Extracting individual grades from group assessments; 
a pilot study.

Sverker Johansson
School of Education and Communication
University of Jönköping
Abstract
A novel method for extracting individual grades from group assessment scores is presented, together with results from a pilot study where the method has been tested for the first time in a real assessment context. The central feature of the method is that students do all the work together in groups, but switch groups multiple times during the assessment, so that each student participates in several different group configurations. This switching makes the contribution from each student to the group results mathematically extractable. The method is intended to be applicable to all assessments where numerical scores are given.

The pilot study was largely successful, resulting in usable grades. Some minor adjustments and tuning, based on the pilot study results, will be implemented before proceeding to full-scale testing of the assessment method.

Sammanfattning
En ny metod för att få fram individuella betyg från gruptentamensresultat presenteras, tillsammans med resultat från en förstudie där metoden för första gången har testats i praktiken, i en verklig examination. Grundtanken i metoden är att studenterna tenterar i grupp, men byter grupper flera gånger under tentamen, så att varje student deltar i ett flertal olika gruppkonstellationer. Dessa gruppbyten gör individuella bidrag till gruppresultatet matematiskt åtkomliga. Metoden avses kunna tillämpas på all poängsatt examination.

Förstudien var i det stora hela lyckad, såtillvida att den gick att genomföra, och resulterade i användbara betyg. Vissa småre ändringar och justeringar, baserade på förstudiesresultaten, kommer att införas innan examinationsmetoden testas i full skala.
Introduction

Traditionally, universities emphasise individual learning, and students are normally assessed as individuals. University degrees are invariably granted to individuals. At the same time, there is a rising interest in the advantages of co-operative learning. The advantages are manifold, but can be grouped in two main categories:

- Co-operation improves individual learning (McKeachie, 1994; Ramsden, 1992; Webb, 1994).
- Co-operation is a useful skill in itself, particularly in view of the highly team- and project-oriented future workplaces of many students (Ramsden, 1992; Lejk et al 1996).

Co-operative learning does not, however, necessarily imply co-operative assessment, where students are assessed in groups. Assessment in groups is significantly rarer than learning in groups (Lejk et al, 1997), and is actively discouraged at some universities: “Group assessment is not generally approved…” (Victoria University of Wellington 1997, p 1). Assessment in groups has many of the same advantages as co-operative learning, and it would be a natural part of many forms of co-operative learning, e.g. Problem-Based Learning. Much of what is called “performance” (Wangsatorntanakhum, 1997) or “authentic” (Wiggins, 1990) assessment naturally takes place in groups. Furthermore, it can significantly reduce the exam anxiety that plagues many students in traditional assessment contexts (Box & Barrett, 1989), and there is some evidence that performance is improved (LeCount & Fox, 1992; Webb 1993).

The main reason that group assessment is not used more than it is, may well be the difficulty in deriving fair individual grades from the group work. In general overviews of assessment methods (McKeachie, 1994; Habeshaw et al, 1993; Jacobs & Chase, 1992), three principal grading methods are presented:

- Collective grading, where all students in a group receive the same grade.
- Individual reports from the group work.
- Peer assessment, where the students assess each other’s contributions to the group.

The collective grading method is perceived by many, teachers and students alike, to be unfair, because different students vary widely in both ability, effort and commitment to the group work. Teachers feel a persistent worry (Lejk et al 1997) that weak students may “hide” and be “carried” by their more able or less lazy companions, though Murray (1990) argues that this never happens¹. Some collective grading systems allow for teacher intervention in flagrant cases of abuse.

Group work with individual reports is a compromise that, in my opinion, combine many of the drawbacks of both traditional and group assessment; the risk that students “hide” remains, but the teamwork is at least partially lost.

