Durham University

Durham E-Theses

Effects of student ratings feedback and a group intervention on the quality of university teaching : a randomised controlled trial

Angela R. Penny

How to cite:

Penny, Angela R. (2004) Effects of student ratings feedback and a group intervention on the quality of university teaching : a randomised controlled trial. Doctoral thesis, Durham University.

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a https://etheses.durham.ac.uk/id/eprint/1735/ is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full Durham E-Theses policy for further details.

Academic Support Office, The Palatine Centre, Durham University, Stockton Road, Durham, DH1 3LE
e-mail: e-theses.admin@durham.ac.uk Tel: +44 0191 334 6107
https://etheses.durham.ac.uk

EFFECTS OF STUDENT RATINGS FEEDBACK AND A GROUP INTERVENTION ON THE QUALITY OF UNIVERSITY TEACHING: A RANDOMISED CONTROLLED TRIAL

ANGELA R. PENNY

A copyright of this thesis rests with the author. No quotation from it should be published without his prior written consent and information derived from it should be acknowledged.

Ph.D. THESIS 2004

2 3 JUN 2004

Abstract

There is evidence that feedback from student ratings combined with individual consultation is an effective means to improve teaching effectiveness. But this traditional model now seems too costly for universities because budgets are being cut. A randomised controlled trial was conducted to investigate the effects on teaching effectiveness of augmenting ratings feedback with consultation in peer support groups, a group-based feedback intervention. A sample of 71 teachers grouped according to their department units were randomly assigned to receive the intervention of ratings feedback and engage in dialogue with colleagues in small groups, or no intervention (controls).

Results of OLS regression and multilevel modelling showed that teaching effectiveness improved significantly in the intervention group compared to the control group (*ES* = .24). Semi-structured interviews revealed that teachers are willing to collaborate with their colleagues over student ratings feedback. But there are factors associated with their work that present challenges for engagement in collaborative learning. A meta-analysis was also conducted to identify the practices and strategies that may be important for effective consultation over student ratings feedback.

The study results give preliminary support for group-based peer consultation as an effective means to improve teaching effectiveness. It is suggested that professional development activities should be organised within academic departments to allow teachers to meet in groups, engage in dialogue, and learn together how to use student ratings feedback to improve their teaching. Further investigation through replication in different settings is needed to provide evidence of the effectiveness of this strategy to improve the quality of university teaching.

Declaration

I declare that this thesis, which I submit for the degree of Doctor of Philosophy at the University Durham, results entirely from my own work and has not previously been submitted for a degree at this or any other university.

Statement of Copyright

Copyright © 2004 by Angela R. Penny

The copyright of this thesis rests with the author. No quotation or data from it should be published without her prior written consent and information derived from it should be acknowledged.

This thesis is dedicated to Vera, my mother,
for the faithful prayers

Acknowledgements

My sincere thanks to Dr. Robert Coe who provided guidance and critical support that helped moved the research project through the final stage to completion in a timely manner. Thank you for the generous comments that helped me to become more careful when interpreting data.

I am grateful to Professor Carol Fitz-Gibbon for advice and support from development through to the intermediate stage of the research project. Thank you for helping me to appreciate the need for experiments to guide educational practice.

I am greatly indebted to the administrators and teachers for their assistance, participation, and superb support that made it possible to carry out the research.

Special thanks to my sister, Marcia and my many friends for their assistance, encouragement, and prayerful support at various points of this research project.

To my Lord and Saviour, Jesus Christ, I am so thankful for the gift of amazing love and amazing grace. All glory and honour to the Giver and Bread of life.

Related Publications

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, *8* (3), 399-411.

Penny, A. R., & Coe, R. (in press). Consultation over student ratings feedback, what is effective? A meta-analysis. *Review of Educational Research.*

Penny, A. R. (in press). Responding to students' views about university teaching: The experience of teachers in Jamaica. *International Journal of Educational Development.*

Table of Contents

Abstract ii

List of Figures, Tables, and Appendices xiv

Chapter 1 Introduction 1

Context of the Problem 1

Purpose of the Study 4

Significance of the Study 6

Research Questions 8

Conceptual Framework 9

Chapter 2 Student Ratings of Teaching: Valid and Reliable? 14

Introduction 14

What is Teaching Effectiveness? 15

What do Student Ratings Measure? 18

Student Learning 19
Evaluation by Different Evaluators 21

How are Student Ratings Used? 22

Are Student Ratings Valid? 25

Student Variables 27
Teacher Variables 29
Course Variables 32
Administration Procedures 34

Are Student Ratings Reliable? 35

Interrater Reliability 36
Stability or Generalisability 37

Chapter 3 Shortcomings of Student Ratings Feedback Research 39

Introduction 39

Shortcomings of Ratings Research 41

Construct Validity and Rating Forms 42
Use of Ratings Data by Administrators 45
Relevance of Ratings Research to Teachers 48
Conceptions of Learning and Teaching 52

Conclusion 55

Chapter 4 The Practice of Consulting over Student Ratings 56

Impact of Consultation over Student Ratings 56

The Practice of Teaching Consultation 61

The Consultation Process 66

Models of Consultation 67

Types of Consultation Programmes 69

Use of Support Groups 71

Chapter 5 Consultation over Student Ratings: What Works? A Meta-analysis 75

Introduction 75

Meta-analytic Reviews 79

Merits of Meta-analysis 80
Limits of Meta-analysis 85

The Study 87

Criteria for Inclusion 87
Identification of Studies 89
Coding Procedures 90
Outcome Measures 91
Moderator Variables 92
Analysis 93

Results 97

Distribution of Effect Sizes 100

Outlier Procedures 102

Publication Bias 104

Effect on Teaching Effectiveness 106

Moderator Analyses: Consultation Practices and Strategies 109

Effect of Process Variables 112

Approach to Consultation 112
Observation 113
Videotape Recording 113
Interview 114

Effect of Context Variables 114

Duration of Consultation 115
Consultant 115
Nature of Consultation 115
Participants 116

Effect of Content Variables 116

Normative Data 116
Self-Ratings 117
Improvement Goals 117

Effect of Study Variables 118

Ratings Forms 118
Publication Type 119
Publication Year 119

Discussion 119

Maximising the Effects of Consultative Feedback 121

Process Variables 122

Interview 123
Observation 124
Videotape Recording 124

Context Variables 125

Duration of Consultation 125
Consultant 127
Nature of Consultation 128
Participants 129

Content Variables 130

Normative Data 130
Self-Ratings 131
Improvement Goals 133

Study Variables 134

Ratings Forms 134
Publication Type and Year 135

Implications for Practice 137

Directions for Future Research 143

Chapter 6 Methodology 145

Introduction 145

Randomised Controlled Trials in Education Research 148

Context of the Study 151

Sample 154

Ratings Collection 156

Mid-semester Ratings 156
End of Semester Ratings 157

Identification of Peer Support Groups 159

Instrumentation 161

The Intervention 166

Ratings Report 167
Interpretation Guide 167
Teaching Ideas (TIPS sheets) 168
Peer Group Meetings 169

The Interviews 169

Treatment for Control Group 172

Letter of Appreciation 172

Statistical Procedures 173

Data Screening 173
Preliminary Analysis 175
Primary Analysis 175
Multiple Regression and Multilevel Modelling 177
Variables 179

Ethical Research Practice 180

Chapter 7 Results 182

Preliminary Analyses 183

Factor Analysis 183
Reliability Analysis 187

Primary Analyses 189

Participant Characteristics 189
Attitude towards the Value of Student Ratings 191
Feedback Rating Scores 193

Analysis I: OLS Regression Model 195

Teacher-level Analysis 195
A Ceiling Effect? 200
Department-level Analysis 202
Subgroup Analysis 204

Analysis II: Multilevel Modelling 205

MLM Models 206

Presentation of Results 208

Results 210

Variance in Teaching Effectiveness 210
Effects on Teaching Effectiveness 212
What Size of Effect? 214
Effect of Teacher Characteristics 215
Effect of Course Characteristics 218

Comparison of Results from OLS Regression and Multilevel Analysis 220

Summary of Quantitative Results 223

Chapter 8 Teacher Perceptions of Consultation in Peer Support Groups 225

Introduction 225

Talking with Colleagues 226

Talking over Student Ratings 227
Factors Affecting Participation 230

Impact of Ratings Feedback on Teaching 232

Impact on Teachers 232
Impact on Teaching Practice 234
Training for Students 235

Conditions for Use of Peer Support Groups 236

Sufficient Learning Time 236
Support from Administrators 237
Feedback Information 237

Summary of Qualitative Results 238

Chapter 9 Discussion 240

Overview of Study 240

Summary of Findings 241

Contribution of the Meta-Analysis 242

Discussion 244

Ratings Feedback on Teaching Effectiveness 244
Departmental and Institutional Influence 246
Collaboration for Teaching Improvement 250
Scholarship of Teaching 253

Evaluation of the Intervention 255

Implications for Practice 258

Directions for Future Research 261

Conclusion 264

Appendices 266

References 309

List of Figures, Tables, and Appendices

Figures

Figure 4.1 Instructional Consultation: A Typology of Programmes 70

Figure 5.1 Stem-and-Leaf Plot Showing Distribution of Mean Effect Sizes 100

Figure 5.2 Effect Sizes, Overall Effect Sizes and their 95% Confidence Intervals by Study 101

Figure 5.3 Funnel Plot of Effect Sizes 105

Figure 5.4 Distribution for the Overall Mean Effect of Consultative Feedback 108

Figure 5.5 Effect Sizes as a Function of Use of Additional Information on Teaching 113

Figure 5.6 Effect Sizes as a Function of Consultation Context Variables 114

Figure 5.7 Effect Sizes as a Function of Consultation Content Variables 117

Figure 5.8 Effect Sizes as a Function of Study Variables 118

Figure 6.1 Model of Nested Sample Design for this Study 148

Figure 6.2 Hierarchical Structure of Ratings Feedback Data in this Study 160

Figure 7.1 Attitude towards the Value of Student Ratings 192

Figure 7.2 Pre-feedback Ratings on SEEQ Scales on Peer Group Means 194

Figure 7.3 Interaction Effect for Individual Teachers 199

Figure 7.4 Post feedback Ratings on Initial Level of Teaching Effectiveness 200

Figure 7.5 MLwiN Output: Results Estimating the Variance in Teaching Effectiveness 210

Tables

Table 2.1	Relationship between Student Characteristics and Student Ratings	28
Table 2.2	Relationship between Teacher Variables and Student Ratings	30
Table 2.3	Relationship between Course Characteristics and Student Ratings	33
Table 2.4	Relationship between Administration Procedures and Student Ratings	35
Table 5.1	Summary of Studies included in the Meta-analysis	99
Table 5.2	Overall Effect of Consultation on Teaching Effectiveness	106
Table 5.3	Mean Effect Sizes as a Function of Moderator Variables	111
Table 6.1	Teacher Participation by University and Intervention Group	158
Table 6.2	Description of Variables used in this Study	180
Table 7.1	Factor Analysis Results of the SEEQ Items for the Total Sample after Rotation	186
Table 7.2	Interrater Reliability Estimates: Total Sample and Different Class Sizes on SEEQ scales	188
Table 7.3	Description of Participants in Percentages by Intervention Groups	190
Table 7.4	Post feedback Ratings by Intervention Group on SEEQ Scales for Teachers	194
Table 7.5	OLS Regression Results for Teacher-level Analysis	196
Table 7.6	OLS Regression Results for Department-level Analysis	203
Table 7.7	Multilevel Results Estimating Effect on Teaching Effectiveness	212
Table 7.8	Multilevel Results Estimating Effect of Teacher Characteristics	216
Table 7.9	Multilevel Results Estimating Effect of Course Characteristics	219
Table 7.10	Comparison of Multilevel with OLS Regression for Effect on Teaching Effectiveness	221

