes the use of such assessment strategies problematic unless better procedures are devised.

The procedures need to deal with complexity, encourage better communication between examinee and examiner, and lead to more consistent and relevant assessment. The procedures should be feasible in current practice and should give information about how confident the examiner is about the precision of the assessments. The assessment strategies adopted for such activities also need to cater for this diversity to meet accountability requirements - fairness to each student so that students are rewarded in accord with their genuine knowledge and skill, providing students with information to support their own learning, and providing trustworthy certification which is informative for those who will use the assessments as part of the selection process for entry to later stages of education or for future employment. However most teachers wish to reduce the time devoted to the assessment task to the absolute minimum and to make the task as simple as possible.

The critical environment

Universities in the UK and elsewhere have developed a system which incorporates internal and external reviews of teaching, learning and assessment procedures associated with undergraduate courses. In the case of assessment these procedures have identified areas for concern which were recognised by the institutions but which were thought to be intractable and therefore not given adequate consideration. An internal review (1987) reported

> "The committee had been concerned to note that projects were not given marks or grades, ... a system that gave no real guide to students on standards"

a statement which is consistent with a reluctance by academics to address the marking of projects in a systematic way. That this is not unique is typified by the following external examiner's report (1991):

> "... students benefitted greatly from doing them ... it is difficult to compare different projects and hence to mark fairly. I would like to see a much more detailed report from the examiners, justifying the marks awarded ... ensuring disciplined marking ..."

Note that in this comment, the academic worth of project work is accepted and suggestions are made to improve the assessment procedures.

Although the report above does not put the view, a common response to the difficulties with assessing complex mathematical tasks is to reduce the weighting of that component within the whole degree scheme:

> "It might be advisable to give some thought to the weighting given to the third year project, as it is
>
> extremely difficult to compare projects on widely differing topics" (external examiner 1990).

Contemporary views treat assessment as a dynamic process which develops student learning, recognising the formative nature of certain modes of assessment (Izard, 1992). In the tertiary sector the internal review procedures (1991) support this view :

> "... students received no feedback on the projects ... The committee considered that the absence of a numerical mark, and of feedback, provided no real guidance ..."

but the response at course level has been slow. It has been difficult for mathematicians to appreciate the importance of identifying key descriptors which can be used to mark projects in a systematic way. The problem of achieving assessment comparability between projects on widely differing subjects remains.

Aggregating marks in a global scheme

Taking the view that in mathematical modelling projects it is important that student can show evidence of achievement in each of the subsections of any global scheme, Haines (1991b) compared the results of aggregating marks using a weighted arithmetic mean and a geometric mean using data from City University for the 1986-88 projects. The results showed that, in practice, marks would be reduced using a weighted geometric mean by at most 3% provided that the variation in the weights was not extreme.

The weights themselves were investigated in a study by Izard (1991) in which the same data were analysed alongside Australian data from assessments of English and from assessments of spatial ability using subtests from a German medical entrance test. Table 1 shows that the actual weights derived from the marks awarded on the Haines 1986-88 data differ significantly from the weights intended by the project marking scheme. In particular the examiner achieved weights for the first two categories have reversed the intended weights and those for the other categories show a significant variation from the intended pattern. Izard (1991) shows that these discrepancies are consistent with the assessment results for other fields; in this respect mathematics is no different from other subjects.

[Table 1 about here.]

Taking into account the results of the Izard (1991) analysis and aware of the criticism of voiced by both internal reviews and external examiners the

instructions to the examiners were modified and the intended weights between the categories were changed. A checklist of key descriptors covering 28 aspects of project work was constructed to assist examiners in awarding credit for student achievement in the five categories, (Haines, 1991a). The checklist served to make examiners aware of the processes which were involved in project work at tertiary level and to begin the move towards a systematic approach to marking such project work. The significance of the checklist was not lost on the external examiner (1991), although the conclusion presented in the resulting report was premature.

> "... a considerable (and seemingly successful) effort has been made to devise a marking scheme to ensure consistency between marks given to the projects ..."

The weights were changed from 1:2:5:1:1 (used in 1986-88) to 5:8:25:4:8 (used in 1989-91) to reflect more accurately the perceived importance of 'Content' with 50% of the marks and to reflect the value attached to the way in which the project was carried out as deduced from judging the written report.

Investigating weights and categories

Data for the 1989-91 projects were gathered and the actual weights for the two three-year periods were compared with the intended weights. The five categories were investigated in depth by assigning them, for each three-year period, to a continuum which gave a measure of whether students were likely to achieve a low mark or a high one for that particular category.

As far as the weightings were concerned, the pattern for the 1989-91 data was similar to the pattern in the 1986-88 data; the change in instructions to the examiners and the additional checklist did not eliminate the discrepancy between the intended and the actual weightings. The intended weights in the second (1989-91) batch, altered from those in the first (1986-88) batch, were not reflected in the actual weights of the marks awarded. Further, even where the intended weight of a category remained the same in the global scheme, a change in the intended weight of other categories had an effect on the actual weights of that category and consequently on the apparent level of student achievement within that category (Haines, 1991a).

Figure 1 shows the data for 1986-88; the left-hand side shows each student project assigned to a position an achievement continuum ranging from - 4.0 to + 4.0. The raw scores for each student were scaled onto a four-point scale before analysis and each student in this data set exceeded the pass mark which was set at 40% in raw score terms. The right-hand side of Figure 1, obtained using the QUEST software package (Adams and Khoo, 1991) shows thresholds where the probability of a particular category score

changes. For example, 10 students are shown on the achievement scale at 0.0. On the right-hand side it can be seen that in the second category (Introduction), students at this achievement level are likely to score 2 marks out of 3. Similarly, the next column shows that in the third category (Content), these same students are likely to score 2 marks. This threshold is indicated in the figure by the code 3.2.

Figure 2 shows the data for 1989-91 on the same basis. Note, for example, that on the right-hand side the threshold 4.1 appears at a lower achievement than for the earlier set of results. This is the level at which a score of 1 is more likely in the fourth category (References and Data Sources). Each threshold for this category has moved down the scale reflecting a greater relative understanding amongst the students.

The quality of communication about the intended tasks within the categories varied considerably. Some information was vague and nebulous while other information was well-defined. Figures 1 and 2 show that the well-structured and well-defined tasks were easy to achieve by the students. The first category (Abstract) in both data sets, and the fourth category (References and Data Sources) in the second data set illustrate this. Conversely, tasks which were ill-defined or more nebulous were more difficult to achieve such as the third category (Content), in both data sets. Where this has happened, it is evident that students had been ill informed about the requirements of each task within the project and about the weightings that these tasks would receive from the examiners. The assessment scheme can be modelled from the data to give insights into the probability that a student at a particular point on the ability scale achieves a score of 0, 1, 2 or 3 for a given category. For example, taking the fifth category (Presentation), for which the requirements are well defined, Figure 2 shows that it is easy for students to achieve 1 mark, few students achieve 2 marks and many students gain the full 3 marks.

This behaviour is modelled in Figure 3 and is interpreted as follows. A student at the low achievement end of the achievement scale (-4) has a probability of about 0.1 of achieving 1 mark and 0.9 of achieving 0 marks. The probability distributions for 0, 1, 2 and 3 marks show a clear threshold from 0 to 1 mark, but notice that the distributions for 1, 2 and 3 are almost concurrent at a particular point on the ability scale. The

central part of the mark scale does not appear to be used when assessing the fifth category (Presentation).

The issue is further complicated where the examiners depart from the assessment guidelines by such as providing incomplete and inconsistent returns. For example, students in some instances have "... attempted a relatively difficult task ..." but "... it has not come off ..." as the supervisors expected. Yet the same supervisor claims that the project does what it set out to do. In other cases the checklist has been used but the checklist assessment on content does not appear to support the high marks.

Surprisingly some projects do not contain "... any logical reasoned arguments ...". There is also a variation in the degree and quality of supervisor support for students during the projects.

[Figures 1, 2 and 3 about here.]