¹ Unfortunately I can testify from personal experience (both as a student and as a teacher) that it does. As a student, I have “carried” others (this being less effort than making an issue of it), and as a teacher I both receive perennial complaints that "he didn’t pull his weight”, and have received occasional admissions from students who were carried.
Peer assessment is in fairly wide use (Lejk et al, 1996), has received much favourable attention (Habeshaw et al, 1993; Lejk et al, 1996) and can show some favourable results (Schechtman & Godfried, 1993). Nevertheless, I am highly uncomfortable with the concept; having the students obliged to report on each other, cannot but hurt the team spirit and solidarity that are at the heart of the co-operative learning experience. The students have an incentive to impress each other and compete with each other, rather than to optimise team success.

None of the methods for deriving individual grades extant in the literature appear satisfactory. Lejk et al (1996) concludes their survey with “Group assessment is an important issue and will become more so with the increased emphasis upon group working which is being adopted by several institutions. The questions raised above illustrate the need for more research in this area.” (p 276). The need for a better method is the starting point for this work.

**The evolution of an assessment idea**

I have been using various forms of group assessments in my courses for several years, with results that have been highly satisfactory from the co-operative-learning perspective, but quite unsatisfactory for grading purposes. Individual take-home exams, which amounts to de facto individually reported group work, just led to a lot of duplicated labour, for both students and teacher, and didn’t measure what I wanted measured². Experiments with “real” group assessment, collectively graded, in various settings, likewise led to grave doubts about the validity and fairness of the grades. It became more and more obvious to me that I was flunking the wrong people.

So I ended up contemplating whether to give up group assessments altogether, or if there could possibly be any acceptable way to make them work. I was reluctant to return to an exclusive diet of traditional individual exams, which have numerous drawbacks (implicit in the previous section), in that they do not foster co-operation, are not effective learning experiences in themselves, and for many students are a major source of stress and anxiety.

The fundamental cause with the general dissatisfaction with available group assessment methods appears to be the difficulty of disentangling the individual contributions to a group. In my contemplations, I spent a fair amount of time considering ways around this difficulty. Being a physicist, my thoughts naturally went in the direction of mathematical methods for modelling and disentangling complex systems. The obvious way in physics would be to isolate the interesting variable, and study it on its own; in an assessment context the interesting variable is the ability of the individual student, so isolating it translates as traditional individual assessment, but that’s not what we’re after here. Another way would be to keep all variables but one fixed, varying just the one, under controlled conditions. Translation to an assessment context: repeated group assessments, with all group members but one kept the same – replacing that one with another student gives a handle on the contributions of both. Could be done, but would have to be repeated for all combinations of students; too cumbersome for practical use.

Nevertheless, here was the seed of an idea. Changing the group settings, letting students work in different combinations, could give clues to their individual contributions. Given a mathematical model for how the abilities of individual group members add up to a group score, even a modest number of group permutations should make it possible to disentangle all

---

² As one (female) student expressed it: “This exam doesn’t measure how much physics you understand; it measures whether your boyfriend studies at the engineering school.”
individual contributions in one go. The mathematical model was the key; it had to be a reasonable approximation of reality, and at the same time it had to be computationally tractable for the rather complex problem of solving a large system of equations, with one variable for each student and one equation for each group.

I spent a considerable amount of time, over the course of a year, in trying out different models and disentangling algorithms in computer simulations. The conclusion from that exercise was that the results are not all that sensitive to the details of the models; a simple general model like the one presented in the next section would handle all reasonable cases. Likewise, a small number of groups was adequate; each student need participate in no more than four groups, in order to receive a reasonably reliable grade.

So, there we had the assessment idea, all worked out in theory and simulation: the students do four separate assignments in groups, each time in different groups. Each group receives a single, collective score, and from the pattern of collective scores individual grades for the group members are mathematically extracted.

This assessment method is intended to work in all assessments where numerical scores are given, which is the vast majority of traditional written exams. It is internally mathematics-based, but this should not be taken to mean that it is limited only to mathematics-based subjects. In fact, the pilot study uses qualitative non-numerical questions.

I was planning to do a systematic experiment, testing this assessment design in reality, as opposed to simulation, with a full set of experimental controls, and individual tests as a baseline, when the opportunity to perform a smaller pilot study serendipitously turned up. The pilot study had to be done without proper controls, but I nevertheless felt it valuable to pre-test the basic machinery of the assessment procedure.