Appendices

Appendix A	Letter to Administrators	266
Appendix B	Letter to Teachers	268
Appendix C	Consent Form	269
Appendix D	Procedures for Administering Student Questionnaires	270
Appendix E	SEEQ Rating Forms	271
Appendix F	Teacher Self-Evaluation Questionnaire	275
Appendix G	Frequency of Talk with Colleagues about Teaching	277
Appendix H	Letter with Ratings Report	278
Appendix I	Guidelines for Peer Support Groups	279
Appendix J	Sample of Ratings Report	280
Appendix K	Ratings Interpretation Guide	283
Appendix L	Teaching Ideas and Practical Strategies (TIPS) Sheets	285
Appendix M	Interview Guide	301
Appendix N	Letter for Control Group	302
Appendix O	Letter of Appreciation	303
Appendix P	Correlations of SEEQ Factors	304
Appendix Q	Teacher Self-Ratings by Intervention Group	306

Chapter 1

Introduction

Context of the Problem

"*What am I supposed to do with this?*" This was the question asked, somewhat angrily, by one lecturer on receiving feedback from student ratings for the very first time in July 2000 in one university in Jamaica. The formal system of teaching evaluation was prompted by the establishment of quality assurance procedures to monitor teaching quality among other things. Symbolically, the envelope bearing the ratings report contained only a single sheet of paper, was sealed, and stamped 'Confidential'. This could be interpreted as, "for your eyes only", "keep it to yourself" or even "you're on your own". After all, there were no procedures in place to help lecturers make sense of the ratings feedback.

Many lecturers, I imagined, examined the report and felt satisfied that the ratings in a number of areas were good and ignored the low ratings. Interestingly, even where course team members shared the same office space, or met for course planning and

programme review, no mention was made of the ratings feedback. This is perhaps not so surprising given the norm of individualism and privacy in university teaching. The responsibility for improving teaching was that of the individual teacher and not a collective one. This experience became the guiding focus for investigation on the applicability of a peer-based model of consultation to help teachers learn from ratings feedback how to improve their teaching.

Faced with unprecedented social, economic and political demands for accountability and assuring the quality of teaching, universities around the world have come to adopt the use of student ratings of teaching as part of their response to these demands. Accordingly, use of student ratings has emerged as a key indicator of teaching quality and an important component in quality monitoring. The growing emphasis on teaching is also part of the broader educational reform efforts that view effective teaching as critical to better learning opportunities for students.

For this reason, university administrations routinely collect student ratings on the tacit assumption that the ratings will provide informative feedback teachers will be able to use to improve their teaching. Of course, it is reasonable to expect that teaching performance can and will benefit from student views. But giving teachers ratings feedback alone only partially supports teaching effectiveness. It should be remembered that many university teachers have had little or no formal training in how to teach effectively. And, there is no relationship between research and teaching effectiveness (Hattie & Marsh, 1996).

University teachers need support if they are to use student ratings to improve their teaching. Research has demonstrated (e.g. Marsh & Roche, 1993; Murray, 1997) that the effect of feedback from student ratings on teaching effectiveness increases when teachers

receive support from a resource person in interpreting, reflecting on, and developing improvement strategies. Obviously, then, learning to use ratings feedback to improve teaching requires dialogue, information, and support to devise improvement strategies, not in isolation but in a collaborative setting.

The traditional approach to supporting teachers has been individual, one-to-one, consultation where a teaching consultant works with the teacher to bring about improvement in an aspect of teaching. There is solid evidence that this model is effective in improving teaching but its use is not feasible in the current context of higher education. For one thing, individual consultation is time-consuming, labour intensive, and fosters teacher isolation and individualism. For another, it is arguably very costly in a climate where universities are struggling with significantly reduced budgets and rising costs. The downside is that, as a consequence, support with student ratings is virtually non-existent in many universities. Many teaching evaluation systems therefore operate without a corresponding system that provides teachers with appropriate support and the opportunity to learn how to use ratings feedback effectively.

One alternative to the traditional individual consultation model and learning in isolation with ratings feedback is group-based peer consultation. This model of consultation over student ratings provides a structured context of learning and mutual support from disciplinary peers organised in support groups for the purpose of problem-solving and bringing about change in teaching on the basis of ratings feedback received. It may be seen as a process through which dialogue and social interaction become the vehicle for change.

Group-based peer consultation capitalises on the fact that university teachers are more or less organised as groups (course teams) in their department units, though not

physically bounded, value their peers' experience and knowledge, and that disciplinary peers are uniquely positioned to provide support to improve teaching in their common field (Boud, 1999; Jenkins, 1996, Hicks, 1999a). And yet, combining ratings feedback with learning in peer support groups is a complex interaction as it is based on a paradigm that is not yet dominant among academics, teacher collaboration on teaching improvement.

Purpose of the Study

By tradition, university teaching is characterised by norms of individualism, privatism, and isolation rather than by sharing with peers one's successes or failures in the classroom, let alone seek their help to enhance teaching effectiveness. Requests for help could easily lead to judgements about competency. And yet, collaboration among teachers is being promoted as a necessary condition for the improvement in teaching and learning envisage by educational reform efforts (Fullan & Hargreaves, 1992; Little, 1990, 2003; A. Hargreaves, 2001).

Examples of academics collaborating with colleagues on some aspects of teaching already exist. To illustrate, an international study into the impact of teaching improvement strategies found that consulting with peers on issues pertaining to course content and organisation, construction of examinations and assignments, and curriculum development and planning, is highly valued among academics (Wright & O'Neil, 1995). There are also examples of academics engaged in collaborative learning experiences which they report on as "professionally enriching" and being able to experience a "different way of being" (Cranton & Carusetta, 2002; Ferman, 2002). But to what extent is it possible to extend the boundaries of collaboration to get teachers to reveal to

colleagues sensitive and confidential data, in the form of student views about one's teaching, and seek their help to become more effective?

The aim of the present exploratory study was to evaluate the effectiveness of group-based peer consultation to improve the quality of university teaching. More specifically, the purpose of the study was to investigate the effect of augmenting feedback from student ratings with consultation in peer support groups on teaching effectiveness, as measured by changes in student ratings of teaching. A secondary purpose of the study was to gain insights into how teachers experienced the intervention and to determine their perceptions about the efficacy of consulting over ratings feedback in peer support groups as a means to improve teaching.

This study is unique in the sense that there is no known research on the use of ratings feedback and consultation in teacher peer support groups. The study was based on the assumption that university teachers are now more willing to engage in dialogue about teaching with their disciplinary peers. There is also the assumption that by talking about their ratings feedback, sharing ideas, experiences, and knowledge with colleagues, teachers would get ideas on what to do to improve their teaching. From this, they would develop a new way of thinking about their teaching, leading to immediate action to improve practice.

Use of this model of consultation means that the following principles may result from the interaction: collaboration and collegiality as teachers talk about teaching with colleagues, problem-solving through dialogue, professional community, knowledge construction rather than knowledge transmission, and reflective practice. Evidently, group-based peer consultation holds considerable promise as an effective strategy to facilitate learning from student ratings feedback to improve teaching.

Significance of the Study

The recent criticism by Dowling (2002), himself a teacher, that the use of student ratings is nothing more than 'customer satisfaction' surveys that undermine genuine teaching and learning, should not be seen as hostility towards the use of student ratings. Neither should Harvey's (2001) finding that some teachers are hostile and cynical towards the use of student ratings, which they have derogatorily labelled 'happy forms', be taken out of context. Teachers do not reject the idea of collecting student ratings.

What has transpired is that, many academics have become discontented and somewhat frustrated because university administrations seem to put more emphasis on the use of student ratings for quality assurance purposes than for teaching improvement. Alongside this, is the criticism that administrators continue to overlook the need to locate teaching evaluation within a comprehensive system that could provide support to teachers to better understand how to use ratings feedback to bring about change in their teaching and what they do for students (Powney & Hall, 1998; Arreola, 2000).

In part, the problem lies with reduced budgets as already mentioned. Yet, it remains important for teachers to receive support to be able to use feedback ratings effectively, given the expectations that they increase the quality of instruction to students. Also, there is the need for new models and practices to help build and maintain collegiality and collaboration among teachers (Fullan & Hargreaves, 1992). It seems, then, that investigation into the potentials of teachers learning with and from peers in their department units, how to respond to feedback from student ratings to become more effective, is justified, timely, and is of great practical interest today.

This study is significant because the findings represent an important contribution to the research-based knowledge to the extent that group-based peer consultation over

student ratings has not previously been systematically examined. If found to be potentially useful, the findings will provide preliminary evidence that this model of consultation could represent a cost efficient yet effective approach to support university teachers to productively use the feedback from student ratings.

The results also hold many benefits for teacher professional development. Teachers may become more confident, take more responsibility for student learning, change their views about teaching and learning, habitually engage in reflective practice, and form the habit of experimenting with new ideas. The findings of the study could also lead to benefits for departmental and course teams. This model of consultation also offers an opportunity for teachers to learn how to work together to improve practice. This could lead to developing a culture of collaboration in which supporting colleagues and the sharing of knowledge and ideas within departments become a common and expected process. This could certainly help to break the norm of isolation and individualism in teaching.

Additionally, the study has significance for administrators and staff developers as they become more aware of the need to support teachers to facilitate continuous quality improvement, and decide how best to plan, implement, and invest resources to provide the conditions that will help teachers teach better. Perhaps most importantly, though, the results of this study could help to provide a real chance to improve the quality of learning for students so that they will be able to experience the 'higher' education promised.

Research Questions

This research was designed to answer the following questions:

1. What consultation practices and strategies are most important to maximise the effect of consultation over student ratings on teaching effectiveness?

2. What is the effect of augmenting feedback from student ratings with consultation in peer support groups on the teaching effectiveness of university teachers?

3. To what extent is the effect of ratings feedback and consultation in peer support groups related to university teachers':

   a. demographic characteristics in terms of age, gender, and teaching experience, and

   b. teaching responsibility in terms of course level and the discipline in which instruction is offered?

4. a. Do university teachers perceive the use of ratings feedback in peer support groups to be a practical strategy to improve teaching effectiveness?

   b. What are the factors that might inhibit participation in peer support groups to learn from student ratings?

   c. What conditions do teachers perceive to be important for effective use of consultation in peer support groups?

CONCEPTUAL FRAMEWORK

Improving teaching effectiveness is about the professional development of teachers, which involves enhancing knowledge, skills, attitudes, beliefs, and behaviours to be able to bring about improvement in classroom practice. A criticism levelled at teacher professional development activities is that they are not, for the most part, theory-based (Webb, 1996; Knapper & Piccinin 1999). So too, consultation over student ratings, as a professional development strategy, is not grounded in a theory of teacher professional development or even a theory of feedback intervention.

