STUDENT DISCOVERY PROJECTS IN DATA ANALYSIS

Mike Forster[1] and Helen MacGillivray[2]
[1]University of Auckland, New Zealand
[2]Queensland University of Technology, Australia
m.forster@auckland.ac.nz

At Queensland University of Technology, students have been doing self-selected group projects in data analysis as a major component of their course assessment for well over a decade. These projects provide experiential learning of statistical problem-solving and the data investigation cycle of plan, collect, process, discuss in topics of their choice. Datasets must involve at least four, and preferably more, variables. Auckland University is introducing similar discovery projects in a second year data analysis course in 2010. In this paper, we discuss the background and motivation for such projects; experiences in setting up, running, assessing and administering self-selected student project work; the types of projects students choose to do and guidance they receive; effects of the students' work and reports; our experiences with project students and finally, our assessments of the value to students of doing self-selected group project work.

INTRODUCTION

I hear, I forget. I see, I remember. I do, I understand (Chinese proverb).

Among statistical educators, there is universal acceptance of the value in having statistics students engage in experiential learning in real contexts with real data, and in the whole problem-solving experience of the PPDAC (Problem, Plan, Data, Analysis, Conclusion) data investigation cycle (Wild & Pfannkuch, 1999). Data analysis projects can play vital roles in fulfilling these objectives. Many different types of project work have been suggested: topics given to the students (Jolliffe, 2002), extra-curricular competition based projects with students collecting their own data (Habbullah, 2002) or using existing data (Starkings, 2002), and activities using data collected by students to tackle questions posed by instructors (Zeleke, Lee & Daniels, 2006). Recent work on virtual systems (Darius, Portier & Schrevens, 2007; Steiner, 2009) is also breaking new ground in providing environments for experiential learning in data investigations in complex realistic situations, although these tend to require students to apply previous knowledge. Bulmer's (2010) virtual island is providing a virtual environment for introductory science students to design, carry out and analyse experiments.

In this paper, we focus on self-selected group project work in early undergraduate courses in statistics. These self-selected projects are on topics chosen by the students, who then design the study, collect and analyse the data and communicate their results and findings in a written report (MacGillivray, 1998). Chance (2005) describes such projects as 'the most valuable learning experiences for my introductory students'. Lee (2005) comments on the change in students' approach to projects, from indifference towards real world problems posed by the instructor to interest in self-selected topics, with consequent improvement in analysis and reports.

The references above are representative of the many that provide the pedagogical framework for such projects. This paper provides a review of the strategy for university instructors. After commenting on frequently-asked questions of the Queensland experiences of fifteen years of conducting such projects in large introductory courses, the experiences of a peer reviewer are reported leading to plans for introduction of similar projects at Auckland University. After a brief update on quantitative and qualitative evaluations, the paper closes with remarks on expectations of the strategy.

ASSISTANCE FROM QUEENSLAND EXPERIENCES

Self-selected group projects began in introductory statistical courses at the Queensland University of Technology more than fifteen years ago. The original concept was to give students hands-on experience with planning and carrying out data collection, together with exploration of the data and reporting features. The motivations were those of experiential learning and communication in data investigations. For example, no matter how real the contexts, data and discussion of examples provided by the instructor, students have to accept the instructor's version

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

of the story, and the learning of the vital PPD components of the cycle now called PPDAC cycle is at best controlled, usually passive, and at worst missing. The strategy was trialled with a moderately-sized first year cohort and then a large (400+) engineering cohort. However the engineering cohort's desire to apply the statistical methods they were learning to "their" data lead to the extension to include data analysis. They are called discovery projects because the experiential learning of the full PPDAC cycle is within a curriculum whose objectives are to develop the skills, capabilities and knowledge to carry out real data investigations in life-related situations with many variables, using technology in the analysis and reporting. The content includes EDA, interval estimation and hypothesis testing, including the use of chisquare, ANOVA and multiple regression, and the strengths and limitations of these introductory statistical data analysis methods.

Students form their own groups, with assistance when needed, and are provided with information, criteria and standards, exemplars in the forms of model projects and past projects. They suggest and instructors advise, and examples taken from past projects and datasets are used throughout the course in examples, exercises, tutorials and practicals.