The fundamental purpose of this pilot study is to determine whether this method of assessment is at all workable in practice. Questions I want answered are:

- What practical problems are involved in administering this rather complex exam?
- Are individual results extractable at all?
- Do the grades make sense, and do the students accept them?
- How do the results compare with simulations? Does reality conform to the model?
- If it basically works, what improvements are needed for full-scale use?

**Method**

The current study was performed in a real assessment context, with students who received real, valid grades from the experimental assessment process. This has obvious advantages (the students are readily available, and are highly motivated to do their best), but would have been ethically dubious, if it weren’t for the fact that the student group voluntarily accepted (and in many cases demanded!) to be part of the experiment.

Two parts of the same course were assessed separately, one astronomy module, and one meteorology module, henceforth called “Astro” and “Weather”. This double assessment was done in two sessions, consecutively during the same evening.
Student population

The population who took part in the study consisted of second-year students in a teacher-training programme aimed at the first seven years of the Swedish primary school. Approximately half of the students were specialising in natural sciences, the other half in social sciences.

The experimental assessment was used in the students’ third attempt to pass the Astro and Weather exams, having failed in two previous attempts. These exams are considered fairly difficult, with a sizeable proportion (up to 50%) of the student body regularly failing at least once. The students trying for the third time are thus selected from the lower part of the curve, but not exclusively from the extreme lower tail; the population is somewhat biased towards mediocre performers, but not heavily so.

The total number of participating students was 47, of whom 42 did Astro and 22 did Weather (17 students did both). These 47 comprise about 40% of the approximately 120 students who participated in the original course (the other 60% having passed one of the previous exams).

After the students’ second failed attempt, I had informal discussions with a number of them, where they expressed a strong desire for a different assessment procedure the third time around (the first two had been with a hybrid multiple-choice/essay exam; see (Johansson 1998) for a description). In order to accommodate the students’ wishes as far as possible, I invited all the students involved to a meeting to discuss assessment procedures. At the meeting, I presented the students with several alternative procedures, including this experimental one, that I was willing to consider. I hadn’t really intended to do the experiment so early — I had it planned for a later date, in a different course, under more controlled conditions — but the students nevertheless selected that option. About half of the students were present; none of those dissented vocally. Of the students absent from the meeting, several have expressed their support, and none has protested the decision. One absentee decided, however, for undisclosed reasons, not to participate in the experiment.

Assessment design

The exam consists of a number of problems that the students are supposed to solve, working in groups of three, and is conducted in four rounds. In each round, the groups solve a single problem, after which the groups’ solutions are collected by the teacher, the groups are dissolved, and new groups are formed for the next round. Each student meets an entirely new set of group peers in each round, never working the same student in more than one round.

The problems are intended to test higher levels of understanding, not simple recollection of facts, and are thus suitable for fruitful group discussions. The emphasis is on reasoning skills, not on either memorisation or numerical calculations. Sample problems are given in Appendix A (translated into English; the exam was given in Swedish).

In order to make the grade extraction procedure numerically tractable, the students are divided into macro-groups of 9 to 12 students. Students are thoroughly mixed within macro-groups, but not between macro-groups, thus breaking down the extraction problem into several smaller calculations. The total number of students in the Weather exam is not an integer multiple of three, so one macro-group had one supernumerary student; in this group, one student had to

---

3 This condition was violated in one case, through a simple book-keeping error on my part. The consequences of that error are discussed further below.
sit out each round.

**Grade extraction**

In the grading procedure, we wish to obtain an estimate of each student’s individual ability \(^4\), which we call \(x_i\), on the basis of which the student is assigned a grade (in this case just Pass or Fail, but the method is trivially generalizable), so that if \(x_i > x_P\) then the student receives a Pass grade. It is assumed here that "ability" is quantifiable, but this is tacitly assumed in all assessments where grades are given; this method does not differ from standard exams in this respect.