One explanation for the lack of theoretical base may be that professional development activities are more concerned with improving teaching practice and aiding policy decisions. Moreover, with limited funding the focus is to demonstrate what is useful to help teachers learn and bring about changes in their practice rather than on establishing theoretically sound programmes (Kember & McKay, 1996). There is also the assumption, as pointed out by Richardson and Placier (2001), that teachers, if shown that a new practice is good will act in their rational self-interest and make appropriate changes. Such programmes do, however, draw on several educational and organisational theories and concepts. In the same way, this present investigation focuses on evaluating group-based peer consultation as an alternative model to individual consultation over student ratings. That means the study is set in the context of an empirical research, that is to say, the collection and analysis of empirical data to provide evidence of the effectiveness of group-based peer consultation.

Although not theory-driven, the study is set against the broad concepts of learning and change through a feedback intervention to facilitate change in classroom practice for improving quality in teaching and student learning. Clearly, then, a limitation of the

present study is that it does not offer a comprehensive and critical review of the theoretical work in these areas but this is believed to be beyond the limits imposed for this thesis. From this perspective, the identification of some related theories should suffice.

Experiential learning theory as posited by David Kolb that views learning as, "the process whereby knowledge is created through the transformation of experience" (Kolb, 1984, p. 38) is applicable to teacher professional learning. Teachers are believed to have a wealth of professional experience that is a valuable source of learning to improve practice. The theory importantly highlights the need to be able to reflect on past and present experience in order to produce learning and change in practice.

Reflective practice, popularised by Donald Schön but first advanced by John Dewey (1933), is particularly useful and is concerned with the careful and deliberate thinking about one's practice. For Schön (1987) reflective practice is a dialogue of thinking and doing through which one becomes more skilled. Schön (1983, 1987) also suggested that learning to improve practice requires that teachers as professionals not only become aware of research-based knowledge but also have the opportunity to experience 'reflection in action' and 'reflection on action' as important forms of knowledge.

Use of action research methods provides a good support structure for teacher learning with supportive colleagues. It involves teachers inquiring into their own practice and producing practical knowledge to improve teaching in collaboration with their peers. As such, it is described as a social, collaborative, reflective, practice improvement-oriented, active, and educational inquiry into practice (Zeichner, 2001; Kemmis & McTaggart, 2000; Zuber-Skerrit, 1996). One form of action research, action learning, is also considered appropriate for

teacher learning in small groups. As suggested by Beaty (McGill & Beaty, 1995; Beaty, 1999) action learning, is a problem-solving process supported by colleagues in a structured group setting that facilitates learning and reflection from experience.

Another particularly useful theory to link teacher learning and change to student ratings feedback is situated learning theory proposed by Jean Lave. The theory argues that learning how to improve practice occurs not as transmission of a set of facts and rules but as a function of the activity, context, and culture in which practice occurs (Lave & Wenger, 1991; Lave, 1993). Learning is social and context related. Central to situated learning is social interaction, participation, and talk in a community of practice (Wenger, 1998; Wenger, 2000). Wenger and Snyder (2000) define a community of practice as a "group of people informally bound together by shared expertise and passion for a joint enterprise" (p. 139). Teachers within departments can become a community of practice if given the opportunity to sustain mutual engagement in their common enterprise and share some significant learning about their practice and how to improve it.

Because teachers are adult learners the principles of adult learning should also be considered when dealing with teachers' professional learning. For example, Knowles' (1980) model of andragogy, defined as "the art and science of helping adults learn" (p. 43). Andragogy, according to Knowles (1980) reflects the adult learner's needs to be a self-directed learner. In this way, the assumption is that teachers are autonomous individuals who desire self-management of their learning about teaching. Related to the principle of andragogy is the theory of transformative learning, forwarded by Jack Mezirow. According to this theory, learning is a process of becoming aware of one's assumptions and revising these assumptions through critical self-reflection

(Mezirow, 1991). The idea is that adult learners have beliefs, assumptions, and values that determine the way they interpret their experiences, which may be challenged allowing transformative learning to come about.

Teacher professional development involves change—change in understandings, beliefs, and classroom practice. Theories about change are therefore useful to inform the design of any feedback intervention. Argyris and Schön's (1974) posited the concept of theories-of-action, which include the underlying assumptions that underpin an individual's general action. Argyris and Schön have suggested that increasing professional effectiveness involves the building and rebuilding of one's theories-of-action that is at two different levels, single-loop learning and double-loop learning. Simply stated, in single-loop learning professionals are likely to learn about new skills and strategies to achieve existing goals and beliefs. Double-loop learning on the other hand, entails a fundamental change in terms of goals and beliefs that allows for professional growth.

The change theory from the seminal work of Kurt Lewin (1951) that considers the process of change as a three-stage process of unfreezing, changing, and refreezing, may also be considered as a useful model for helping staff developers make sense of planning and implementing interventions to facilitate improvement in classroom practice. The first stage, unfreezing, begins with recognition that a situation or behaviour is in some way inadequate. The changing phase occurs when a new system or plan is implemented. The third stage, refreezing, occurs when newly created process or patterns of behaviour and techniques become the new mode of practice.

Another but more contemporary approach is the three-phase approach of initiation, implementation, and institutionalisation, to planning and managing educational

change as suggested by Fullan (2001). The initiation phase consists of the process that leads up to and includes the decision to adopt a change. The implementation phase involves the initial experiences of attempting to put the change idea into practice. The third stage in the change process, institutionalisation, is concerned with whether the change gets built in as an ongoing part of the system or disappears because of the decision to discard it or through attrition. The three phases, noted Fullan (2001), should be seen in relation to some outcome, particularly whether student learning is enhanced, and whether the experiences with the change idea increase the capacity to deal with future changes. Fullan is also careful to point out that this approach is not a linear process but one in which events at one phase can feedback to alter decisions made at a previous stage, which then proceed to work their way through in a continuous interactive way.

It is important to note, as pointed out by Richardson and Placier (2001) that a "one solution" conception of change, top-down mandates that ignore local contexts, and change that threaten teachers' sense of autonomy no longer holds. Deep and lasting change should be viewed as an ongoing and local process that requires consideration of a multitude of aspects and interests.

Chapter 2

Student Ratings of Teaching: Valid and Reliable?

This chapter discusses issues pertaining to the validity and reliability of student ratings of teaching by looking at the research evidence. The chapter opens with an attempt to answer the question "what is effective teaching?" followed by a look at the purposes of student ratings. The chapter then presents a concise review of the extensive research literature on the validity and reliability of student ratings of teaching.

Introduction

Modern use of student ratings of teaching goes back to the mid-1920s in a limited number of North American universities as one way to improve teaching. This was closely followed by research into issues of reliability, validity, halo effects, and bias in 1927 by Herman Remmers, now seen as the Father of Student Evaluation Research (Marsh, 1987; Centra, 1993). Today, the practice of collecting student ratings

is almost commonplace in universities throughout the world but its use is not without disputatious discourse among academics. Many academics are still not convinced student ratings are valid indicators of teaching effectiveness.

The research literature on student ratings of teaching—also called students' evaluations of teaching (SETs), teacher course evaluations (TCEs), student ratings of teaching effectiveness (SRTEs), student evaluation of faculty performance (SEFP), or teacher ratings forms (TRFs), to name a few—is voluminous, and the subject is described as the most researched and most disputed topic in higher education research over the last three-quarter century or so (e.g. Theall & Franklin, 2001).

Even so, in recent times emphasis on evaluative information on teaching has once again placed student ratings at the centre of a lively debate and active research on their validity, with original studies and interpretation of findings (e.g. Kwan, 1999; Kerridge & Matthews, 1998; Greenwald & Gillmore, 1997; Shevlin, et al. 2000; Kember, Leung & Kwan, 2002). This renewed debate coincides with the pressure on universities to improve the quality of student learning with implications for teaching.

What is Teaching Effectiveness?

Fundamental to the controversial use of student ratings in assessing teaching are the persistent questions; What is effective teaching?, How can it be measured?, Can students assess teaching effectiveness?, Are student ratings related to learning? As pointed out by Abrami, d'Apollonia and Cohen (1990), student ratings are rarely criticised as measures of student satisfaction but routinely so as measures of teaching effectiveness. A definition for effective teaching is problematic in that, effective

teaching is a hypothetical construct but it is perceived to be multidimensional, complex, and dynamic, for which no one single measure can assess it (Marsh & Dunkin, 1992).

Various researchers have indeed made attempts to explore the concept of teacher effectiveness but there is as yet no universally agreed and applied definition. Insofar as there is no agreement it is difficult to give a precise definition of effective teaching. For this reason, a common argument is that we do not know what constitutes effective teaching, so we are unable to evaluate it properly (e.g. Patrick & Smart, 1998; Adams, 1997). Seldin (2000) sees it differently. He contended that the benchmarks of effective teaching have been reasonably and consistently identified through numerous research studies.

So too, McKeachie (1997a) argued that research has generated enough knowledge on effective teaching to facilitate its assessment, at least on those aspects that can be captured with student ratings forms. McKeachie (1997a) went on to say that teaching effectiveness should be defined in terms of the goals for the particular context of teaching. In this sense, effective teaching is more appropriately judged in terms of its intended goals and progress toward such goals over time.

In the context of higher education there is near unanimity that effective teaching provides a learning environment that leads students to discover and construct knowledge, enabling them to become critical and creative thinkers with the capacity to go on learning. It is not teaching that merely transmits knowledge causing students to passively absorb information (Ramsden, 1992; Light & Cox, 2001; Barr & Tagg, 1995).

Biggs (2001), for example, described quality or effective teaching as, "… teaching that will transform students' perception of their world, and the way they go about applying their knowledge to real world problems" (p. 222). This is the sort of

teaching and learning that is believed to be needed to prepare students for the uncertainty, unpredictability, challenging, and super-complexity that characterises the world in which they will find themselves after university (Barnett, 2000).

There is no unanimity, however, on the characteristics of effective teaching or the behaviours that good teachers engage in. A major review of the research literature on the characteristics of effective teaching undertaken by Ramsden, et al. (1995),

included the following statements:

- Good teachers are also good learners; for example they learn through their own reading, by participating in a variety of professional development activities, by listening to their students, by sharing ideas with their colleagues, and by reflecting on classroom interactions and students' achievements

- Good teachers display enthusiasm for their subject, and a desire to share it with their students.

- Good teachers recognise the importance of context, and adapt their teaching accordingly; they know how to modify their teaching strategies according to the particular students, subject matter, and learning environment.

- Good teachers encourage deep learning approaches, rather than surface approaches, and are concerned with developing their students' critical thinking skills, problem-solving skills, and problem-approach.

- Good teachers demonstrate an ability to transform and extend knowledge, rather than merely transmitting it; they draw on their knowledge of their subject; their knowledge of their learners, and their general pedagogical knowledge to transform the concepts of the discipline into terms that are understandable to their students.

- Good teachers set clear goals, use valid and appropriate assessment methods, and provide high quality feedback to their students.

- Good teachers show respect for their students; they are interested in both their professional and their personal growth, encourage their independence, and sustain high expectations of them.

Yet, although having a list of qualities is probably useful and relevant, it does make it pretty clear that teaching is indeed multidimensional and complex, and that evaluating teacher effectiveness is no easy task.

Looking beyond the traditional view of teacher effectiveness with its focus on observable classroom behaviours, Trigwell (2001) posited a contemporary model of assessing university teaching. In this model, defining and assessing teaching effectiveness would account for all the processes underlying the teaching act. The view is that the elements of the teaching process—teaching strategies, planning, teacher thinking, effects of the wider teaching/learning context, and the impact on student learning—are connected and should be seen as such.