A concern of instructors is the workload for staff and students (for example, see Chadjipadelis & Andreadis, 2006; Darius et al., 2007). Certainly students need assistance in selecting the topic and planning the collection to ensure an educationally-productive dataset is obtained efficiently, but shepherding wonderfully enthusiastic ideas into manageable investigations is a rewardingly collegiate experience for students and instructors. Overly-complex categorical variables, excessive numbers of observations on too few variables, and heavily time dependent topics are best avoided. Workload for instructors during semester is controlled through resources, curricula design, use of past projects, information and criteria, sound administrative systems, and tutors with training or personal experience doing these projects. Assessment workload is controlled in two ways: by peer-validated criteria and standards and use of criteria sheets; and, because the projects are rich assessment tasks, by the other assessment focussing on operational knowledge through itemized questions or short response items. Such assessment is readily amenable to marking schemes suitable for many and/or less experienced markers, and is faster to mark than conventional questions that try to combine operational knowledge and statistical thinking.

With more than 2500 projects carried out, there are many examples of types of projects (see, for example, MacGillivray, 2005). The students' imaginations are stimulated, not necessarily by their discipline. For example, engineering students often pursue topics with no apparent relevance to engineering content, but which demonstrate development of creativity and enquiry. Tables 1 and 2 provide a classification of the 120 engineering student projects of 2009 by topic and by type of investigation.

Table 1. Types of topics and percentage chosen in 120 projects in 2009

Types of topics	Transport	Experiments	Student environment (incl. housing)	Clothes, food, drink	Media	Other (incl. sport, general interest)
Percentage	21%	20%	20%	13%	13%	13%

Table 2. Types of investigations and percentages in 120 projects in 2009

Types of investigations	Observational	Data researched and collected	Designed experiments	Surveys	Workplace sources
Percentage	31%	26%	20%	16%	6%

Concerns that such projects may not be carried out "correctly" demonstrate misunderstanding of their roles as learning experiences alongside more structured activities and items. In contrast, although research topics and case studies can motivate valuable contexts for examples, extracts from these adapted for early undergraduate activity, tend to be either a controlled and sanitized learning experience or dominated by complicated information.

Doing these discovery projects in groups is advantageous for students and instructors. Many aspects, such as brainstorming, planning, collecting data and interpreting the results in context, benefit from a cooperative approach, and the peer and self-directed learning fostered by groups helps to lighten instructors' loads. Group problems can never be completely eliminated, but can be reduced by good support and system procedures, including some discussed in the next section. Concerns that groups encourage free riders who do not benefit from the learning experience are alleviated by analysis of results for individual students such as that reported in Section 6. The nature of the project facilitates collaborative learning. In the third project stage of analysis and reporting, advice on task division can assist in cases of group tension.

Tutor confidence in assistance on projects comes from experiencing such projects themselves or needs to be developed through training. Graduates in research or the workplace report that their early experiences in these self-select discovery projects are invaluable in developing statistical communication as well as learning statistical research and enquiry, including how to turn research questions into statistical questions and tackle complex real problems.

RECENT COLLABORATIVE DEVELOPMENT

All practical, ethical and health and safety issues were handled through students reporting their plans in writing and receiving advice during practical classes. However in 2008, it was decided to formalize this in a form developed by the authors during collaborative visits. This form includes statements about ethics and health and safety as well as incorporating the description of the plan in the usual required format including identification of variables, their types, and details of data collection or access. The requirement that the form must be signed by all group members and must be submitted in order for the project report to be accepted later, not only commits the students to seek permission for changes to their plan, but also helps to avoid group problems. While visiting Queensland University of Technology, the Auckland author vetted over 120 project forms and found that he could do about 20 per hour, including providing written feedback where necessary.

Surveys are permitted only under strict instructions. Design of good survey instruments is difficult and time consuming (Chadjipadelis & Andreadis, 2006). In 2009, the only surveys were of students. Questions must be non-invasive as well as unambiguous, and conducted face-to-face. Auckland University is proposing to mandate even more strictly on surveys.