Developing a comprehensive assessment strategy

A group of leading practitioners in the development of project schemes and in the assessment of such projects met in Exeter, United Kingdom, in April 1991 to

1 Review current project schemes and their assessment of project work across a wide range of institutions and levels, and

2 Begin to develop criteria-based assessment procedures for use on a wide range of topics within mathematics.

Schools, colleges, polytechnics (now universities) and universities were represented (See Appendix), ensuring that the projects which were available were completed within course frameworks which dealt with diverse aims and objectives. The projects brought to this workshop dealt with different topics, were completed at different ages, and demonstrated a variety of achievement levels. All this ensured that any assessment was not too narrowly focused on content and year level, and that any process which rated projects could be applied to real differences in achievement (Berry and Haines, 1991).

A preliminary consideration was given to the descriptors or indicators which might be appropriate for objective assessments of projects. Two working groups began the task by first reading the nine selected projects to form an initial impression and to rank all nine pieces of work. A second working session identified three main areas or themes:

1 Activity of the Investigation: processes in the constituent information structure,

2 Integration: ability to integrate knowledge and skills to tackle a particular problem, and

3 Communication: delivery encompassing the written report, group work and oral presentation.

Descriptors of each of these themes were then developed and rating sheets produced. The descriptors included phrases like 'fluent and effective use of language' and those using the rating scale were asked to assign a rating of 2, 1, or 0 to that descriptor according to whether the project being assessed showed comprehensive evidence, some evidence or no evidence of this behaviour. An inappropriate category,

rated X, was included for cases where the marker felt the descriptor to be inapplicable to that project. Six of the nine projects were remarked using the rating sheets and a detailed item response model analysis of this activity was carried out by Izard (In Berry and Haines, 1991) using the FACETS computer package (Linacre, 1990).

There were three facets in the design of the analyses: Judges, Persons/Projects, and Items (descriptors). (This report uses the term judge to indicate a person who is faced with the task of assessing students work within the investigational framework of the research). The analyses used a rating scale model (with allowed categories: 0, 1, 2) for all judges, for all persons/projects, and for all items with a weight of 1. (That is, each judge was regarded as of equal importance, each person/project was regarded equally, and each of the 29 descriptors was regarded as of equal importance.) The 'inappropriate' category shown as X on the sheets was excluded from the analyses. There were 676 usable ratings (defined as a 0, 1, or 2 rating on an item by a judge for a person/project).

There were significant differences between the achievement levels of students as inferred from the evidence of the projects. Figure 4 represents the projects on a linear scale as a measure of achievement. The numerals immediately above the axis are frequencies and the numerals linked to them by vertical lines are codes to identify the projects.

[Figure 4 about here.]

The rankings for the projects are given in Table 2, which shows that changing the assessment process through discussion, and concentrating on descriptors, has resulted in a change in the overall rating. Projects 9 and 2 were ranked some distance below the other projects on the achievement scale (See Figure 4) although it was not possible to determine a consensus position for a 'pass' mark since the 'pass' mark for each institution was not known.

[Table 2 about here.]

Figure 5 shows the judges on a linear scale which could be described as a leniency-stringency dimension (in logit units and in the same metric as for Figure 4). The numerals immediately above the axis are frequencies and the numerals linked to them by vertical lines are codes to identify the judges. Note that all judges except judges 7 and 4 have similar patterns but vary in their stringency. judge 4 differs from the others by making less extreme distinctions between the projects; judge 7 differs by making more extreme distinctions between the projects. No inference should be made that one type of pattern is better than another. The remarkable consistency between

judges (reliability 0.67), even though some are more lenient than others, is cause for celebration but not complacency. Developing strategies which improve such assessment of complex tasks remains an objective.

[Figure 5 about here.]

At first sight Table 2 highlights disagreement between the judges on the assessment of the projects prior to the development of the descriptors, but Figure 6 shows a positive correlation, 0.57, between the two groups of judges on the 7 ranked cases (the regression line passes through the two points marked R on the axes). The individual behaviour discussed above, and the group behaviour of the judges gives some ground for confidence in the quality of their assessment.

The 29 descriptors themselves were analysed for consistency and for the demands which they placed on the students. (The descriptors were placed on a continuous demand scale similar to Figure 5). Of the three themes, the descriptors under 'Activity of Investigation' included the extremes of demand. The descriptors

'identified the main features of the task' and
'relationships between variables established'

were the easiest to satisfy, whilst

'hypothesis stated and established' and
'good use of human resources made'

were the indicators on which it was most difficult to achieve a good score. In addition, redundancy was detected under the same theme between

'results interpreted and validated' and
'conjectures analysed'

whilst the model did not fit

'generalisations made and proofs attempted'

which suggests that this descriptor should have been separated into two distinct items.

The workshop explored the central issues behind marking projects and suggested categories and indicators, with associated rating sheets, which might be used for this purpose.

[Figure 6 about here.]

Further data were collected following the Exeter workshop (Izard and Haines, 1991) in order to develop and test these strategies for producing effective rating scales in the tertiary sector. As part of this exercise, the projects which were analysed at Exeter were assessed by a sub-group of the same judges using a larger different collection of descriptors which were not clustered into categories. Figure 7 shows that the assessment process using the two different

sets of descriptors gave almost identical results, with a correlation of 0.92. We consider that this reflects the importance of the judges sharing the same meanings, rather than attempting to find a unique set of descriptors.

[Figure 7 about here.]

Developing the oral descriptors

Oral communication descriptors developed at Exeter (Berry and Haines, 1991) were tested in a pilot investigation alongside the official assessment procedure at Brighton University.

In the first stage of the investigation two judges took the thirteen descriptors to the group presentations of a project on insurance given by diploma students. At this early stage the shortcomings of a three point scale with an inappropriate category were evident. Difficulties were also found with the assigning of individual and group measures of performance, for example it was not clear that referring to visual aids, the term 'well explained' was or was not an individual attribute. The data obtained from this exercise were too sparse and inconsistent to give any reliable indication of performance.

In the second stage four judges used the same descriptors in group presentations given by the same students on a modelling project. Of the four judges two were internal examiners. The topic and the associated assessment was the same for all groups but there was no requirement that the presentation of the report should be handled in the same way. A video of the presentations was made to allow further follow up analysis.

At a group level each of the four judges ranked the 10 groups by reference to the communication skills guidelines but not by using any quantitative measure. A FACET analysis was carried out on the data provided by three of the four judges using the descriptors suggested in Berry et al. (1991). The data provided by the fourth judge was very sparse and so was omitted as atypical. Although global assessment marks correlated with scores obtained using the FACET analysis there were substantial discrepancies at the extremes. An inference which might be drawn is that the judges did not have shared meanings for the descriptors. They did not describe those qualities which the internal judges held to be important, at least where work of

the highest and lowest standards were concerned.

[Figure 8 about here.]

Turning now to individual student assessments, the responses from three of the judges were analysed using FACETS and their performance

was compared. It can be seen that there was some agreement between judges 1 and 2, (Correlation 0.46, see Figure 5) although the comparison between judge 1 and judge 3 (See Figure 6) shows an even lower correlation (0.11), of about the same order as that between judge 2 and judge 3 (0.16).

[Figure 9 about here.]

The internal examination scheme required that each student be graded, fail, pass, merit or distinction. In the analysis, these four grades were coded 0, 1, 2 and 3 respectively. The grades awarded by the internal examiners are compared with those which would have been awarded had the FACET analysis been used on descriptors of performance for each individual student (See Figure 10). The vertical scale in Figure 10 shows the students clustered at the four levels according to the global assessments of the internal examiners, and the horizontal scale shows each student rated on an ability scale using the FACET analysis. Whilst there is a positive correlation (0.71) between the two measures, there are substantial discrepancies which give cause for concern.

At the right-hand side of the Figure 10, there are three students who appear to be candidates for a distinction, one of whom achieved a bare pass. There is evidence of a considerable overlap in the merit and pass categories, which suggests that the criteria used internally may not be well-defined and/or that the descriptors either do not measure student achievement as intended or that the judges do not share the same meanings for the descriptors.

[Figure 10 about here.]

Issues to be addressed

There are many areas which require further information, research and analysis. Developmental changes in the approach to learning as described above require different assessment strategies, but which are nonetheless well defined, easily understood by the students and the examiners and can be implemented in a straight-forward fashion. The project scheme at City University allows students in their final year a wide choice of subject area provided that there is a sound mathematical basis in the chosen area. Each student completes an individual project on different topics from one another over a period of six months. This is quite different from a first year group project completed in a few weeks on a specific modelling topic which is the same for all groups. Given the diverse nature of project work it is important to understand the context within which the project has been set, the quality and extent of supervision, and therefore to develop an assessment scheme that takes account of these factors, perhaps through rating scales which include such contextual

indicators.