For Trigwell (2001), this model serves to develop a holistic view of teaching, making it easier to identify and judge the quality of teaching. Trigwell's model might indeed allow for a more comprehensive approach to evaluating teaching but it does seem more applicable for peer review of teaching and the use of teaching portfolios, where the different elements could be more readily assessed. Yet, the model draws attention to the need to broaden the scope of teaching evaluation practices to include other aspects of the teaching activity apart from that of classroom behaviour, which is the only aspect that most student ratings forms seem to capture.

What Do Student Ratings Measure?

Marsh and Dunkin (1992) summarised the views in the research literature when they stated that no single criterion measure satisfies the notion of effective teaching. In the absence of a single criterion of effective teaching, researchers have been content to adopt the construct validation approach, demonstrating that student ratings correlate with other indicators of effective teaching. The most widely accepted criterion of effective teaching is student learning. In this way, student ratings are used to evaluate teachers on how well they have induced learning.

It is important to keep in mind, though, that student ratings do not measure learning per se. Rather, they are simply proxy measures of learning, representing student perceptions of learning from the instruction received. In other words, ratings feedback reflects student perceptions of good teaching and their satisfaction with the teaching in a course. This perception and satisfaction, is of course, influenced by an interplay of factors including the student's expectations and motivation, attitudes, abilities and effort, the teacher's instructional techniques, as well as the institution's facilities and services (Neumann & Neumann, 1981; McKeachie, 1997a).

Student Learning

An established finding in student feedback research is that student ratings are positively and moderately related to student learning, inferred to effective teaching. This relationship has been investigated through a number of studies using multisection validity designs. In the strongest multisection design, different sections of the same course are taught by different teachers, using the same syllabus and textbook, and assessed with a common final examination. The average of ratings are then correlated with the average exam scores to determine if high ratings are related to more learning, higher exam scores. Comparisons of effective teaching are made on the basis of the level of learning that can be related to student ratings.

Despite methodological complications of the design (Marsh & Dunkin, 1992; Abrami, d'Apollonia & Cohen, 1990; Abrami, 2001), the results of meta-analytic reviews of these multisection validity studies do show that student learning correlate, moderately to highly, with student ratings. Students tend to give higher ratings to teachers from whom they learned more, and give low ratings to teachers from whom

they learned less (Cohen, 1981, 1987; Feldman, 1989b). This finding has been broadly accepted as providing strong support for the validity of student ratings.

To illustrate, in Cohen's (1981) meta-analysis of 41 studies across 68 multisection courses useful correlations were found: Teacher overall rating, $r = 0.43$, Course overall rating, $r = 0.47$, Skill, $r = 0.50$, and Structure $r = 0.47$. A critical analysis and re-analysis (Cohen, 1987) also suggested that the relationship between learning and student ratings were stronger for multi-item scales than for single-item ratings. Koon and Murray (1995) also found correlation of $r = 0.41$ between student ratings and overall teacher rating for learning on an objective test.

Cohen's (1981) and Feldman's (1989b) findings were supported by d'Apollonia and Abrami (1997a) who on review of an earlier meta-analysis of 43 multisection studies (d'Apollonia & Abrami, 1996), suggested that under appropriate conditions, more than 45% of the variation in student learning could be explained by student perceptions of teaching effectiveness. The conclusion drawn is that there is reasonably strong support for the validity of student ratings as an indicator of teaching effectiveness.

The use of learning as a meaningful criterion in higher education is however problematical (Feldman, 1997; Braskamp & Ory, 1994). First, an association between student ratings and learning does not establish causality. A third variable such as student motivation, ability, or attitudes and even the amount of independent work undertaken, may influence learning. Second, this correlation does not mean that student ratings will consistently differentiate instructors who facilitate high and low levels of learning, as demonstrated with in the ''Dr. Fox'' type studies. The ''Dr. Fox'' studies

will be explained in a later section. In short, student ratings are not perfect estimates of the effect of a teacher on student learning.

Third, the assessment of learning in higher education may not be a practical indicator of teaching competence as standards are not comparable across institutions, departments, or even courses. Moreover, higher education is concerned with the development of higher order thinking skills whereas the assessment of learning in multisection studies relied on objective tests, which tapped lower level skills such as knowledge and comprehension. On this basis, learning as a criterion of teacher effectiveness is probably more suited for use at the level of the individual teacher. This should provide good feedback on the extent to which students are achieving course goals and objectives.

Evaluation by Different Evaluators

Because student learning is an imprecise measure, ratings of teaching effectiveness from other measures have also been investigated. Researchers emphasise that teaching effectiveness should be measured from multiple perspectives. The multitrait-multimethod (MTMM) design is used to assess the extent to which student ratings correlate with other measures such as, alumni ratings, teacher self-ratings, ratings by external observers, administrators, and ratings by peers. Use of the MTMM method is based on the principle that two measures of the same construct should be highly correlated. That is, they should have convergent validity to indicate that they measure the same thing. Another condition of the design is that unrelated measures should not be highly correlated, this means they should have discriminant validity.

In a meta-analysis of these MTMM studies that investigated the relationship between student ratings and these other measures, Feldman (1989a) found evidence of

moderate to high positive correlations between student ratings and alumni ratings (r = 0.69), ratings from trained observers (r = .50), teacher self-ratings, (r = .29), colleagues (r = .55), and administrators (r = .39). Although ratings from students might agree with that from administrators, teacher self-ratings, and peer ratings these measures are viewed as less satisfactory criteria of teaching effectiveness (Kulik, 2001).

Qualitative (written or group interview) evaluations are also highly correlated with quantitative ratings (Braskamp & Ory, 1994).

In summary, research has established that student ratings are strongly associated to learning and so the ratings represent a reasonably good measure of teaching effectiveness. The recommendation, however, is that the ratings represent only one source and should be combined with other sources of information for a more accurate indication of overall teaching effectiveness.

How Are Student Ratings Used?

Student ratings have become synonymous with teaching evaluation. Although methods such as student focus group and written evaluations could be used to collect student views about teaching, individual student ratings have turned out to be the most commonly used source of information to evaluate teaching. This is largely because they are seen as relatively more objective and trustworthy than assessment from administrators, peers, or even teacher self-evaluation; students are convenient sources; and the ratings are relatively easy to collect and interpret. Put simply, they are fast, fair, and cheap (Arreola, 2000; Seldin, 1999a).

In discussing the reasons for collecting student ratings, four common purposes are cited (Marsh, 1987; Cashin, 1999; Feldman, 1997):

- To provide diagnostic feedback to teachers that will be useful for teaching improvement;
- As one measure of teaching effectiveness for promotion and tenure decisions;
- To provide information for course selection by students; and
- As an outcome measure for research on teaching.

In recent times a fifth purpose has been discerned, to provide information for external monitoring of teaching quality in universities, putting student ratings on the political agenda. To illustrate, in its recent White Paper, "The Future of Higher Education" (Department for Education and Skills, 2003), the UK government made clear its intention to give higher priority to the use of student feedback to inform on quality and to link funding. Similarly, in Australia the Course Experience Questionnaire is used to collect feedback from recent graduates about their learning experience and monitor quality in all publicly funded universities.

The first two purposes are easily seen as the primary purposes of student ratings, that is, for formative and summative evaluation, or improvement and accountability. The purposes co-exist but do get in each other's way from time to time as data intended for teaching improvement purposes may be inappropriately used by administrators for accountability regarding personnel, courses, department units, and the institution (Johnson & Ryan, 2000; Stake & Cisneros-Cohernour, 2000). Different types of information are needed to address the different purposes appropriately.

For the summative evaluation function student ratings data are used by department chairs, promotions and tenure committees, and deans as supporting evidence to arrive at decisions concerning teaching effectiveness for salary increments, tenure, renewing or terminating contracts, promotion decisions, assigning teachers to courses,

selecting teachers for awards and honours, as well as a performance indicator in external quality monitoring. For this purpose summary data from multiple sources for more than one course over an extended period of time should be considered. Recognising that the nature of effective teaching could vary across teachers, courses, classes, students, and contexts, Abrami (1989) recommended that global measures (the overall evaluation items) are best used for summative judgements about teaching quality.

For formative evaluation, student ratings represent diagnostic information to teachers that can serve as a basis for reflection, identifying strengths and weaknesses, and to plan possible changes in courses and teaching behaviours. Teachers generally approve of this use of student ratings (Baxter, 1991; Murray, 1997). By contrast, frustration, cynicism, and even hostility are directed towards the use of ratings for summative evaluation. The concern of teachers is the way in which ratings are often misused and misinterpreted by administrators.

Anecdotal reports abound of administrators using ratings quite selectively, over-interpreting and over-relying on the data, classifying and ranking teachers rather than assessing teaching effectiveness (e.g. Adams, 1997; Sproule, 2000). Many administrators simply lack the skills and knowledge needed to use the ratings properly (Franklin & Theall, 1989). Taken together, accounts of how ratings are generally used in practice have led to the view that criticisms about the validity and reliability of student ratings are due mainly to the inappropriate use by administrators than any potential biases in the ratings themselves (e.g. Theall & Franklin, 2001; McKeachie, 1997b).

Are Student Ratings Valid?

Student ratings as a valid indicator of teaching effectiveness are questionable if there is reason to suspect they are influenced by factors unrelated to what the teacher does, and so might not measure what they are intended to measure. Validity is crucial to the use of student ratings. The conclusion from meta-analyses, research studies, and comprehensive reviews of the literature is that student ratings are relatively reliable and valid indicators of teaching effectiveness, provided they are carefully collected, interpreted, and used appropriately (e.g. Marsh, 1987; Cohen, 1981, 1983; Kulik 2001; Koon & Murray, 1995; McKeachie, 1997a, 1997b; d'Apollonia & Abrami, 1997a; Feldman, 1978, 1979, 1984, 1988, 1989a, 1989b, 1997; Marsh & Roche, 1997, 2000; Cashin, 1995; 1999; Braskamp & Ory, 1994; Wachtel, 1998; Centra, 1993; Marsh & Dunkin, 1992; Abrami, 2001; Arreola, 2000).

These proponents of student ratings have insisted that student ratings are not perfect measures but they are not greatly influenced by extraneous factors. Or, as d'Apollonia and Abrami (1997a) put it, "...student ratings of instruction are not plagued with biasing variables" (p. 1203). The ratings, it is maintained, represent a valid and useful source of information on teaching for teachers, students, and administrators. Besides, there is little evidence of the validity of any other source of data on teaching effectiveness.

This view does not go unchallenged. Critics of student ratings such as Dowell and Neal (1982) have suggested that the evidence is weak and inconclusive to support claims that student ratings are good measures of teaching effectiveness. Others question validity in terms of students' ability to objectively judge teaching performance

and the curriculum when their primary concern is that of getting high grades and easy courses rather than learning itself (e.g. Stone, 1995).

Some critics have also argued that ratings are based on teacher popularity, teacher expressiveness, and teacher charisma, rather than on the substance of teaching itself (Shevlin, et al. 2000; Wilson, 1998; Williams & Ceci, 1997). There is also the argument that ratings are not valid as there is no confirmation of the underlying construct that ratings forms are measuring, and are influenced by variables such as course characteristics, student characteristics, and grading policies (Sproule, 2000; Husbands & Fosh, 1993; Kwan, 1999; Trout, 2000).