Critiquing of the plan and quality of data has always been part of the reporting, but in 2009, more explicit emphasis and examples in lectures on adjudging the representativeness of data resulted in improved comments on this in student reports.

AN INSTRUCTOR'S FIRST EXPERIENCE WITH SELF-SELECTED STUDENT PROJECTS

Apart from supervising graduate dissertations, the only hands-on experience of the Auckland author in such data analysis projects was two visits made to Queensland University of Technology. The aspects that were most striking in working with the Queensland students were: their enthusiasm for their projects; their sense of ownership of their ideas and their data; their desire to understand and learn through their project work; but mostly, watching them think through problems and reach consensus, with minimal, if any, guidance from instructors.

A group of engineering students doing a study on effect of leaflet colour on rejection or acceptance invited the visitor to observe their data collection one afternoon. They had already done a pilot run that morning and began the afternoon by discussing if and how they should modify their data collection process. Their pilot and general approach had been planned in previous discussion with the course coordinator. From their pilot, problems were identified and solutions suggested. They revisited the debate on their objectives which led to questioning what their data actually represented. They realized they really had two data sets: one at the individual level (the characteristics of the person taking, or not taking a leaflet); another at an aggregated level, as they were also recording the time it took to dispose of a fixed number of leaflets of different colours. When they realized just how rich their data was, their enthusiasm grew. Throughout this, the visiting academic was merely an observer.[1]

A member of another student group approached the visitor in a tutorial class and showed his group's data. The group wanted to investigate the effect on the daily quantity of petrol sold of a pricing scheme that, as an economist, the visitor could not believe. These students had a difficult

time convincing the visitor that petrol prices in Australia (except apparently in Western Australia) varied from day to day in a regular weekly (seasonal) pattern. Apart from horror that price fixing like that existed in Australia, this was worrying because the course the students were doing did not include any formal time series analysis. On a subtle enquiry as to whether the topic had been approved, the response was "not yet". Since it was obvious that the students had already spent considerable time recording their data and had come up with an interesting economic situation to investigate, the visitor decided to risk the wrath of the course coordinator and suggested he teach the students about modelling trends and seasonality in time series. The initial discussion concerned missing values. It was explained to the students that time series is ordered data, and that missing values were a problem and, where possible, need to be imputed. The students decided to use the average quantities from the same day from the previous week and from the future week. As these students encounter only an introduction to ANCOVA, they were shown how to do a time series plot and how to fit a time trend using simple linear regression. Once they understood that any patterns in their data not captured in their model will be transferred to the residuals (they had met this in both the ANOVA and multiple regression sections of their course), they were shown how to estimate seasonal effects using a one-way analysis of variance on the residuals from their linear regression. They even wanted to know how to assess the quality of time series predictions. They were told there were many statistics that can be used, for example, RMSEP (a concept commonly introduced in third year applied time series), but that they already had more than sufficient work for their project. When they sent a copy of their final report to Auckland, we were amazed to see they had done a week's worth of predictions, and assessed their usefulness using RMSEP.[2]

PLANS FOR AUCKLAND

The service course in data analysis (1300 students per year) at Auckland is the second largest course after the introductory course (4500 students per year). There was a fear that if students didn't like the idea of doing group project work as a major part of their assessment, introducing projects unilaterally in this second year course could cost dearly. In order to win student support, and to minimize instructor workload while setting up and refining the necessary course infrastructure, it was decided to offer a new version of the course that differs only in the assessment: reduced assignment work, the group project and an exam, compared to the current scheme that includes a set of assignments, a test and an exam (Forster & Smith, 2007). The new version of the course will have an enrolment limit initially, and, if the self-selected projects are well received by our students and feasible administratively, we will extend it. As well as setting up an on-line system that will allow students to register themselves as a group, and discuss their project online amongst themselves and with the instructor, we are planning online registration of their project for approval, and an open forum for students to float project ideas and form groups.