In a student-centred learning environment it is increasingly important to understand and appreciate what has occurred in the interval between a student embarking upon a task and the completion of that task. Many assessment schemes concentrate on examining the end result paying little attention to the processes which exercise the student. In embarking on project and investigational work, each student whether on the same course or not, whether at the same formal achievement level or not, starts at a particular point in their personal mathematical

development. Even where students have followed the same introductory courses leading to the project, the starting point for each student is different.

It is important that these starting points are identified so that improvements and other changes can be detected and so that student achievement can be rewarded. In order to fix these starting points an appropriate measure is needed which deals accurately with the issues involved. The improvement may then be judged as a measure on this achievement scale.

Project work is firmly student-centred and as such it is natural for the student to be involved in the assessment process. Different strategies are appropriate for assessments which are formative rather than those which occur at the end of a particular activity. If students are to make real contributions to their own assessments, then academic staff must be sure that those factors have some integrity. Descriptors which reflect the student's own assessment of achievement will need careful development and testing in order to be accepted within any global scheme.

Questions of consistency and accountability highlight the need to monitor whether examiners share the same views about student achievement. Examiners look for different features, give those features differing emphasis, have different views of complexity, and therefore assign differing values to the work being assessed and vary in their assigned marks. The research described in this paper highlights problems associated with global assessment schemes, but also helps gain an understanding of the intended goals set for students within each category. Global assessment schemes, in particular, allocate marks to different categories which are weighted before aggregation. There is a need to test the assumption that the actual marks assigned by the examiners to the categories achieve the weights intended by the course scheme. In our research, changing the intended weights for the various categories did not resolve the discrepancies between intended and actual weights. Using a checklist or rating scale with agreed descriptors gave more reliable indications of achievement, and as a corollary, more control over the weighting

problem. The rating scale needs to be comprehensive (so that all high achievement is recognised) and to be relevant (so that curriculum-intended achievement is rewarded). It must also be internally consistent so that assessments do not vary too much for work that is of comparable standard.

Item response theory is a powerful tool in the quest for credible assessment schemes for mathematical modelling and other projects. The analyses presented in this paper show that in developing such schemes, thorough testing and research is necessary. However, the rating scales developed in one group of educational institutions may not transfer readily to other institutions. Agreed descriptors must be devised which make sense in the learning context of the particular institutions. The descriptors themselves can be seen as an interpretation of the detailed aims and objectives of the particular course or courses. Investigations are necessary to devise descriptors and to test those descriptors during development. Further, the development of agreed descriptors has the effect of clarifying shared aims for the teaching staff involved in the process. This experience has to be gained through applying descriptors to real student work, the same work that other colleagues have assessed also. Those who judge the quality of mathematical projects should share the same meanings for all or most of the descriptors.

As projects may differ in the demands they make of candidates, choosing a project which makes less demand may give a student an advantage by allowing mastery of a topic to be demonstrated more easily. If each person does a different project, and no person does

more than one project, there is no way in which the influence of the project can be separated from the achievement of the candidate. In a similar way, the leniency/stringency and consistency of rating by judges cannot be assessed unless each piece of work is rated by more than one judge and the judges overlap in the work they rate.

Although progress has been made, the central questions relating to learning mathematics through project work remain. It is important that the intended skills are identified in the descriptors and that the assessment process does not distort the curriculum intentions.

References

Adams, R.J. and Khoo, S.T. (1991). Quest: The interactive test analysis system. Hawthorn, Vic.: Australian Council for Educational Research.

Berry, J.S. and Haines, C.R. (1991). Criteria and assessment procedures for projects in mathematics. Workshop report

CTM75, Exeter, 15-17 April 1991. Plymouth: Poly South West. 26pp.

Clarke, D.J. (1989). Mathematics Curriculum and Teaching Program (MCTP): Alternatives in Mathematics. Canberra, ACT.: Curriculum Development Centre.

Gasking, D.A.T. (1948). Examinations and the aims of education. Carlton, Vic.: Melbourne University Press.

Haines, C.R. (1991a). Developing assessment strategies for mathematics projects. Keynote address at conference on Assessment in the Mathematical Sciences, Geelong, Australia, 20-24 November, 1991. [Now published in M. Stephens and J. Izard, (Eds.) (1992) Reshaping assessment practices: Assessment in the mathematical sciences under challenge, (pp. 127-141). Hawthorn, Vic.: Australian Council for Educational Research.]

Haines, C.R. (1991b). Project assessment for mathematicians. In M. Niss et al. (Eds.) Teaching of mathematical modelling and applications, (pp. 209-305). Chichester: Ellis Horwood.

Haines, C.R. (1991c). Assessing mathematical science projects. International Journal of Mathematics Education for Science and Technology, 22, 97-101.

Izard, J. (1991). Issues in the assessment of non-objective and objective examination tasks. In A.J.M. Luitjen, (Ed.). Issues in public examinations, (Proceedings of the Sixteenth IAEA Conference. Maastricht, The Netherlands, 18-22 June 1990), (pp. 73-83). Utrecht, The Netherlands: Lemma, B.V.

Izard, J.F. (1992). Assessing learning achievement. (Educational studies and documents, 60.) Paris: United Nations Educational, Scientific and Cultural Organisation.

Izard, J. and Haines, C.R. (1991). Marking schemes for projects in mathematics and its applications: Interpreting the results. Paper presented at ICTMA5, Utrecht, The Netherlands, 9-13 September 1991.

Linacre, J.M. (1990). Modelling rating scales. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA., United States of America, 16-20 April,1990. [ED 318 803]

Madaus, G.F. (1988). The influence of testing on the curriculum. In L.N. Tanner (Ed). Critical Issues in the Curriculum, 87th

Yearbook of the National Society for the Study of Education, I, (pp.83-121). Chicago,IL: National Society for the Study of Education.

Appendix

Members of the Assessment of Mathematics Projects Research Group

R.L. Abrines	School of Mathematics, Kingston University.
J.S. Berry	Department of Mathematics and Statistics, Plymouth University.
R. Crouch	Division of Computer Science, University of Hertfordshire.
A. Davies	Division of Mathematics, University of Hertfordshire.
S. Dunthorn	formerly Head of Mathematics, Sheppey School, Kent.
E. Forrest	School of Computer and Mathematical Sciences, The Robert Gordon University, Aberdeen.
G.W. Goodall	Department of Mathematics and Statistics, Brunel University.
C.R. Haines	Department of Mathematics, City University, London.
K.E. Hirst	Faculty of Mathematical Studies, Southampton University.
S.K. Houston	Department of Mathematics, University of Ulster.
P.C. Hudson	School of Computing and Mathematics, Teesside University.
J.F. Izard	Australian Council for Educational Research.
A. Kitchen	Department of Education, Manchester University.
D. Le Masurier	Department of Mathematical Sciences, Brighton University.
S. Prestage	School of Education, Birmingham University.
R. Summers	Department of Systems Science, City University, London.
J. Vine	Head of Mathematics, Eggbuckland Community College,

Plymouth.