The concerns for validity are discussed within the conceptual framework of four types of validity: construct, convergent, consequential and discriminant (Ory & Ryan, 2001). Discourses on the different types imply different answers to questions about the use of student ratings as an assessment device and as an improvement tool. There is a level of agreement on issues pertaining to construct validity, the degree to which student ratings accurately measure the construct teaching effectiveness; convergent validity, the extent to which ratings correlate with other indicators of effective teaching; and consequential validity, the effective use of ratings in teaching improvement. More disputed, however, is the issue of discriminant validity, that is, the degree to which ratings are affected by variables unrelated to effective teaching. Greenwald (1997) criticised that the literature tends to treat student ratings as if their discriminant validity is perfect, that is, as if they are not affected by biasing factors.

What is bias in student ratings? There are different views on what constitutes bias. One common definition is that it is anything not under the control of the instructor

that may influence the ratings (Cashin, 1995). Marsh (1987) rejected this definition as too broad and suggested instead that bias be seen as variables that influence ratings but which are not related to teaching effectiveness. Typically, the literature views bias as variables related to student ratings but not teaching effectiveness that might contaminate the ratings. Much research energy has been directed at examining the extent to which bias is introduced from variables related to the characteristics of students, teachers, courses, and even how ratings are administered and for what purpose.

The literature on student ratings feedback research is voluminous. Drawing on the work of researchers, reviewers, and meta-analysts mentioned at the start of this section, a brief summary of the more pertinent findings is presented here.

Student Variables

It is argued that effective teaching may be influenced not only by the competence of the teacher but also by the characteristics of the students. Table 2.1 presents findings for the relationship between student variables and ratings.

The debate on the influence of student variables is not very controversial except for the relationship between student ratings and expected grade. The biasing effect of grades on student ratings has received considerable attention from researchers. There are findings of a positive significant relationship between expected course grades and student ratings (r = .10 to .30). Ratings tend to be higher for teachers when students know their final grades before making ratings but there is no firm conclusion whether the source of the bias lies with the ratings or the student themselves. This effect has been interpreted as a grading leniency bias.

Table 2.1
Relationship between Student Characteristics and Student Ratings

Variable	*Finding*
a. Student motivation	Students with prior interest in the subject give higher ratings. Students who selected a course as a major elective also tend to higher ratings (r = .40).
b. Expected grade	Students expecting higher grades tend to give more favourable ratings (r = .10 to .30)
c. Gender	No significant difference in gender and overall ratings but there is slight tendency for a same-gender preference (r = .03).
d. Personality	No meaningful and consistent relationship exists between personality traits and ratings.

Greenwald and Gillmore (1997) offered five possible reasons for the grades-ratings relationship:

- *Teaching effectiveness influences both grades and ratings*: quality instruction facilitates better learning therefore higher ratings. This relationship supports the validity of student ratings.

- *Students' general academic motivation influences both grades and ratings*: highly motivated students work harder and should get higher grades and in turn will give higher ratings.

- *Students' course-specific motivation influences both grades and ratings*: probably on the basis of prior interest or choosing the subject as an elective

- *Students infer course quality and own ability from received grades*: high grades will be attributed to intelligence and diligence and low grades will be attributed to poor teaching.

- *Students give high ratings in appreciation for lenient grading*: teacher gives high grades, especially if greater than expected, in turn students reward this generosity with high ratings. This constitutes a bias and a threat to the validity of student ratings.

The grading leniency argument is further discussed under the section teacher variables.

More recent studies on the influence of student characteristics have looked at ethnic background, age, and academic status but have reported that even where differences are statistically significant they are not large enough to make much difference in personnel decisions (Young & Shaw, 1997; Worthington, 2002; Centra & Gaubatz, 2000).

Another perspective of the effects of student characteristics is provided by Kember and Wong (2000) who suggested that students' conception of learning might be a biasing factor. Using a qualitative research design, the researchers found that students with a passive approach to learning are likely to give lower ratings to teachers who adopt student-centred approaches that might require them to think and develop meaning and understanding for themselves. This point is further elaborated in Chapter 3.

Teacher Variables

The objective of teaching evaluations is to assess the extent to which the teacher stimulated learning as perceived by students. The influence of teachers' reputation, ranks, research activities, and teaching experience on ratings have also been examined as possible biases. Table 2.2 summaries the findings.

Of all teacher related variables, grading leniency is perhaps the most controversial. The argument of supporters for a lenient grading bias (e.g. Stone, 1995; DuCette & Kennedy, 1982; Greenwald & Gillmore, 1997; Redding, 1998; Worthington & Wong, 1979), is underpinned by the assumption that teachers can "buy" high ratings by offering less demanding courses and grading more leniently. The likelihood that this might indeed be happening is seen as the cause of grade inflation, a primary contributor to declining academic standards, and the "dumbing down" of the curriculum.

Table 2.2
Relationship between Teacher Variables and Student Ratings

Variable	*Finding*
a. Rank and experience	Regular teachers tend to receive higher ratings and first-year teachers receive lower ratings than those with more seniority (r = .10) but the relationship is not consistent.
b. Personality	Few personality traits correlate with student ratings but positive for self-esteem (r = .27) and energy and enthusiasm (r = .30).
c. Gender	No significant relationship exists between teacher gender and overall ratings although students might rate same gender teachers a bit higher than opposite gender (r = .02)
d. Research productivity	Weak positive correlation with student ratings (r = .12)
e. Expressiveness (The ''Dr. Fox'' effect)	Highly expressive lecturers may receive higher scores on global ratings but this is not a bias if it helps students to learn.
f. Grading leniency	No consistent findings that giving higher grades will result in higher ratings.

Although many other researchers agree with Marsh and Roche (1997, 2000) that the grading leniency effect is at best weak, there are some who conclude otherwise. In their oft-cited study among 200 courses across three consecutive terms, Greenwald and Gillmore (1997) found that teachers who get high ratings tended to give out higher grades, whereas teachers who get low ratings give out lower grades. Students also seemed to work harder in classes where teachers give low grades and less hard in classes where teachers give high grades. This finding was interpreted to mean that the use of student ratings can produce lenient grading and that higher grades may not be a reflection of better teaching and more learning.

Marsh and Roche (2000) re-analysed Greenwald and Gillmore's data along with data from the Students' Evaluation of Educational Quality (SEEQ) questionnaire over a 13-year period. These researchers explained that workload is positively related to student ratings, in that the more difficult and challenging courses that require more hours are evaluated more favourably. This makes a workload bias untenable. Also, grade/ratings correlations are modest which varies from $r = 0$ to .30 on individual ratings factors which does not support a simple grading leniency bias. Marsh and Roche (2000) concluded, "there is absolutely no support whatsoever for doom and gloom implications derived from grading leniency and workload bias theories" (p. 222).

Nonetheless, Greenwald and Gillmore (1997) called for statistical adjustment of ratings, to take account of grading leniency and other variables such as class size, when ratings are to be used for promotion and tenure decisions. Even against the background that some teachers might choose to inflate grades to get high ratings, other researchers have dismissed the idea of adjustment (d'Apollonia & Abrami, 1997a; Marsh & Roche, 1997). The reasoning is that it would be introducing bias of another kind, unfair removal of elements of the valid influences of good teaching.

Instead of statistical adjustment, Cashin (1995) usefully proposed that peer review of course materials and assessment practices should be used as a control for lenient grading. Redding (1998) and Eiszler (2002) suggested that institutions should adjust their practices and incentive systems. Here, it is argued that administrators' heavy reliance on student ratings in decision-making about promotion and merit raises for example, actually tempts teachers to inflate grades.

The other issue that has received much attention is the 'Dr. Fox' effect or the overriding influence of instructors' expressiveness on student ratings. The concern is

that, displays of enthusiasm, expressiveness, and entertainment can unduly influence students into giving favourable evaluations regardless of content, as first suggested by the much discussed study by Nauftulin, Ware, and Donnelly (1973, in Marsh 1987). This study was later shown to be flawed and was dismissed on the grounds of methodological weaknesses but led to many others.

More recently, Williams and Ceci (1997) followed a similar line to show that ratings data were not biased-free indicators. They attempted to show that being more enthusiastic and expressive could result in high ratings even though students might not have actually learned more. In response, d'Apollonia and Abrami (1997b) criticised the study for its faulty design including the lack of proper controls. The study was dismissed by the greater part of the research community but is popularly cited by critics of student ratings.

Analyses of the earlier studies using the 'Dr. Fox' paradigm (Marsh & Ware, 1982, in Marsh, 1987; Abrami, Leventhal & Perry, 1982) found that the 'Dr. Fox' effect was exaggerated. It was recognised, though, that enthusiasm and expressiveness have a strong influence on the ratings of instructors' enthusiasm but not the other dimensions of teaching. The conclusion is that students are able to tell the difference between style and substance and that expressive teachers act in ways that positively influence learning so it is not a biasing factor (Murray, et al. 1990).

Course Variables

Because student ratings reflect students' perception of the teaching-learning interaction, the influences of class and course characteristics have been considered as possible biases. Table 2.3 presents a summary of the research findings.

Table 2.3
Relationship between Course Characteristics and Student Ratings

Variable	*Finding*
a. Class size	Some studies have found a small tendency for larger classes to receive lower ratings but this relationship is not consistent.
b. Course level	Advanced courses tend to receive higher ratings than beginning courses but the difference is small (r = .07)
c. Class meeting time	No consistent relationship between time of day when the course is taught and student ratings
d. Academic field	Weak tendency for mathematics, physical and natural sciences, and engineering to receive lower ratings than humanities, arts, and the social sciences.
e. Electivity	Teachers for elective courses are likely to receive higher ratings than those for required courses
f. Workload/difficulty	Positive but weak correlation. High ratings associated with difficult courses and with higher levels of workload/difficulty (r = .11 to .29).

Among the course variables, the effect of class size on student ratings is the most widely studied. Possible reasons for a class size relationship is that smaller classes more than larger classes, tended to provide a better opportunity for students to interact with teachers, receive more individual attention, receive written work, and engage in interactive teaching sessions, all of which contribute to better learning.

In this sense, class size is not biasing the ratings but simply reflecting increased learning. But should this be interpreted as 'effectiveness'? Class experience is also significantly influenced by the nature of the course. Ratings are generally higher for advanced-level than for introductory-level courses. The explanation is that, with a low

level of difficulty, students might become easily frustrated by the course material and exhibit low satisfaction with the class. At the same time, greater difficulty might act as a motivator for some students, causing them to work harder and thus learn more from the course (Shiparo, 1990).

Recent research by Kwan (1999) showed that the effects of academic discipline and class size are larger than that of course level, mode of study, and type of course requirement. In another study, Santhanam and Hicks (2002) found that students in the arts/humanities/social sciences rate their courses higher than students in the sciences/mathematics. The ratings in higher year level courses were also more positive than lower year levels. The implication of these findings is that for summative evaluation ratings data cannot simply be used to compare teachers without due consideration to contextual factors such as the discipline, year level, or even class size.

Administration Procedures

The circumstances under which ratings are collected have also been examined as possible sources of bias (Table 2.4). The general conclusion in the literature is that the effects of factors associated with administration are not large. Nonetheless, it is good practice to follow a standard set of procedures when administering ratings questionnaires.

Table 2.4
Relationship between Administration Procedures and Student Ratings

Variable	*Finding*
a. Time of administration	Little effect on ratings although ratings may be lower if collected during final examination period.
b. Student anonymity	Ratings are more positive if students are required to identify themselves than ratings provided anonymously.
c. Teacher's presence	Ratings are more favourable if the teacher remains in the room.
d. Stated purpose	Ratings are more favourable if the stated use is for promotion.