Communication of statistical findings is an important and integral part of any statistical analysis (for example, see Davies and Connor, 2005). At Auckland, the focus in our second year service course in data analysis has been to develop the course around case studies with highly stylized concise written reports (Forster et al, 2005). The reports were designed in a one-size-fits-all style to reinforce in our students the key components of any statistical analysis that should be reported. Our experiences with graduate student dissertations has alerted us to the fact that writing more comprehensive reports that are more tailored to a particular situation or scenario is something we also, regrettably, need to teach. If we expose our students at an earlier stage to the kind of report writing that is required when doing scientific data investigations, be it a dissertation or a report for their boss, our students should have more appeal once they move out into the world.

Today's students want exemplars (MacGillivray & Hayes, 1997), and exemplars that are in a consistent style (Francis, 2005). We are seriously considering using different styles to further reinforce the idea of context as the report style depends in part on the context.

Another advantage we see in having students do project work is to reinforce key underlying concepts of statistics (such as variation, representativeness, statistical reasoning and context) and to deepen their understanding of statistical techniques, their importance and their place in the wider scheme of things (e.g. plotting data, creating numerical summaries, modelling and the underlying process of scientific investigation that uses statistics).

Perhaps the biggest advantage is that group projects force students towards holistic learning of statistics while also developing many other useful skills: research skills (Loi, 2002); problem solving and decision making skills (Del Mas, 2005); team work (Matis, 2006). These externalities (third party benefits) not only help our students (and us), but they will also be of substantial benefit to colleagues teaching in other disciplines. In essence, projects make students think! Self-selected group projects make students think more.

EFFECTS ON COURSES AND STUDENTS

Information from interaction with students on their projects and from the reports is constantly feeding back into the design of the curriculum and resources in Queensland. Because of increased student engagement and motivation to learn more, the curriculum has been both streamlined and extended. As reported in Johnson and MacGillivray (2009), the culmination of many gradual improvements has resulted in significant changes to the structuring of the content, and analysis of quantitative measures of student performance demonstrates improvement in student performance in both projects and the operational knowledge assessed in examinations.

The Auckland author discovered that neither of the experiences of Section 4 was unusual. In particular, the strategy of self-selected data investigation projects facilitates student learning across the full range of student capabilities. The less capable students are able to apply, with some assistance, the core principles and methods of the course and gain a valuable sense of achievement. The most capable students typically self-extend, again with some assistance, and their rapidity in developing statistical thinking and confidence with statistical methods results in reports that amaze observers and are often used by the students in portfolios attached to their curriculum vitae.

The emphasis in the projects, their criteria and assessment is on formulating ideas, planning and carrying out data investigations, choice and use of statistical techniques, and synthesis and reporting of results in context. The building blocks of statistical knowledge, interpretation and communication are assessed by quizzes and examinations consisting of itemised responses and short response items. However, as reported in Johnson and MacGillivray (2009), analyses of the end of semester examination marks demonstrate the value of the group projects for individuals, and in their learning of statistical knowledge and skills. These effects are again demonstrated in analysis of the 2009 cohort of more than 500 engineering students. The quizzes provide exemplars for the end of semester examination, so it is not surprising that in a regression model of the examination mark on all other assessment components, the quizzes ($p = 0.000$) are significant, but in the presence of these, the project mark, which is the same for all group members, is also highly significant ($p = 0.000$) in predicting the exam marks which are individual. Because of slight heteroscedascity, a square root transformation of the exam marks was used, giving an almost perfect normal residual plot. No other residual problems were present and the model explained 33% of the individual exam mark variation

SUMMARY OF THE VALUE OF GROUP PROJECT WORK

- Brings all aspects of statistics together from the initial problem to the final written report.
- Makes students think about statistics and about real world problems in context.
- Gives students experience in team work, which is valuable for the modern work environment.
- Enhances students' written and oral communication skills.
- Provides learning for all students and extends better students beyond the course syllabus.
- Provides externalities to all other disciplines.

CONCLUSION

The Auckland author's position is that if the change in attitude towards statistics and the overall performance of the students who are doing self-selected group projects in data analysis takes us half-way to what was encountered in Brisbane with the Queensland students, we will not only be very happy but will get further confirmation of the value of self-selected group projects in data analysis. More than anything, we expect to learn as much as our students do.