The extraction of individual grades is based on a very simple model of group work. It is assumed that the group score is proportional to a linear combination of some function of the individual abilities, plus a noise term:

\[
s = a \left[ \sum_{i=1}^{n} f(x_i) \right] + A
\]

where \(a\) is a constant, relating the (arbitrary) scale of \(f\) with the scoring range; in the actual analysis, \(a\) is subsumed into \(f\). \(A\) is a stochastic noise term, with some reasonable amplitude and distribution.

The function \(f\) is quite arbitrary, except that it is assumed to be monotonically increasing throughout the relevant interval. It need not be (and likely isn’t, in reality) linear. The monotonicity assumption means that \(f(x)\) is adequate for grading purposes: \(x_i > x_P\) implies that \(f(x_i) > f(x_P)\). Thus there is no need to actually reconstruct \(x\) itself, nor any need to assume any specific explicit form for the function. The function is assumed to be the same for all students.

An additive model is used, because it appears more reasonable than a multiplicative one (the presence of a zero-ability student doesn’t kill a group in reality, whereas it would in a multiplicative model), and because more complex mathematical models are less computationally tractable without being obviously more realistic.

All complications arising from group interactions and other processes that do not add linearly are for this purpose regarded as “noise”, and included in \(A\). In this sense, the model is clearly incomplete, but for this purpose it may still be adequate, assuming that linear or near-linear effects dominate.

The model assumes, furthermore, that the group results scale linearly with group size, which is not a realistic assumption; on a problem-solving task (where division of labour isn’t practical) I find it highly unlikely that e.g. a six-member group should get twice the score of a three-member group. Group size must therefore be held constant; all work is done in groups of three students. In the Weather exam, the number of students wasn’t a multiple of three; nevertheless groups of three were used, and one student had to sit out each round.

For the extraction procedure, the solutions that the groups have given to the exam problems

---

\(^4\) “Ability” is for this purpose operationally defined as simply that upon which we wish to base our grading, whatever it may be. It normally includes knowledge of the subject matter, understanding of and ability to use that knowledge, and may well here include social talents that facilitate group work.
are teacher-scored in the usual fashion, on a scale from 0 to 10, with the score determined by
my judgement of the level of understanding displayed by the students.

The ability \( x \) (or rather \( f(x) \)) for each student is then derived through a minimization procedure. The \( \chi^2 \)-function:

\[
\chi^2 = \sum_{k=rounds} \sum_{j=groups} \left( \frac{s_{jk} - \sum f(x_i)}{\sigma_S} \right)^2
\]

is minimized with respect to the free parameters \( f(x_i) \), one parameter for each student. The \( s_{jk} \)-values are the actual scores received by group \( j \) in round \( k \). The sum over students runs only over the members of group \( j \) in round \( k \). The \( \sigma_S \) in the denominator is the standard deviation of \( s_{jk} \), which unfortunately is unknown. For the minimization, though, the absolute value of \( \sigma_S \) doesn’t affect the result; it is set to 1. An estimate of the real value of \( \sigma_S \) can be obtained \( \textit{a posteriori} \) through comparison with simulation, where \( \sigma_S \) is known, and equal to the RMS of the noise term \( A \).

The set of \( f(x_i) \)-values which minimizes \( \chi^2 \) is taken as an estimate of the students’ individual abilities, and is used for grading. The actual minimization is performed with the standard function minimization package MINUIT (James 1994), developed at CERN (Conseil Européenne pour la Recherche Nucléaire) for applications in particle physics (for a more typical application than this one, see ch 11 in (Johansson 1990)). MINUIT uses a sophisticated minimization algorithm to estimate, with confidence limits, the best \( f(x_i) \)-values

The individual results \( f(x_i) \) are restricted to positive values; without this restriction a few students (e.g. student 8 in Table 2) would have obtained negative values, which can be interpreted as their presence in a group doing more harm than good. In rare cases (students who actively disrupt group work) this may be true, but in my judgement this is not the case here. No upper limits on \( f(x_i) \) were imposed.