To summarise, the dominant view in the student ratings literature is that the ratings are not unduly influenced by supposedly biasing factors. In fact, such factors account for no more than 12 to 14 percent of the variance in the ratings (Marsh, 1987). In this way the ratings represent valuable indicators of teaching effectiveness and can provide useful information to both academics and administrators, with the caution that they should be used as only one of several measures in assessing teaching performance.

Are Student Ratings Reliable?

Unlike the discourse on the validity of student ratings, there is probably unanimous agreement that student ratings are quite reliable. Reliability indicates the extent to which ratings items consistently measure the same component of effective teaching. It is a pre-condition for validity. For student ratings the measures; a) interrater reliability or consistency, b) the stability of ratings over time, and

c) generalisability of ratings are all important in considering the reliability of ratings (e.g. Centra, 1993; Cashin, 1995; Braskamp & Ory, 1994).

Interrater Reliability or Consistency

The general agreement in ratings feedback research is that interrater reliability, or agreement among different students within the same class rating the same instructor—rather than consistency among responses on questionnaire items by individual students, internal consistency—is the most appropriate estimate of reliability. As reliability varies according to the number of raters, the rule of thumb is to have at least ten raters to obtain the minimum acceptable estimate of reliability of $r = 0.70$, using well-developed ratings questionnaires.

Estimated reliabilities for standardised instruments such as the Students' Evaluations of Educational Quality (SEEQ) are $r = 0.95$ for average class response from 50 raters, $r = 0.90$ from 25 raters, $r = 0.74$ from 10 raters, $r = 0.60$ from 5 raters and $r = 0.23$ from 1 rater (Marsh, 1987; Marsh & Dunkin, 1992). Similarly, estimates for the Instructional Development and Effectiveness Assessment (IDEA) form are approximately $r = 0.70$ with 10 students, $r = 0.80$ with 20 students and $r = 0.90$ with 40 students (Cashin, 1999).

Commenting on the consistency with which student ratings measure the construct teaching effectiveness, Marsh and Dunkin (1992) wrote, "Given a sufficient number of students the reliability of class-average SETs compares favourably with that of the best objective tests" (p. 158). These claims of reliability, though, are based on well-developed questionnaires. Therefore, with many universities using "home-made" ratings forms, which generally have not demonstrated psychometric soundness to show

they are valid and reliable, it becomes difficult to counter criticisms that student ratings are not valid measures of teaching effectiveness.

Stability and Generalisability

Student ratings are also highly correlated from one administration to another. This means they have stability. Stability in student ratings is concerned with agreement between sets of ratings over time. Generalisability reflects how well the ratings assess the instructor's general teaching effectiveness, not just in a particular course or term. Referencing a number of studies in the research literature, Rindermann and Schofield (2001) for example, interpreted the findings of studies as indicating that the teacher has a stronger effect on student ratings than the course. The overwhelming evidence is that a teacher in different courses, with different topics and at different measurement times, teaches in a similar way. There are marked differences between instructors with respect to teaching effectiveness.

The ratings collected for an instructor over time have also been shown to be stable over courses, time, and across course levels, with average correlation $r = 0.83$. Research by Marsh (Marsh & Hocevar, 1991b; Marsh & Bailey, 1993) on the ratings of 195 teachers over a 13-year period demonstrated that teachers have distinct profiles, for instance, high on rapport but low on organisation. Although researched over the shorter period of two years, Hativa (1996) reported a similar finding and went on to suggest that this stability of profile might even be resistant to teaching improvement interventions.

An exception to this finding is Clayson's (1999) recent study, which offered only partial support for the stability of student ratings. Clayson (1999) instead found that students perceive instructors as changing in knowledge, fairness, and organization

skills over time but that it was the attribute relating to personality they perceive as stable. This research was however conducted in one subject area over a shorter time period and used an unstandardised instrument.

Referring to a number of early ratings feedback studies, Cashin (1995) recommended that for summative evaluation purposes ratings from an adequate number of students in at least five different courses from every term for at least two years for an instructor should be considered to get an accurate assessment of teaching performance. Where a teacher provides instruction for the same course for which he/she has been previously evaluated, use of results from at least two different offerings of each course should be taken into account.

In summary, the evidence from the research literature is that with use of well-developed and administered ratings forms coupled with correct interpretation, student ratings represent a fairly valid and reliable source of information on teaching effectiveness. Rindermann and Schofield's (2001) analogy to the drawing of a portrait seems to capture the essence of the argument that student ratings of teaching are reasonably valid and reliable:

> A portrait should never describe an exact picture of the outward appearance of a person, but it is more exact and near to reality than any attempt that can be made at imagining the outward appearance of that person based only on hearsay (p. 396).

Chapter 3

Shortcomings of Student Ratings Feedback Research

This chapter looks at the gap between research on ratings feedback and practice. It presents four perceived shortcomings of the research and calls for more relevant and practical research to explore how to increase the usefulness as well as knowledge and skills of users of ratings data. The chapter concludes by pointing out that the continued poor practice in using ratings data coupled with complaints from teachers, administrators, and students is a clear indication that the research agenda needs to be reconfigured.

Introduction

Almost eighty years on from the first use and early study into the use of student ratings in universities, the issue remains one of the most sensitive, divisive, and political in higher education (Theall & Franklin, 2001). Any mention of student ratings of

teaching is likely to lead to a discussion on problems associated with their validity, in particular, often overlooking their potential usefulness.

Ironically, while the use of student ratings is increasing, supposedly with the aim to improve teaching, research on how practitioners can use the data more effectively remains very limited. A great deal of research attention continues to be centred on providing 'evidence' that ratings data are biased, even in the light of extensive empirical evidence that they are not unduly so. This search has aptly been described as a 'witch hunt' rather than a search for truth (Theall & Franklin, 2001; Marsh, 1987).

In part, the underlying problem is the unresolved tension between the two primary purposes of student ratings, summative and formative evaluation, or accountability and improvement. The frequent teacher complaint is that their interests and needs are not reflected in the use of student ratings and the focus is more on satisfying the requirements of quality monitoring. From teachers' viewpoint, it appears that the concern is more about demonstrating to those outside the institution that something is being done to assure quality, probably more than is the concern about quality improvement inside the institution (Powney & Hall, 1998; Harvey, 2002).

In this sense, ratings are perceived as simply a managerial tool that serves to threaten teacher professional autonomy, rather than a tool for improving teaching quality in the short or long term. The ratings are also seen as a disciplinary measure of some sort to keep teachers in line but linked to the rhetoric of improving quality. There is also the frustration that there is a heavy reliance on student ratings in many instances as the primary measure of teaching performance.

Given the current quality movement in which student ratings of teaching to play a central role, it is now pertinent for researchers to shift focus from issues pertaining to

validity and reliability to how to improve the quality of their use by teachers and administrators alike. This is not to say that the issues of validity and reliability are not important. But after almost 80 years of debate and research it seems unlikely that the argument on whether student ratings should be used for promotion and tenure decisions will be resolved any time soon.

Shortcomings of Ratings Research

Ratings researchers now need to take account of the shifts that have occurred in higher education and change the research focus to increase the practical usefulness of student ratings. What is needed is a combination of the traditional commitment to rigour and recognition of the changing character of the higher education landscape in areas such as: instructional methods; demographics of students; emphasis on the responsibility of students in the learning process; stakeholders' need for useful and understandable information; need for engagement of practitioners in decisions about what to investigate; use of evaluation data for teacher development; and institutional accountability and continuous quality improvement (Theall & Franklin, 2000).

As a key element in quality monitoring it becomes not only prudent but also ethical to find out how ratings data can be used more effectively by practitioners. To start with, researchers need a research agenda that will have a greater influence on policy and practice. Four perceived critical issues need further exploration from researchers: (1) construct validity and ratings forms, (2) use of ratings by administrators, (2) relevance of ratings research to teachers, and (4) conceptions of learning and teaching.

Construct Validity and Ratings Forms

Despite decades of research into the use of student ratings as a measure of teaching effectiveness there is no consensus about the validity of student ratings. For student ratings to be seen as valid they should reasonably relate to criteria of effective teaching. Student learning is seen as the most important criterion of effective teaching, which implies that teachers who promote learning should receive better ratings than those who might fail to do so. The issue of validity becomes particularly important in the use of ratings for summative evaluation. Such decisions must be seen to be based on information that reflects the competence of the teacher, relatively free from the influence of factors unrelated to good teaching, or not under the control of the teacher (Kulik, 2001). There is strong empirical support for the validity of student ratings as one indicator of teaching effectiveness with well-constructed and validated instruments.

For the most part, however, the emphasis of research has been on the issue of discriminant validity, the extent to which variables unrelated to teaching systematically affects ratings. The issue of construct validity, the extent to which student ratings measure the construct teaching effectiveness, has received less attention. Consequently, there is much concern that the items on ratings forms, critical for construct validity, often fail to measure what they claim to be measuring. Marsh (Marsh & Roche; 1997; Marsh, 1991a, b) lamented that many ratings questionnaires are simply ad-hoc lists of items, which do not clearly represent the underlying multifaceted nature of teaching, let alone represent a theoretically sound basis.

Perhaps more disheartening is the fact that the majority of "home-made" ratings questionnaires have not been subjected to rigorous psychometric procedures required for well-designed instruments (Seldin, 1993), thereby failing to reduce uncertainty that

they measure what they purport to measure. Without the support of content relevance, adequacy of coverage, empirical, and theoretical analyses for ratings forms, the scores and any action based on them cannot be interpreted as valid (Ory & Ryan, 2001; d'Apollonia & Abrami, 1997a).

Student ratings can be no more valid than the instrument used to collect the information. Citing research by Tagamori and Bishop (1995), Trout (2000) contended that ratings forms might not be measuring what we assume they should be, that is, teaching effectiveness. Examination of over 200 questionnaires used in American universities revealed that 90 percent of the instruments contained evaluation items that were ambiguous, unclear or vague; 76 percent contained subjectively stated items; and over 90 percent contained items that did not correlate with classroom teaching behaviour. Similarly, Weimer (1997) found from examination of 75 instruments that components of teaching effectiveness were uniformly not covered as some important components were not represented, while unimportant ones were over-represented. The situation is not much different with the use of ratings forms for summative evaluations, as many tend to ask the wrong global questions (Scriven, 1995).

It stands to reason, that with so many varied and poorly designed forms being passed off as assessing teaching, one logical conclusion is that, student ratings might not be valid measures of teaching effectiveness as ratings forms are not valid instruments. Although some might be critical of such a broad-brush characterisation of student ratings of teaching, it points to a larger issue of the absence of a "common language" on the characteristics that constitutes effective teaching and the absence of a coherent set of benchmark standards for ratings forms. In one sense, this absence might not only be

responsible for the variation in the quality of student ratings forms but linked to teachers' reluctance to accept and use ratings data.

There is no unanimity on the characteristics that represent effective teaching across contexts and courses, but there is a fair level of consensus on the general qualities, skills, and behaviours that research has identified to be important for effective teaching and that can be judged fairly accurately by students (Ramsden, 1992; Biggs, 2001, Young & Shaw, 1999). Equally, Centra (1993) usefully identified models of well-designed ratings forms such as the Student Instructional Report (SIR), the Instructional Development and Effectiveness Assessment (IDEA), and the Students' Evaluation of Educational Quality (SEEQ) forms.

Of course, this does not suggest that ratings forms are easily transferred across contexts. With no 'blueprint' to draw on and the lack of research effort in developing a shared understanding of teaching effectiveness, ratings research has had little influence on policy and practice. The result is, institutions design one standard form, which might emphasise one teaching method over another, thereby threatening validity (Weimer, 1997; Ory & Ryan, 2001). There is a clear need for a set of standards which identify the components of ratings instruments, subject to update as situations change.