Table 1 Regression beta weights for categories with traditional aggregate (Izard, 1990)

N=52 (traditional aggregate) (Rsq=0.97)	Beta weight	Actual Weight	Intended Weight
Summary or abstract	0.22	2	1
Problem statement/ introduction	0.11	1	2
Content	0.59	6	6
References/ data sources	0.15	1.5	1
Presentation	0.14	1.5	1

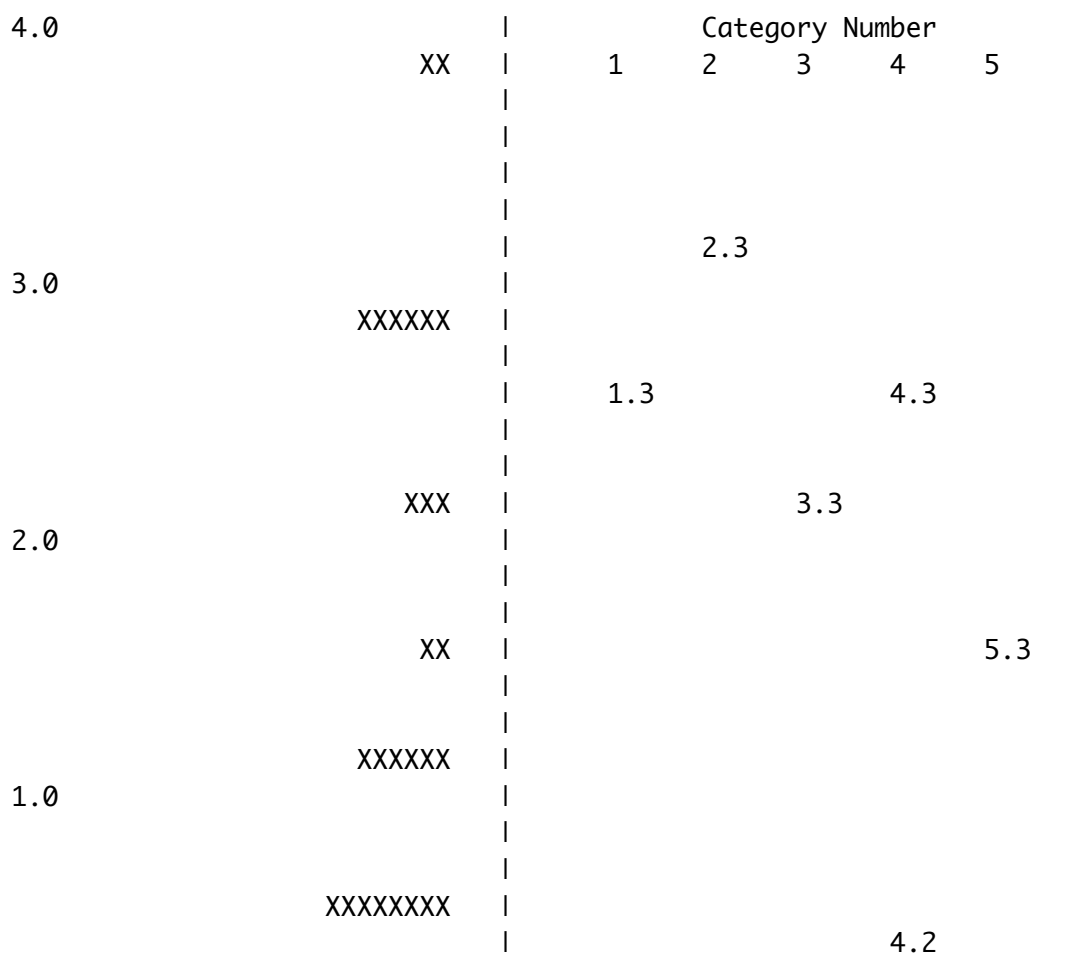
Table 2 Initial ranking of specimen projects and ranking following Facet analysis (Berry & Haines, 1991)

	Source 1		Group 1	Group 2	FACET rank
1	History	UG	6	5	2
2	Modelling 1	UG	3=	?	5

3	Modelling 2	UG	1	2	3=
4	Modelling 3	Dip	8	7	
5	Complex analysis	UG	?	?	
6	Investigation 1	GCSE	3=	3	1
7	Investigation 2	GCSE	3=	1	3=
8	Investigation 3	UG	7	4	
9	Investigation 4	UG	2	6	6

---

1 UG Undergraduate; Dip Diploma; GCSE Secondary



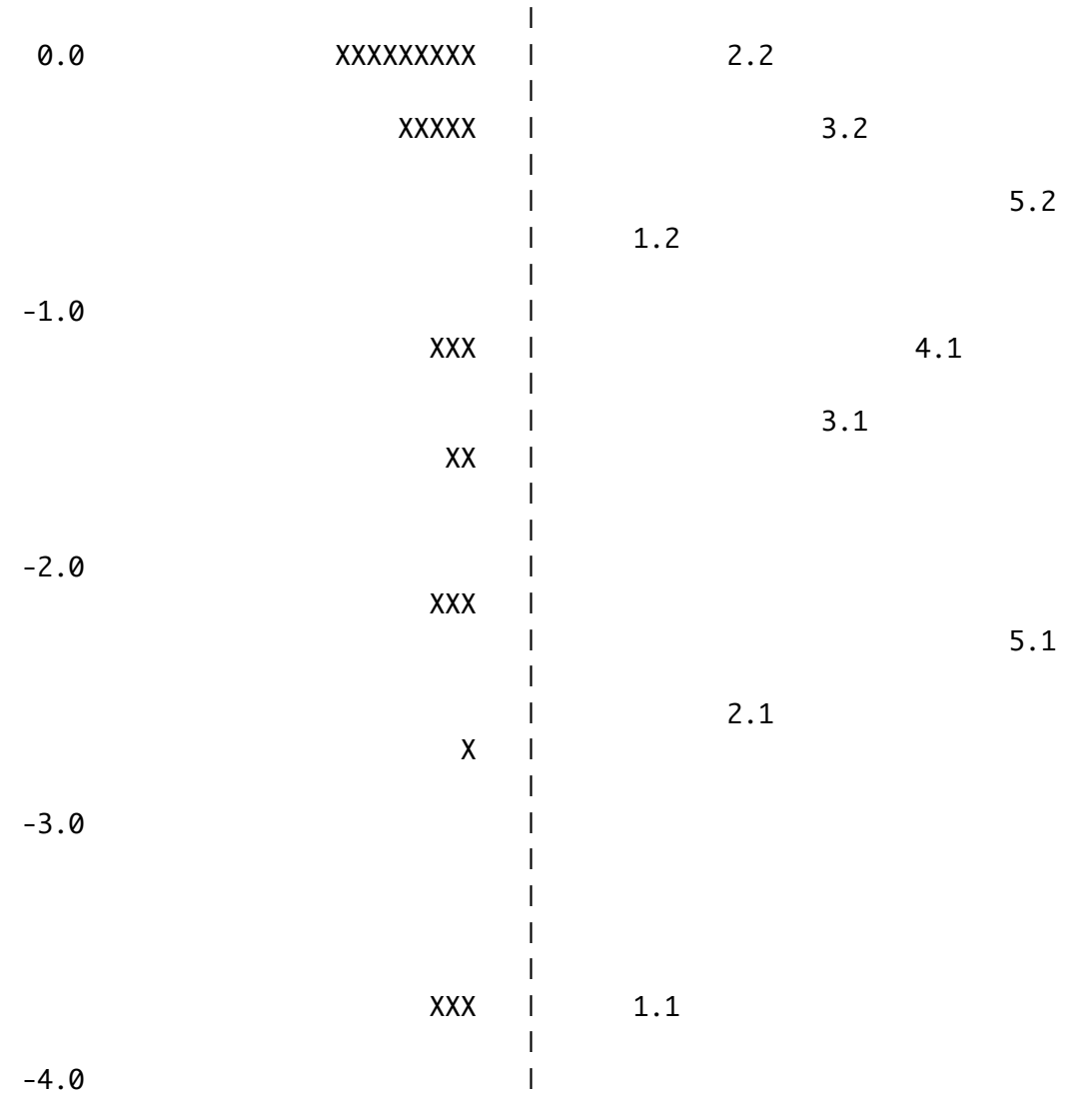
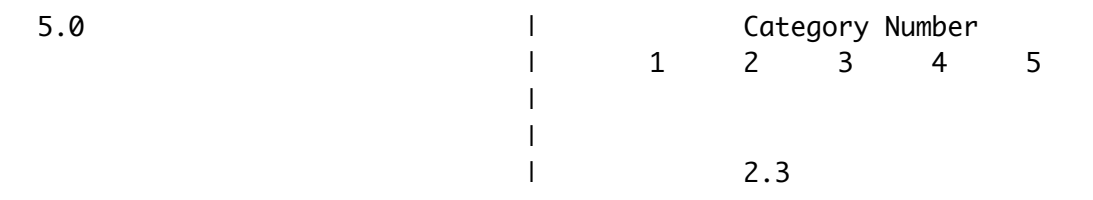
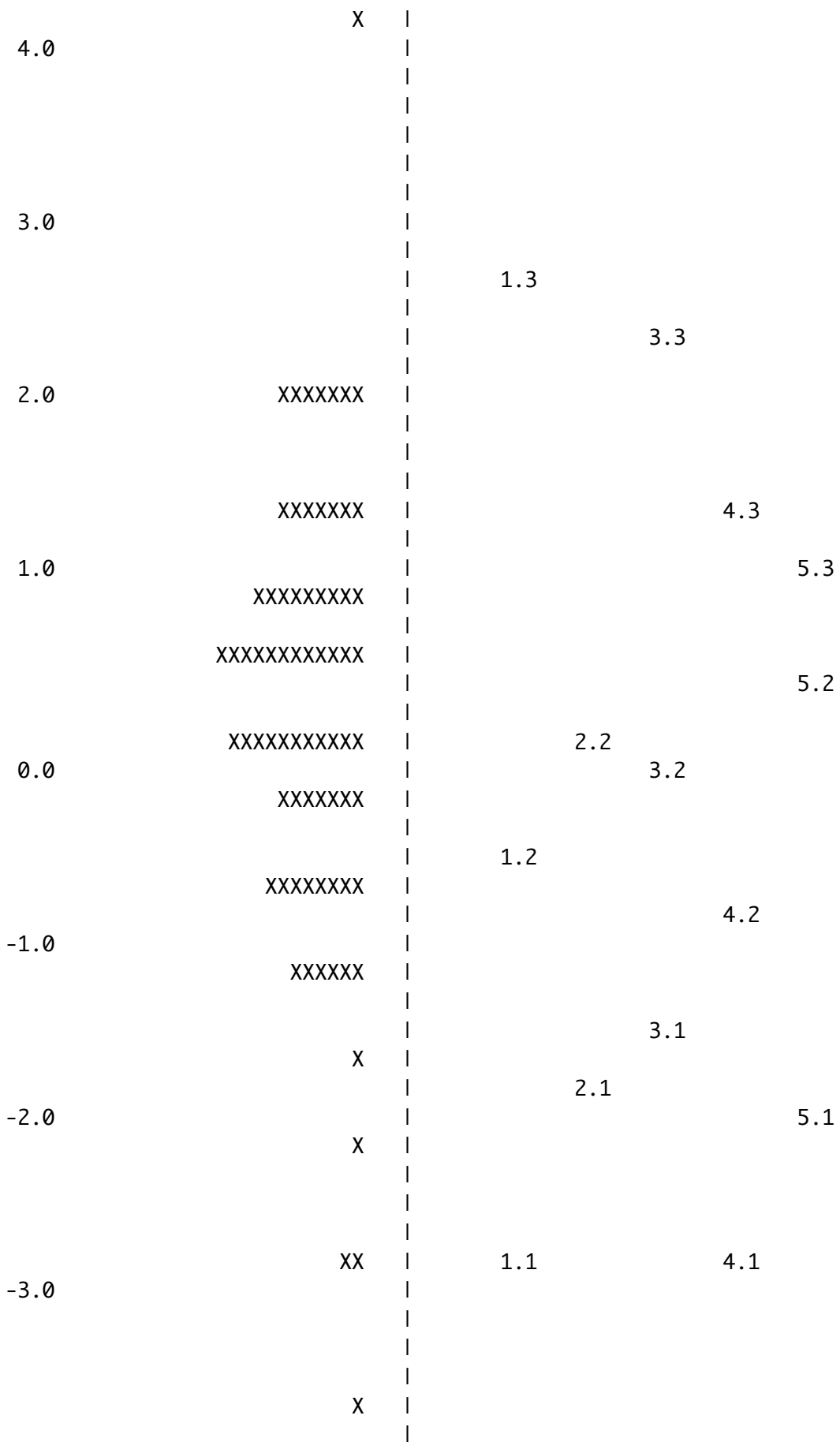