In order to actually assign grades, the value of \( f(x_p) \) is needed. A normal grading practice on problem-solving exams is to set the pass limit at 50% of the maximum score. This practice is translated into the group context here in the following manner: \( 3f(x_p) = 5 \) (5 being 50% of the max score), meaning that a group composed of three members who are all exactly on the pass limit individually should achieve 50% of the maximum score.

Each student’s \( f(x_i) \)-value is finally translated into a grade (Pass/Fail), and into a score on a percentage scale: \( S\% = 50* f(x_i)/ f(x_p) \% \), with a score of 50% meaning a barely passing grade. When assigning grades, I use the confidence intervals on the reconstructed scores that are calculated by the reconstruction program; in order to give a Fail grade, the student’s score must be \textit{significantly} (84% c.l.) below 50%. The system is thus biased in favour of the students, which I think is only fair when using an experimental procedure.

\textbf{Computer simulations}

In order to test and evaluate the reconstruction process, the whole group procedure is repeatedly simulated, with virtual students who conform to the model described above. In the
simulation process, each virtual student is randomly assigned a value for \( f(x_i) \) (flat distribution of \( S\% \) between 0 and 100%). The score for each group in each round is then calculated according to equation (1), with a Gaussian noise term \( A \in \mathbb{N}(0,\sigma) \), with adjustable width \( \sigma \) (expressed in the following as a percentage of the maximum score). The output from the simulation is then used as input to the reconstruction program, and virtual student grades are reconstructed in accordance with the procedure described above. The reconstruction results can then be compared with the “true” \( f(x_i) \)-values from the simulation, giving an estimate of the reconstruction quality.

### Results

The results are presented in two sections; first, the actual results from the assessment and reconstruction, and then comparison with simulation results.

#### Reconstruction results

The rather complex practical arrangements around the exam, with group-switching and multiple solution-collection, worked out fairly well, with no major glitches. All groups, despite their ephemeral character, functioned well enough to produce a solution (if not always correct) to the problem at hand. Likewise, the scoring procedure caused no special problems; I did all the scoring myself, and observed nothing really remarkable. The scoring went quite fast; the total volume of group output was noticeably less than that expected from a similar individual exam.

The raw scores were then fed into the reconstruction program, one macro-group at a time. One obstacle that turned up here, was that the problems used in the different rounds were not of equal difficulty, which is implicitly assumed in the reconstruction program; a correction had to be added to the scores, equalising the average score received by all groups on the same problem. After that correction, the program ran smoothly, with no numerical difficulties, and produced stable results, not sensitive to free-parameter initial values or minor perturbations.

One exception, though, was a macro-group where a grouping error had occurred, with one student working together twice with each of two others. The program had marginally sufficient data on the separate abilities of these three, resulting in a sizeable covariance between their results. Perturbing the system caused substantial redistribution of percentage scores between them, and some perturbation also of other students’ scores in the macro-group. The problem was exacerbated by the fact that this was the macro-group with a supernumerary student, so that some students did only three rounds. The grades of two of the three students involved, as well as that of one other in the group, were unstable, changing across the Pass limit when perturbed; in fairness to the students, these three received a Pass grade.
Table 1: Group scores from macro group Astro 1.

<table>
<thead>
<tr>
<th>Round</th>
<th>Group</th>
<th>Actual score</th>
<th>Reconstructed score</th>
<th>Group member 1</th>
<th>Group member 2</th>
<th>Group member 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>7.0</td>
<td>6.7</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5.0</td>
<td>4.8</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>5.0</td>
<td>5.8</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4.0</td>
<td>3.7</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>5.0</td>
<td>4.8</td>
<td>1</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4.0</td>
<td>3.7</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10.0</td>
<td>9.7</td>
<td>3</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5.0</td>
<td>5.7</td>
<td>2</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>6.0</td>
<td>5.7</td>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.0</td>
<td>2.8</td>
<td>2</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.0</td>
<td>1.7</td>
<td>3</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5.0</td>
<td>4.7</td>
<td>6</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>8.0</td>
<td>8.7</td>
<td>1</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6.0</td>
<td>5.8</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.0</td>
<td>6.7</td>
<td>3</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.0</td>
<td>6.8</td>
<td>6</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>
Table 2: Individual results reconstructed in macro-group Astro 1.