As researchers have not been able to promote such a system for ratings forms, it is now up to institutions to take action together and develop this system, I suggest. This issue is so important that the idea of a *Ratings Form Taskforce*, for example, sponsored by universities is not specious. The idea of standards for ratings forms does not mean having a standard instrument for use across departments or institutions. What it means is the development of a set of core items that characterise effective teaching along with a bank of items sensitive to different teaching methods, contexts, and academic fields

that could be added to the core items. The result is a product fit for the purpose of its users. This bears some resemblance to the Purdue's Cafeteria System (Centra, 1993; Braskamp & Ory, 1994) in which a pool of items is available from which teachers can select items suited to their needs.

Having standards for ratings forms is likely to result in a common framework to help in determining what is being assessed and what should be assessed for the purpose of generalising about student ratings of teaching as a good indicator of teaching quality across departments and institutions. Until the validity of ratings forms improves, the credibility and impact of ratings feedback will continue to be compromised.

Standards are essential for the process of quality enhancement and continuous improvement. Meeth (2000a) explained, "If there are few standards, there are even fewer criteria against which … improvement can be evaluated" (p. 347). He added, "the effort to improve teaching will not be complete until there is a better base for evaluating that improvement" (p. 349). To be true to the rhetoric of valuing student views in the quest for quality, it becomes imperative to collect relevant and representative information to serve as a firm base for change and improvement.

Use of Ratings by Administrators

The original intent of student ratings of teaching might have been for teaching improvement and teacher development, but ratings data have for a long time now been used as a major input for tenure and promotion decisions. Marsh (1987) provided evidence to indicate that from the late 1970s student ratings have been considered as an important element in personnel decisions about academic staff. It seems, however, that in recent times the emphasis might have shifted to almost exclusive use of student ratings to shape judgments of teacher performance in promotion, tenure, and retention

decisions. And yet, still very little is known about how administrators actually use the information to make decisions about teaching performance. Marsh (1987) elaborated:

> Debates about whether students' evaluations have too much or too little impact on administrative decisions are seldom based on any systematic evidence about the amount of impact they actually do have (p. 350).

The general feeling in the literature is that, probably the greatest threat to the validity of student ratings lies in the far from perfect ways in which administrators engage with student ratings of teaching. Nonetheless, this area of concern has received relatively little attention in the hundreds of studies on student ratings. This is indeed surprising when one considers that misinterpretation or misuse of ratings data in administrative decisions could 'derail' the personal and professional lives of teachers, as well as the functioning of institutions in its focus on continuous improvement.

A study by Franklin and Theall (1989) found that teachers and administrators charged with the responsibility of using ratings data to make personnel decisions lacked relevant knowledge on critical issues about student ratings of teaching that would allow for informed decisions. As would be expected, consistent and serious errors in interpreting ratings data were often made. A distinction, though, must be drawn between unintended and intended misuse and abuse of ratings data. The latter constitutes a violation of moral and ethical standards.

The problem is, simply, that many of these users lack the knowledge, skills, and information required to appropriately use ratings data, and may not even be aware of their own ignorance. The literature does offer guidelines for good evaluation practices (e.g. Braskamp & Ory, 1994; Centra, 1993; Cashin, 1999). Here again, though, there is no good research base concerning what works and what does not work so that

interventions can be designed to support administrators in changing or improving their practice.

Anecdotal reports suggest that administrators simply eyeball ratings results and decide that scores below the mean are bad and those above it are good (Adams, 1997). Even worse, administrators treat percentiles as though they represent a ratio scale and consider differences of one or two decimal places as being real, resulting in the '3.516 is better than the 3.515' syndrome (Theall, 1996; Cashin, 1996). Clearly, the continuing questionable practices of administrators are symptomatic of the need for more attention to be focused on this area by researchers. As McKeachie (1997b) noted, "... until those making the decisions become more sophisticated, the nature of the instrument and possible biases are not likely to make significant differences" (p. 1222).

The validity of student ratings also relies on the interpretation and use of such data. Even if ratings forms are valid in terms of their coverage and relevance of items, for example, intended or unintended inappropriate interpretation and use of the data will threaten the validity of student ratings as measures of teaching effectiveness (McKeachie, 1997b; Ory & Ryan, 2001). It seems, then, that validity concerns also rest with users of ratings data, not just with students alone.

One solution is to provide users of ratings data with the opportunity to improve their skills, knowledge, and evaluation practices. Coinciding with McKeachie's (1997b) call for research on how to train members of personnel committees and other users to be better evaluators, Villaescusa, Franklin, and Aleamoni (1997) provided preliminary empirical evidence of a training workshop approach for administrators and teachers. The researchers found that training in the interpretation and use of ratings data improved knowledge and led to positive attitudes

towards student ratings. It could not be clearly demonstrated, however, that the training led to improvement in practice. More research is needed to solidify these findings.

Another solution is for the use of consultants. For teachers, ratings data is often combined with consultation with an 'expert' as a means of increasing knowledge and skills in using the data effectively. There are no reports of administrators or personnel committees using such approach to assist them in interpreting and using data appropriately. McKeachie (1997b) raised the interesting point of whether these users would be willing to draw on consultants to assist them in using the data. However, this idea might be resisted on the grounds that it violates a supposedly confidential process. Engaging these users in training might be more appealing. In fact, these users should see training as part of their obligation to teachers, just as teachers are obligated to their institutions and students to seriously consider student ratings of teaching as an important element in their teaching improvement efforts.

Relevance of Ratings Research to Teachers

Robinson (1992) is of the opinion that research evidence will not be of much relevance if such research is not conducted as a collaborative inquiry into problems as experienced by practitioners. The flaw in educational research, Hargreaves (2000) argued, is that it is the researchers and not practitioners who determine the research agenda. This type of inquiry is necessary to close the gap between research and practice and facilitate an understanding of the practice of practitioners. While student ratings have received much scholarly attention the actual practice of teachers has been more or less ignored. On this issue, Weimer (1997) criticised ratings feedback research for being uninformed, as it is unresponsive to the needs of teachers and the complexities of the teaching practice in higher education. In the same vein, Menges (2000) described

this lack of understanding of teachers and their practice as one of the most serious deficiencies of student feedback research.

Then, there is the presumption that research findings are generalisable across populations, settings, and times. Like Abrami, d'Apollonia, and Cohen (1990), Centra (1993) questioned this assumption on the basis that the majority of studies are conducted at the undergraduate level in introductory courses with the lecture mode. Therefore generalising to higher-level active learning environments might be problematic.

It is also questionable whether or not inferences about student ratings can be made across different contexts for example:

- paper and pencil ratings versus computer rating;
- campus-based versus distance learning courses;
- modular versus semester long courses;
- part-time versus full-time courses; and
- mainstream versus mature students, as being the same.

The needs of teachers are different in these contexts with implications for the use of student ratings of teaching. As such, the assessment process runs the risk of unfairly penalising some teachers (Theall & Franklin, 2001). It is certainly not unreasonable to expect that, in the face of changes in the higher education environment for both teachers and students, research explores how these changes impact on the use of student ratings of teaching. Unfortunately, this is not the case. Why the inertia in responding to the changing needs of teachers in relation to the use of student ratings?

The gap between research and practice can also be seen in the use of consultation with student ratings. Empirical evidence has shown that use of student

ratings without consultation is likely to result in minimal improvement. For instance, Cohen's (1980) influential meta-analysis reported an effect size $d = .20$ for student ratings without consultation. Instead, only when ratings data were combined with individual consultation that teaching improvement was substantial, effect size $d = .64$. However, it could be interpreted that the aim of this feedback intervention is to improve ratings rather than promote serious change in teaching behaviours. In most instances, suggestions for improvement come in the form of a 'prescription' of techniques and strategies to remedy the problem of low ratings.

From this perspective, there is the temptation to see consultation as 'training' to increase student ratings (cf. Williams & Ceci, 1997). Trigwell (2001) explained that emphasis on strategies will be misplaced unless the teacher is concerned with facilitating learning, but this is rarely considered in the research on consultative feedback as no attempt is made to address teachers' underlying beliefs. So, combining consultation with student ratings might have little to do with learning for improvement but rather learning the tricks of the trade which acts only as a 'band-aid' treatment. This could help to explain how Stanley, Porter and Szabó (1997) were able to conduct research among 'consistent' users of consultation services.

Arguably, what is needed is a conceptual change approach in helping teachers examine their beliefs about teaching and learning in higher education, and the way they view their roles (e.g. Kember, 1997). Support for this comes from Feldman's (1983) findings that overall teacher ratings are negatively associated with age, and years of teaching experience. In his reflection on this finding, Feldman (1983) pointed to teachers' beliefs and values about teaching as one reason for the association. So, rather than a focus on increasing knowledge on 'teaching tips', the consultation process should

also engage teachers in becoming aware of their beliefs and values about teaching. Rowland (2001) went further to suggest that staff developers "stimulate a questioning approach … not only to teaching, but to the very purposes of higher education itself, how it is managed and the wider social context in which … student learning takes place" (p. 165). The extent to which this is possible, though, has received little attention from researchers.

For the most part, traditional research has been preoccupied with demonstrating the effectiveness of consulting with feedback as compared to providing ratings alone or no ratings, to the extent that they have overlooked the needs of practitioners for information on the practices and strategies that work and those that do not. The field of ratings feedback research is littered with varied consultation/feedback intervention procedures used by different researchers. However, reports contain such sparse details on contexts and variables that it is difficult to determine what works and in what contexts. Further, there is a serious lack of replications of the different forms of consultation from which to draw conclusions and provide information on the practices and strategies that are effective in consultative feedback. Research has certainly not been very helpful in this regard as staff developers now have to proceed by trial and error as it relates to designing consultative feedback programmes.

The landscape in higher education is now very different. This means that research needs to reflect an understanding of the pressures and complexities of the teaching practice. It is only on this basis that research can design interventions that will have a likely impact on influencing practice so that long-term improvement will follow.

Conceptions of Learning and Teaching

The most widely accepted criterion of effective teaching is student learning. That effective teaching should stimulate students' curiosity, encourage creative thinking as well as increase students' capacity and desire for learning is not a contentious issue. Unfortunately, not much attention is paid to the role of students in the learning process. Many researchers (e.g. Ho, Watkins, & Kelly, 2001; Kember & Wong, 2000) are however beginning to recognize that there is need to take account of both teachers' and students' underlying beliefs in the assessment of teaching effectiveness.

From all indications in the literature, ratings feedback research has traditionally been conducted on the assumption that all students, as with teachers, share common underlying beliefs regarding teaching and learning. This assumption, however, is not supported by empirical research on teachers' and students' conceptions about teaching and learning in higher education (Kember & Gow, 1994; Marton & Säljö, 1976, 1997; Samuelowicz & Bain, 2001).

The evidence is that, teachers' beliefs and conceptions of teaching range from thinking about teaching as transmission of knowledge to thinking about teaching as producing student learning. In like manner, students' approach to learning ranges from "surface" with a tendency for passive learning with emphasis on the reproduction of facts, to "deep" or active learning which stresses reconstruction of meaning and the fact that the teacher is the facilitator of independent learning. It is these approaches that guide an individual's intentions and behaviours in the teaching and learning interaction.