Figure 1 Student Achievement Estimates and Item Estimates (Thresholds) for 1986-88 Projects (N = 53, L = 5). (Each X represents 1 student.)





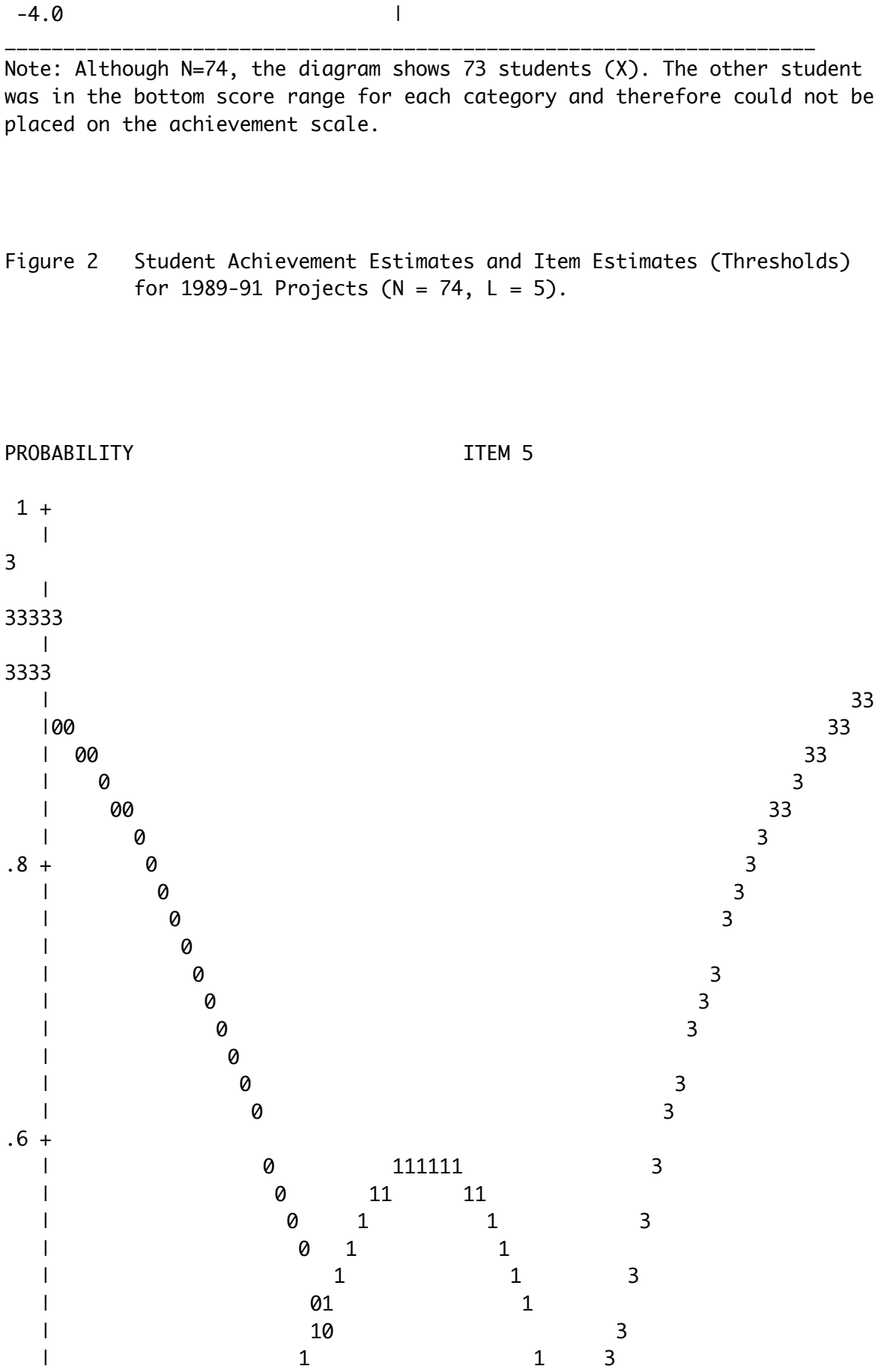


Figure 2 Student Achievement Estimates and Item Estimates (Thresholds) for 1989-91 Projects (N = 74, L = 5).