NOTES

1. Derriman, G., Praeger, K., Vearncombe, D., & Webb, B. Flyers. *QUT student project.*
2. Phabmixay, M., Rovere, D., Szetu, A., & Yang, D. Petroleum. *QUT student project.*

REFERENCES

Bulmer, M. (2010). Technologies for enhancing project assessment in large classes. In Reading, C. (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics*, Ljubjlana. Voorburg, The Netherlands: ISI. (to appear)

Chadjipadelis, T., & Andreadis, I. (2006). Use of projects for teaching social statistics: Case study. In Rossman, A., & Chance, B. (Ed.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador. Voorburg, The Netherlands: ISI.

Chance, B. (2005). Integrating pedagogies to teach statistics. In J. Garfield (Ed.), *Innovations in Teaching Statistics*, MAA Notes #65, Mathematical Association of America, Washington.

Darius, P., Portier, K., & Schrevens, E. (2007). Virtual experiments and their use in teaching experimental design. *International Statistical Review, 75*(3), 281-294.

Davies, N., & Connor, D. (2005). Helping students to communicate statistics better. In L. Weldon & B. Phillips (Eds.), *Proceedings of the ISI/IASE Satellite on Statistics Education & the Communication of Statistics*, Sydney. Voorburg, The Netherlands: ISI.

Del Mas, R. C. (2005). Teaching statistics to under-prepared college students. In J. Garfield (Ed.), *Innovations in Teaching Statistics*, MAA Notes #65, Mathematical Association of America, Washington.

Forster, M., Smith, D., & Wild, C. (2005). Teaching students to write about statistics. In Weldon, L. & Phillips, B. (Eds.), *Proceedings of the ISI/IASE Satellite on Statistics Education & the Communication of Statistics*, Sydney. Voorburg, The Netherlands: ISI.

Forster, M., & Smith, D. (2007). Assessing large second year undergraduate service courses in data analysis. In Chance, B. & Phillips, B. (Eds.), *Proceedings of the ISI/IASE Satellite on Statistics Education*, Guimares. Voorburg, The Netherlands: ISI.

Francis, G. (2005). An approach to report writing in statistics courses. In Weldon, L. & Phillips, B. (Eds.), *Proceedings of the ISI/IASE Satellite on Statistics Education & the Communication of Statistics*, Sydney. Voorburg, The Netherlands: ISI.

Habbullah, N. H. (2002). Data analysis talent award. In B Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, Cape Town. Voorburg, The Netherlands: ISI.

Johnson, H., & MacGillivray, H. (2009). Constructing environments for early learning of statistical thinking in higher education, *RSS Conference*, Edinburgh, September.

Jolliffe, F. (2002). Statistical Investigations – Drawing it all together. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, Cape Town. Voorburg, The Netherlands: ISI.

Lee, C. (2005). Using the PACE strategy to teach statistics. In J. Garfield (Ed) *Innovations in Teaching Statistics*, MMA Notes #65, Mathematical Association of America, Washington.

Loi, S. L. (2002). Final year business students experiences of data analysis in projects. In B Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, Cape Town. Voorburg, The Netherlands: ISI.

MacGillivray, H., & Hayes, C. (1998). *Practical Development of Statistical Understanding: a project based approach*. QUT press, Brisbane.

MacGillivray, H. (1998). Developing and synthesizing statistical skills for real situations through student projects. In *Proceedings of the Fifth International Conference on Teaching Statistics*, 1149-1155, Singapore. Voorburg, The Netherlands: ISI.

MacGillivray, H. L. (2002). One thousand projects. *MSOR Connections, 2*(1), 9-13.

Matis, T.I. (2006). Conceptualizing applied probability through project-based learning. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador. Voorburg, The Netherlands: ISI.

Starkings, S. (2002). Pedagogic issues required for successful statistical project competitions. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, Cape Town. Voorburg, The Netherlands: ISI.

Steiner, S., & MacKay, R. J. (2009). Teaching process improvement using a virtual manufacturing environment. Submitted to *The American Statistician*.

Wild, C. J. & Phannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67(3),* 223- 265.

Zeleke, A., Lee, C., & Daniels, J. (2006). Developing projects based on student's data in introductory statistics. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador. Voorburg, The Netherlands: ISI.