<table>
<thead>
<tr>
<th>Student</th>
<th>Rec-score</th>
<th>Lower limit</th>
<th>%-score</th>
<th>Upper limit</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.25</td>
<td>51%</td>
<td>68%</td>
<td>84%</td>
<td>P</td>
</tr>
<tr>
<td>2</td>
<td>0.38</td>
<td>11%</td>
<td>21%</td>
<td>30%</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>0.89</td>
<td>8%</td>
<td>27%</td>
<td>45%</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>3.59</td>
<td>91%</td>
<td>96%</td>
<td>100%</td>
<td>P</td>
</tr>
<tr>
<td>5</td>
<td>3.63</td>
<td>91%</td>
<td>96%</td>
<td>100%</td>
<td>P</td>
</tr>
<tr>
<td>6</td>
<td>1.95</td>
<td>40%</td>
<td>59%</td>
<td>77%</td>
<td>P</td>
</tr>
<tr>
<td>7</td>
<td>3.23</td>
<td>79%</td>
<td>90%</td>
<td>100%</td>
<td>P</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0%</td>
<td>2%</td>
<td>4%</td>
<td>F</td>
</tr>
<tr>
<td>9</td>
<td>2.92</td>
<td>71%</td>
<td>86%</td>
<td>100%</td>
<td>P</td>
</tr>
<tr>
<td>10</td>
<td>2.53</td>
<td>58%</td>
<td>76%</td>
<td>94%</td>
<td>P</td>
</tr>
<tr>
<td>11</td>
<td>0.78</td>
<td>5%</td>
<td>24%</td>
<td>42%</td>
<td>F</td>
</tr>
<tr>
<td>12</td>
<td>0.92</td>
<td>11%</td>
<td>28%</td>
<td>44%</td>
<td>F</td>
</tr>
</tbody>
</table>

As an example, the results from one macro-group (not the problematic one) is shown in Table 1 (group scores; real, and predicted\(^5\) from the reconstruction) and Table 2 (individual results, anonymized, with confidence limits (1 \(\sigma\)) on the percentage scores). In Table 1, the third column contains the actual scores awarded to the groups, and the fourth column the scores “predicted” from the reconstructed abilities of the group members, as given in column 5-7 (the numbers in column 5-7 correspond to the student numbers in Table 2).

---

\(^5\) “Predicted” means the group score calculated from equation (1), using the reconstructed student scores, assuming no noise.
The total distribution of individual scores, from all the macro-groups, is shown in figure 1. Figures 2 and 3 display the relation between the real and the predicted score for each group, which is an indirect measure of how closely the actual students conform to the model described in the previous section. Figure 2 is a direct comparison of the two, showing a clear linear relationship, with a modest spread. Figure 3 is a histogram of the residuals corresponding to the spread around the straight line in figure 2, which will be used in the comparison with simulated data. The RMS of this spread, for each macro-group separately, as well as the reconstruction $\chi^2$ is given in Table 3.

**Comparison with simulation**

When used on simulated students, the reconstruction program performs tolerably well, as long as the “noise” level in the simulation is set at a moderate level. With no noise, the reconstruction is of course perfect; with increasing noise the performance deteriorates, as shown in figures 4-7. Figures 4 and 5 show quantities that are available also for real data: reconstruction $\chi^2$ and difference between real and reconstructed group scores. A comparison between figures 4 and 5, and the corresponding values for real data, listed in table 3, gives an estimate of the effective noise level in the data.
Figure 2: Reconstructed versus actual scores for all groups.
Figure 3: Difference between real and reconstructed group scores, for all groups.