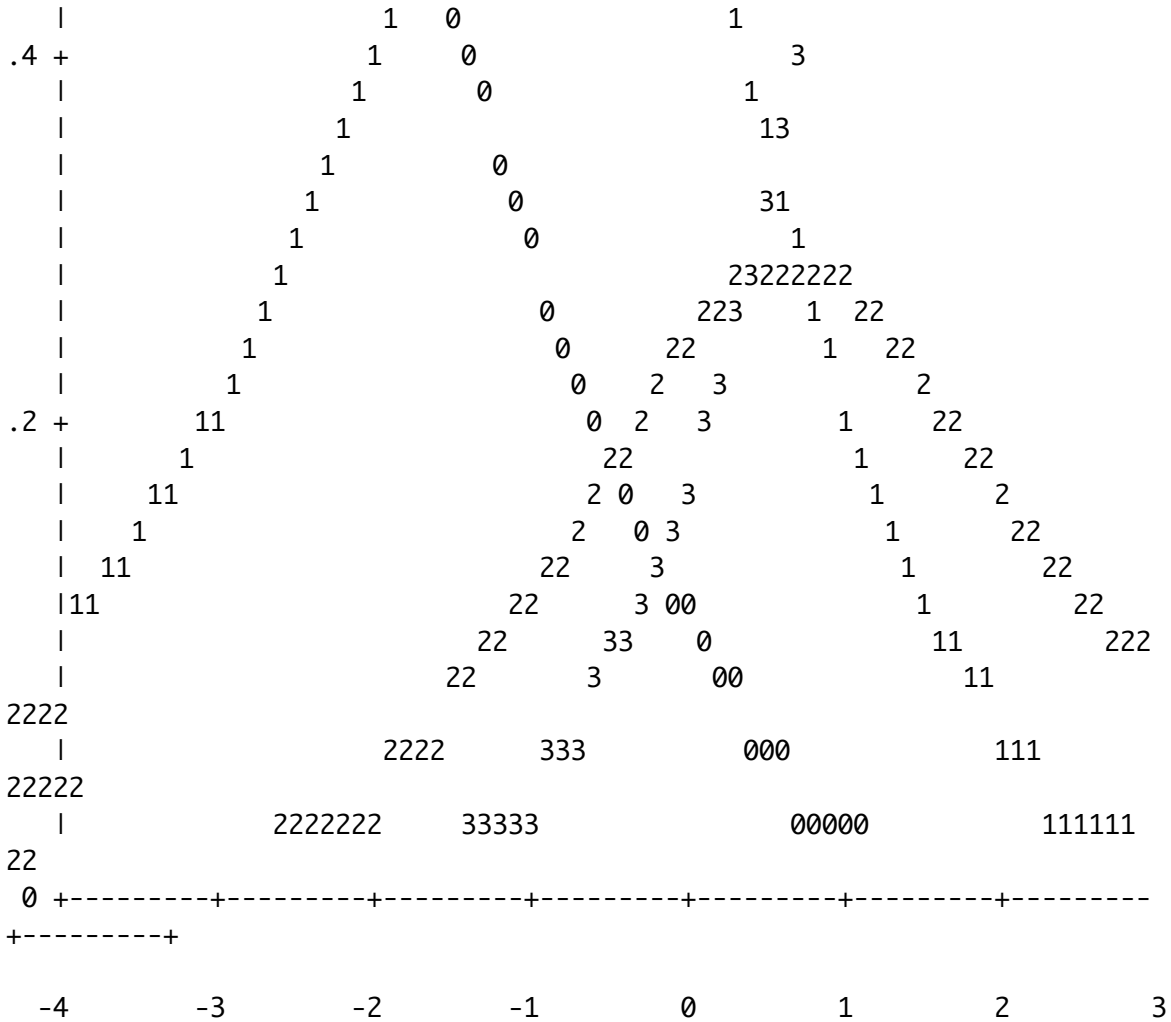


Figure 3 Probability distributions for the modelled assessment scheme for the fifth category (Presentation) 1989-91.

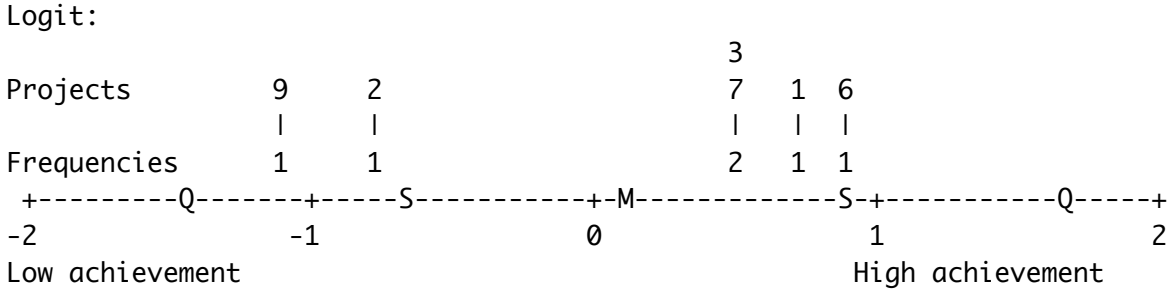


Figure 4 Projects ranked on an achievement scale.

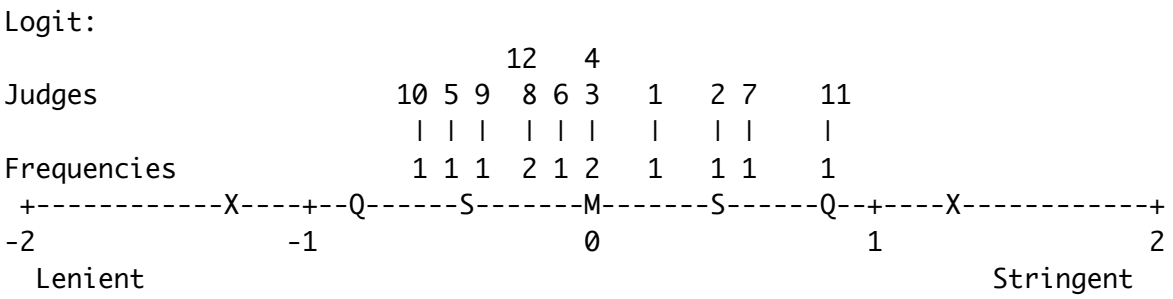
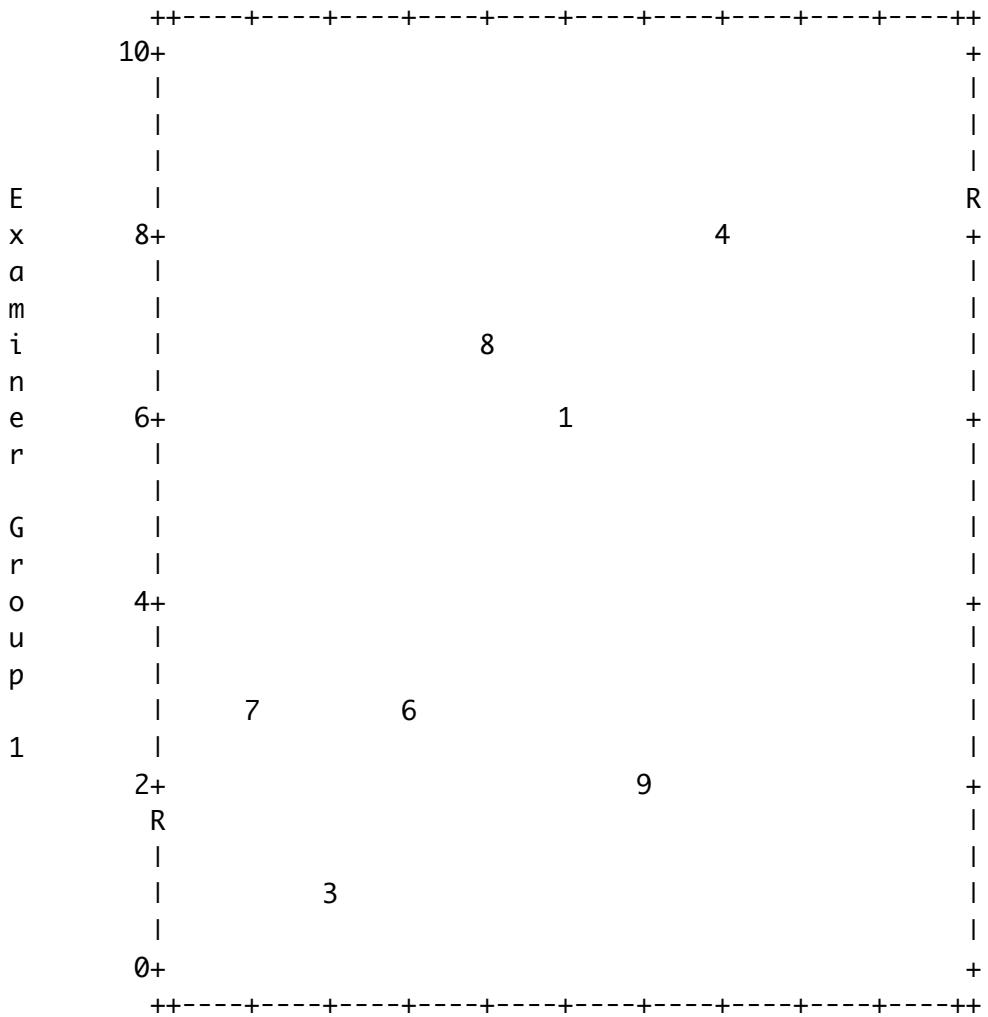
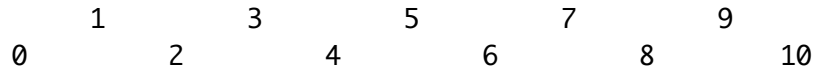