Figure 4: Average reconstruction $\chi^2$ as a function of noise level, for two different macro-group sizes.
For macro-groups of size 12, there is a fair bit of spread in the values listed in table 3, but taken as a whole they are consistent with a noise level of approximately 15%. The situation for 9-groups is a bit different; both 9-groups are consistent with there being no noise at all, which is highly unlikely. Possibly this is a statistical accident; the average simulated values are those shown in figures 4 and 5, but a fair fraction of simulated 9-groups do get a $\chi^2$ very close to zero, even with noise levels above 15%. Also, the number of degrees of freedom in the 9-groups is very small, making the reconstruction problematic, possibly resulting in artificially low values for $\chi^2$. In any case, there are grounds for some scepticism concerning the reliability of 9-groups. The one 10-group lies close to the 12-groups, but is problematic in other ways, as discussed in the previous section.

Figures 6 and 7 show the grading accuracy for simulated students, as a function of noise level. The grading procedure is identical to that used for real students, including the student-friendly bias described above. The grading accuracy with non-zero noise is in all cases in the vicinity of 80% correct grades, slowly worsening with increasing noise. Without bias, the percentage of correct grades would be somewhat higher, but half of the incorrect ones would be wrongful Fails. With bias, as can be seen in the figures, very few students undeservedly fail. Macro-groups of 9 students give marginally better performance than macro-groups of 12 students, but not enough to outweigh the suspicions raised in the previous paragraph.

<table>
<thead>
<tr>
<th>Macro-group</th>
<th>Group size</th>
<th>$\chi^2$/NDF</th>
<th>RMS of group score residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astro 1</td>
<td>12</td>
<td>0.75</td>
<td>0.43</td>
</tr>
<tr>
<td>Astro 2</td>
<td>12</td>
<td>0.33</td>
<td>0.29</td>
</tr>
<tr>
<td>Astro 3</td>
<td>9</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Astro 4</td>
<td>9</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Weather 1</td>
<td>10</td>
<td>0.81</td>
<td>0.37</td>
</tr>
<tr>
<td>Weather 2</td>
<td>12</td>
<td>1.51</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 3: Reconstruction $\chi^2$ and difference between real and reconstructed group scores for the six macro-groups in the real data set.
Figure 5: Group score reconstruction accuracy as a function of noise level, for two different macro-group sizes. The vertical axis measures the width (RMS) of the simulated distribution (average of 1000 runs) corresponding to that shown in figure 2.
Figure 6: Grading accuracy in simulated 9-groups, as a function of noise level.
Figure 7: Grading accuracy in simulated 12-groups, as a function of noise level.
Discussion

By and large, this pilot study worked out as planned. As an assessment it gave results that are perhaps not perfect, but nevertheless appear to be of adequate validity. As a pilot study, it uncovered minor flaws in the procedure, which will be corrected before it is used again. The grades that came out of the reconstruction were reasonable, and were accepted by the students with no more than normal grumbling. The fact that the grading procedure is effectively opaque to the students\(^6\) had worried me, but didn't cause any major difficulties.

The principal practical problem that I found was that of time. I had allocated 20 minutes for each round, which was nowhere near sufficient. The groups need time to organise their work, and time to reach a consensus, above and beyond the time needed to work on the actual problem. An hour per round appears more reasonable, with significant breaks in between.

A larger number of rounds would improve the statistical stability and accuracy of the reconstruction, by increasing the number of data points. Particularly in the 9-groups, the number of degrees of freedom in the reconstruction is very small, making it highly sensitive to perturbations. Four rounds is, however, the maximum possible with nine students, without the same students re-encountering each other in the groups. Larger macro-groups appear to be called for, and will be tested in simulation. With at least 15 students, six rounds are possible. Great care must be taken when assigning the students to groups, in order to avoid the kind of mishap that perturbed one of the macro-groups here.

Supernumerary students are highly undesirable from a reconstruction point of view; if the number of students is not a multiple of three, it may be preferable to leave vacant seats, letting some students work in groups of two. A by-product of such an arrangement would be data on how performance varies with group size.