Figure 5 The judges shown on leniency-stringency scale.





Examiner Group 2

Figure 6 Assessment of written reports by two groups of examiners. Project number plotted.

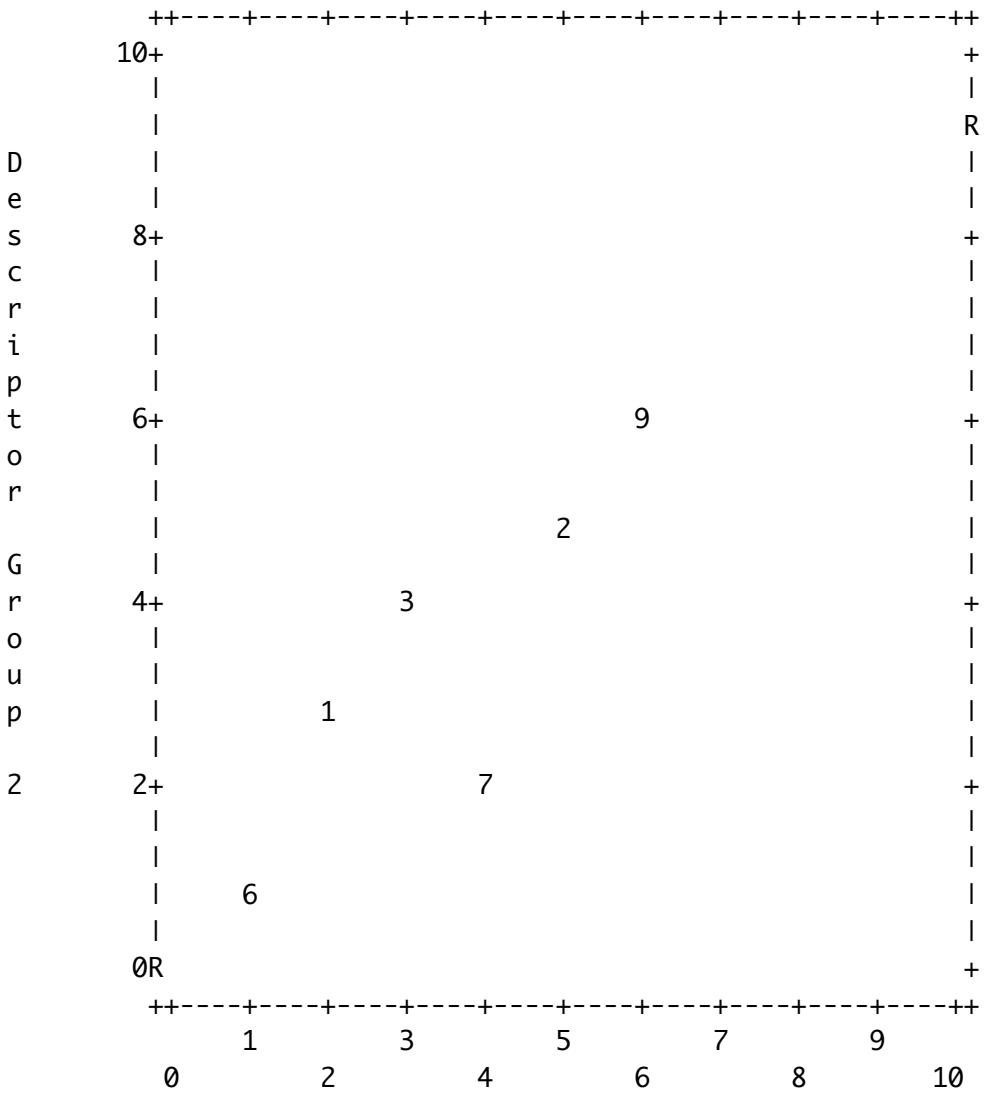


Figure 7 Rank assessment of written reports from two sets of descriptors. Project number plotted.

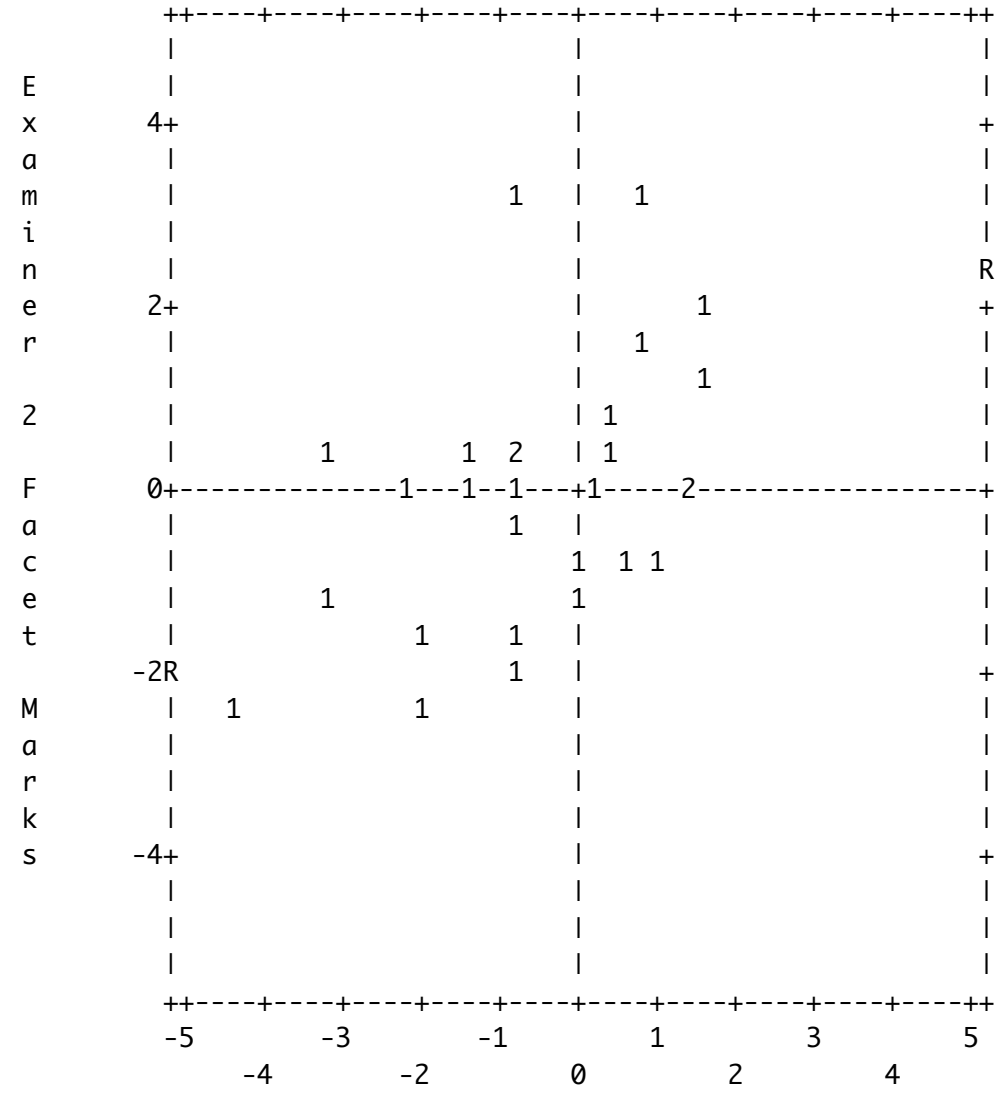
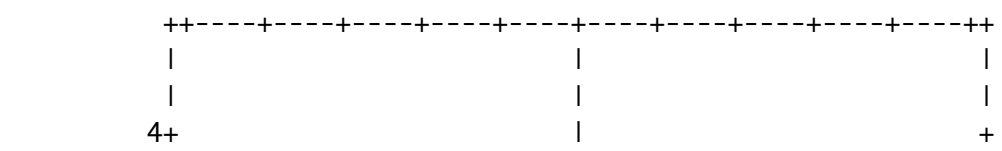


Figure 8 Assessment of individual oral reports: judge 2 compared with judge 1. (Correlation 0.46).