The results of the experiment are reproduced fairly well by simulations with a noise level around 15%. This inspires some confidence in the model, simplistic as it is, and in the grading as well. The grading accuracy in simulation, with nearly 80% correct grades, is quite competitive with that of other assessment procedures, where inter-grader consistency is often significantly worse (Abedi et al, 1995).

All in all, this pilot study has been fruitful and generated useful answers to all the questions raised in the introduction. The assessment method basically works, and appears ready for a full-scale trial, where it will be more systematically evaluated. This next-stage experiment will take place in early 1998, with a larger and less biased group of students, who have agreed to be doubly assessed, with a traditional individual exam as well as with the experimental group procedure. The individual exam will serve as a control, giving a solid baseline for comparison, and for firmer conclusions about the reliability and validity of the group-derived individual grades.

---

\(^6\) The students who wished to had full access to all aspects of the procedure, but few of them could follow either the mathematics or the computer code involved, so that didn't help much. I did spend a fair bit of time explaining the algorithm to those who asked for it, but most of them took it on faith.
References


Habeshaw, S & Gibbs, G & Habeshaw, T (1993) 53 interesting way to assess your students (Rev. Ed.) Bristol: Technical and Educational Services


Johansson, Sverker (1998) Utvärdering av en innovativ examinationsform i två steg för stora studentgrupper Submitted to Insikt


Lejk, M & Wyvill, M & Farrow, S (1997) Group learning and group assessment on undergraduate computing courses in higher education in the UK: results of a survey Assessment & Evaluation in Higher Education 22(1)267-280


Murray, J P (1990) Better testing for better learning College Teaching 38(4)148-152


Wiggins, Grant (1990) The case for authentic assessment ERIC Digest ED328611
Appendix A: Exam questions

Sample from Astro:

2. **A time machine** could be useful for a lot of things. But the one that you borrow here for this exam has some limitations. You can’t adjust where and when you’re going; instead you are limited to a few pre-set times. But you remain at the same spot in space no matter how you move in time — the machine is smart enough to compensate for the Earth’s movement in space, so that you remain at the same place relative to the Earth. Anyway — you get into the machine, and test the pre-set times. Below are descriptions of what you can see outside the machine, at the various times. Your assignment is to identify, for each description, when you are.

a) No Earth below your feet (but you’re still where the Earth should be). You are suspended in empty space. You can see a tiny blue-white dot shining very sharply, where the sun ought to be. Around you is a normal starry sky, even though you can’t recognize any constellations. But the sky isn’t quite clear, you see it through a multicolored veil.

b) No solid ground directly below your feet, but a long way further down. You are at a high altitude above the surface of the Earth of that time, high enough so that you can see that the Earth is round — but not as round as it is today, noticeably irregular. The sky is full of crap: clouds, dust, gravel, small and large rocks. To the right you can see a hint of a glow behind the clouds, but to the left it’s all crap. But straight up- and downwards the clouds appear thinner, and you vaguely perceive a starry sky beyond, with several bright blue, obviously nearby stars.

c) Everything you see is glowing red-hot. No Earth below your feet, no moon, no stars in the sky. Everything is red. It’s hot! (There are actually two or three solutions to this one.)
**Sample from Weather:**

1. Hur kan man förklara för barn i lägstadiet åldern som undrar:
   a) Varför regnar det, och var kommer allt vattnet ifrån? Tar det inte slut?
   b) Varför blir det kallt på vintern, och varför blir det mörkt så tidigt då?
   c) Vad händer när ett flygplan krockar med ett moln?

Jag vill att du ger vettiga svar till ovanstående frågor som både är fysikaliskt korrekta, och begripliga för den sju-åttaåring som frågar.

**Translation**

1. Hur can you explain for kids in the early school years who wonder:
   a) Why does it rain, and where does all the water come from. Doesn’t it ever run out?
   b) Why does it get cold in the winter, and why does it get dark so early then.
   c) What happens when a plane collides with a cloud?

I want you to give sensible answers to these questions that are both physically correct and comprehensible to the 7-8-year-old who’s asking.