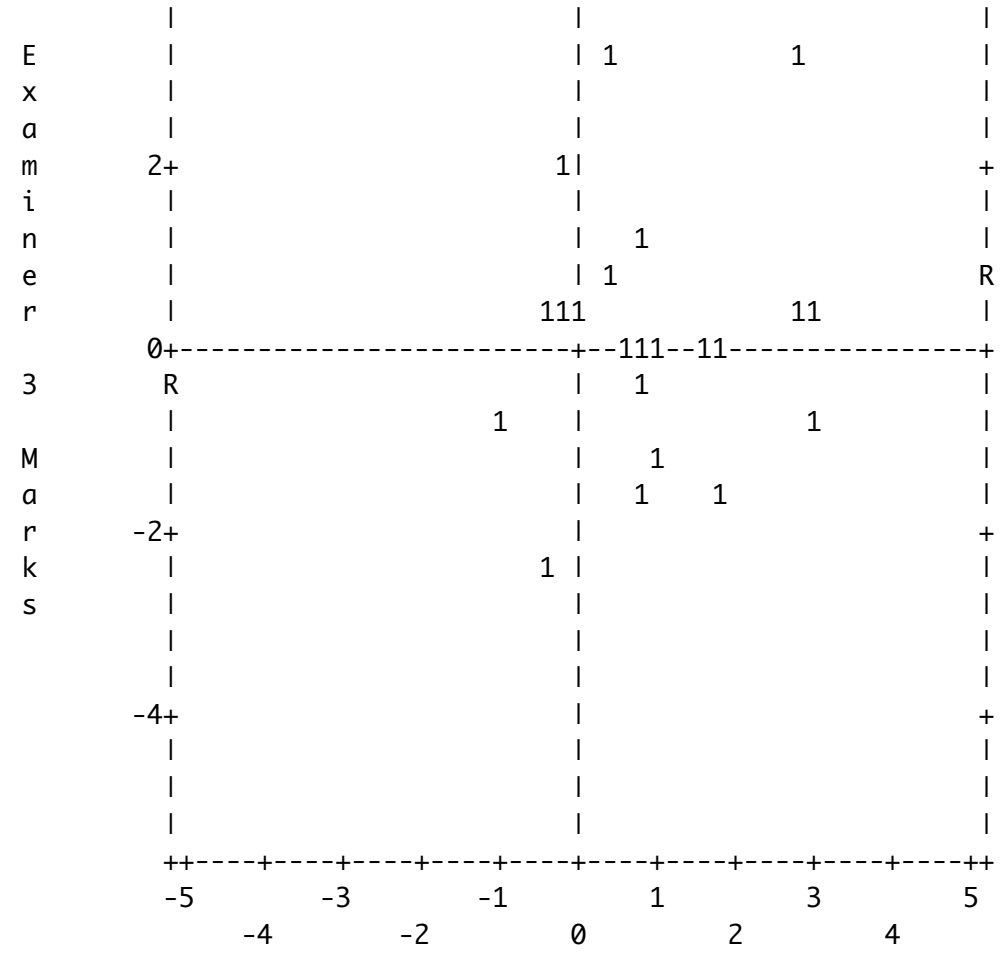
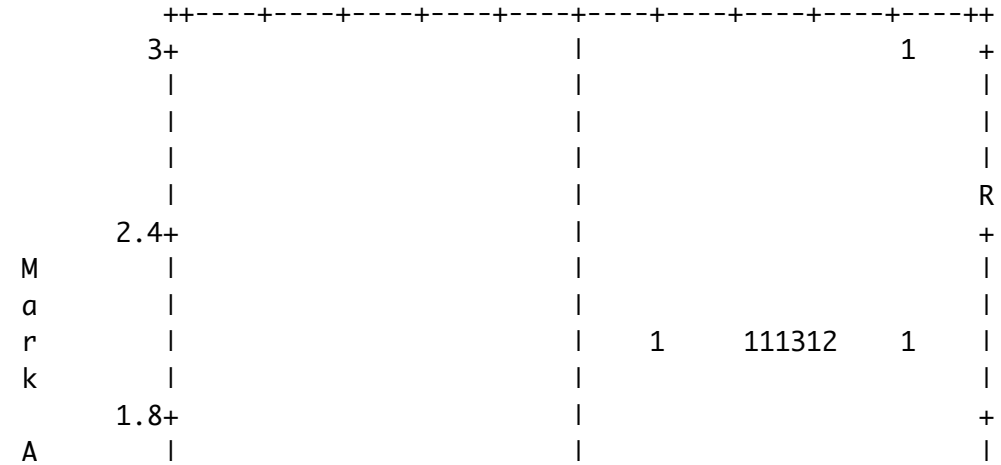


Figure 9 Assessment of individual oral reports: judge 1 compared with judge 3. (Correlation 0.11).



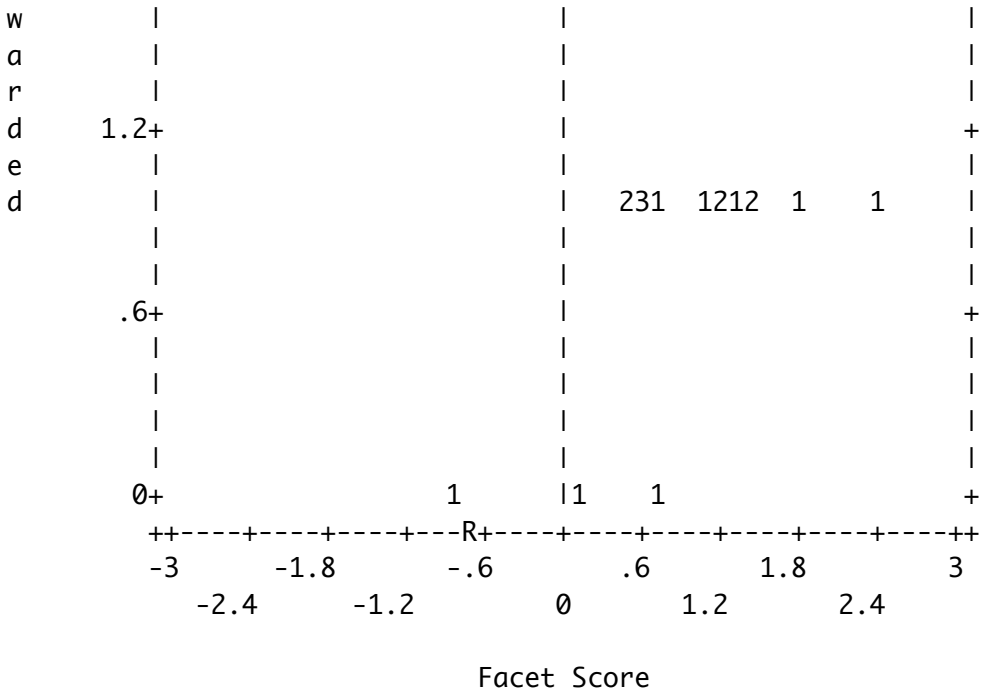


Figure 10 Individual student performance in oral presentations: Comparison of the grades awarded with those suggested by FACET analysis.