Student ratings of teaching are influenced by students' own definition of 'good teaching', in turn grounded in their own conception of learning. If the student's underlying belief and approach to learning is underpinned by the perception that

learning is absorbing facts to pass examinations, then this perception creates an expectation from the teacher that then influences the ratings of teacher effectiveness. This is not to say that students cannot make reasonably good judgements about teaching behaviours but any assumption of objective observation on the part of students is seriously flawed.

Given the present transition from teacher-centred to student-centred learning there is the likelihood of a mismatch between teachers' styles and students' preference. So, when teachers choose to adopt strategies and activities that might promote understanding rather than knowing dry facts, and that could enhance students' learning experience, some students might not value this. Most students' preference is for teachers to neatly 'package' the main points in ways that are entertaining and easy to grasp.

Failure by teachers in this regard could mean the receipt of low ratings for choosing to adopt innovative strategies. To think, then, that students who simply desire to reproduce material to pass exams and students who desire to develop a deep understanding in the learning process would use the same criteria to judge teaching effectiveness is simply absurd. Especially, when viewed against the background that students are not trained in how to rate teaching or even how to assess their own learning experience for that matter (McKeachie, 1997a).

Research by Kember and Wong (2000) provided preliminary empirical support that students' conception of learning might be a biasing factor in assessing teaching effectiveness as teachers might be unfairly judged by students with a passive tendency towards learning. This makes the interpretation and utility of student ratings rather

difficult as such ratings cannot be taken at face value. Given the present competitive job market, for example, the majority of students are anxious to receive 'good' grades.

This may mean that students with preference for passive learning might make up a fair proportion of the student body in classrooms as McKeachie (1997a) suggested. This sometimes places teachers in a hostile environment, which might even cause some to adopt "the customer is king" principle and lower academic standards. Still, they might resort to other tactics, not to increase interest and learning but with the intention that they might be rewarded with high ratings. This is, of course, counter-productive to the intention that underpins teaching evaluations.

Simpson and Siguaw's (2000) recent study on whether teachers engage in activities designed to influence student ratings, provides evidence as to how academics use tactics such as:

- hosting student parties;
- telling the class "what a great class they were and how they were all going to be big successes" immediately before administering the evaluations;
- offering easy or no exams, unchallenging course materials, and retakes on exams; "spoon feeding students with lots of information about the examination"; and
- having an unusually great/fun class the day before administering the ratings forms.

Obviously, the game playing with student ratings of teaching is exacerbated by its link to promotion, tenure, and job retention decisions. It is clear then, even if students do take teaching evaluations seriously as Kwan (2000) observed, the

assignment of ratings by students is not a straightforward activity as some might want us to believe and therefore deserves attention from ratings researchers.

Conclusion

It has been argued in this chapter that the increasing use of student ratings of teaching in universities coupled with the disquiet among teachers and lack of good practice by administrators warrants a new direction for ratings research. Teachers are more dissatisfied and frustrated with the use of ratings data at this point in time than ever before. Administrators still do not know how to use ratings data properly and often see student ratings as a 'weapon' to get at teachers. Students are complaining that their views are not taken seriously because there is no closure of the feedback loop. So, within universities there is a growing chorus of complaints from students, teachers, and administrators concerning the use of ratings data.

It is easy to suggest that ratings research become more relevant to the needs of practitioners through, for example, more sophisticated research designs, more qualitative studies, more collaborative efforts between researchers and practitioners in framing problems, or even that research build on theoretical frameworks. Yet, it is naïve to think that research evidence will be used in a straightforward way to influence policy and practice, or that there is a direct link between relevance and quality.

The fact is, researchers operate in a changing ideological climate, evidence may be misused in whatever shape or form it appears, and some users of ratings data have their own agenda for which they desire evidence to justify decisions, and evidence to the contrary has no appeal. There are simply no "quick fixes" available but ratings research has remained essentially researcher-centred for too long.

Chapter 4

The Practice of Consulting over Student Ratings

Consultation over student ratings, in short consultative feedback, is a teaching improvement intervention. This chapter focuses on how feedback from student ratings is combined with consultation to support teaching effectiveness. First, it reviews the research evidence on the impact of consultative feedback. This is followed by a look at the processes and practices involved in teaching consultation. The chapter closes with a brief look at the use of support groups in education.

Impact of Consultation over Student Ratings

Early reviews and research into the use of student ratings feedback in colleges and universities reached different conclusions about the usefulness of student ratings to improve teaching effectiveness (Centra, 1973; Pambookian, 1974; Rotem, 1978; Kulik & McKeachie, 1975; Levinson-Rose & Menges, 1981; Miller, 1971;

Rotem & Glasman, 1979). For instance, from their review of the research literature Costin, Greenough, and Menges (1971) firmly stated that ratings feedback can be used to improve teaching performance. Meanwhile, in their review of the literature Rotem and Glasman (1979) concluded, "... feedback from student ratings does not seem to be effective for the purpose of improving the performance of university teachers" (p. 507). However, the review by Peter Cohen seemed to have settled the debate on the usefulness of student ratings.

Using a meta-analytic approach, Cohen (1980) reviewed 17 studies that compared the effects of providing mid-semester ratings feedback versus no ratings feedback on teaching improvement. The findings of the meta-analysis suggested that providing university teachers with student ratings indeed had a positive effect on teaching but only a modest one, effect size = .20. In contrast, when ratings feedback was combined with individual consultation the effect on teaching was much larger, effect size = .64. Cohen (1980) concluded that teachers need more than just feedback from student ratings to improve their practice; they need suitable support to use student ratings effectively.

An update of this early meta-analysis found an even larger effect. In their meta-analysis Menges and Brinko (1986) reported a mean effect size = 1.10 for studies that combined ratings feedback with individual consultation. For illustration, this effect size means that those teachers who received support with their ratings feedback improved in their teaching effectiveness to a level that was greater than 86% of teachers in the control group who received no consultation.

Other research studies have looked at the effects of combining ratings feedback with strategies such as; student focus groups (Tiberius, et al. 1989), self-help printed

material (Cohen & Herr, 1982), and workshops/seminars (Rader, 1995). Even so, it is individual consultation that has had the most significant effect on teaching effectiveness, in a variety of settings and using different strategies (Hampton, 2001; Piccinin & Moore, 2002; Piccinin, Cristi, & McCoy, 1999; Rozeman & Kerwin, 1991; Cohen, 1991). Commenting on the consistency in findings Marsh and Roche (1993) wrote, "The most robust finding from the SET [students' evaluation of teaching] feedback research is that consultation augments the effects of written summaries of SETs" (p. 223).

One limitation of the research on consultative feedback is that, the focus has been almost entirely on evaluating individual consultation, even though it has long been recognised as time consuming and costly, and so could only be made available to a few teachers. In part, the use of peers as consultant emerged because staff development units had limited financial and human resources to be able to cope with the high demands for assistance through consultation (Weimer & Lenze, 1997).

A few studies have examined the effects of peers as consultants (Annis, 1989; McKeachie, et al. 1980; Rozeman & Kerwin, 1991). For instance, Annis (1989) used a peer partner approach in which teachers were paired for reciprocal class observation and assessment of teaching. The results from this study suggested that teaching improved from the use of this group process. However, it is impossible to infer a causal inference because of methodological weaknesses. For example, the one-group pretest-posttest design opens the results to alternative explanations, not least of which was the promise of monetary reward. McKeachie, et al. (1980) used the stronger experimental design and found that consulting with peers indeed had a positive effect on teaching effectiveness.

Even more noticeable is the paucity of research on group-based consultation. Only one known study comes close to being regarded as group-based consultation. The study by Hoyt and Howard (1978) experimented with a consultation model in which teachers from various academic units were assigned to groups of eight members and were engaged for 10 weeks in a variety of activities, including discussion on teaching methods, classroom observation, and assessment of teaching. The results indicated that the approach had a significant effect on teaching effectiveness. This model though, did not involve teachers disclosing their student ratings feedback to peers but simply involved teachers in active learning with and from one another.

There are, however, issues in the use of ratings feedback that require some attention. A key issue is that ratings feedback studies are typically conducted within a short-term one-semester time frame, so that feedback to teachers is based on mid-semester ratings of teaching. There are at least two problems associated with the use of mid-semester ratings. The first is that there are questions concerning the legitimacy of mid-semester ratings as a treatment and whether mid-semester ratings generalise to end-of-semester ratings.

In an examination of the problems that affect the interpretation of findings in ratings feedback research, L'Hommedieu, Menges and Brinko (1990) argued that because the normal practice of universities is to collect student ratings near the end of the semester, the effects of mid-semester ratings feedback might not generalise to end-of-semester ratings feedback. The purposes of mid-semester and end-of-semester ratings are considered to be different and therefore establish two different situations for the ratings. Mid-semester ratings are formative in nature and are expected to be used to effect immediate changes to teaching. Conversely, end-of-semester ratings are

summative in nature, provide a judgment on teaching in that semester, and are intended to be used to improve teaching in subsequent semesters (see also Marsh & Roche, 1993).

The second problem is that intervention effects may be diminished as teaching cannot be easily modified within one semester (Marsh, 1987; Marsh & Dunkin, 1992). Related to this is the fact that because teachers might only be able to make minor alterations to teaching within the one semester, this might not readily be perceived as improvement by students when making end-of-semester ratings. In this respect, it is possible that students might penalise teachers with low ratings for what might appear as unwillingness to incorporate their suggestions from mid-semester ratings. As pointed out by L'Hommedieu, et al. (1990), students provide mid-semester ratings with the intent that immediate changes will be made to enhance their learning, not for use in subsequent semesters.

More research is needed to compare the effects of mid- and end-of-semester ratings feedback. In an initial study Marsh and Roche (1993) found that the effects of end-of-semester ratings feedback were stronger than those of mid-semester ratings. This resulted even though teachers in the mid-semester group received the feedback intervention both at mid- and end of semester while teachers in the end-of-semester group received the feedback intervention only once near the end of the semester. These concerns about the suitability of mid-semester ratings feedback challenge the evidence for the effectiveness of mid-semester ratings feedback for evaluating and improving teaching (Marsh & Roche, 1993).

Another issue that L'Hommedieu, et al. (1990) suggested has been overlooked in ratings feedback research is the reporting of whether participants actually

experienced the feedback intervention as intended. In other words, studies do not state whether participants paid any attention to the feedback. This is important to consider in drawing conclusions about the effects of any feedback intervention. Related to this is the lack of information in studies on the way in which ratings results are reported. Some ratings reports may contain only brief summary details of ratings feedback, which may provide little useful information for teachers to work with. On the other hand, some reports may provide adequate and good information to facilitate correct interpretation and use of the ratings feedback to change and improve teaching behaviours. The format of ratings reports is therefore likely to moderate intervention effects even with support from a teaching consultant.

Despite an impressive empirical record that endorses the effectiveness of consultation over student ratings, some 30 years on, the practice is still not fully developed. It is not yet well understood what constitutes effective support, and there is the appearance that practice is guided simply by common sense and trial and error than by a conceptual base or even by an evidence base. Much of what is known about consultation is based on the experience of practitioners and not research, and only a small handful of specific information on teaching consultation is in the public domain (Brinko & Menges, 1997; Knapper & Piccinin, 1999).

The Practice of Teaching Consultation

Like workshops and seminars, the use of teaching consultation has been used in staff development from the mid-1960s. Centra (1978) found that a high proportion of teachers used consultation. But the first widely known teaching consultation programme was the Clinic to Improve University Teaching, which began at the

University of Massachusetts in the United States, in the early 1970s (Tiberius, 1995). Since then similar programmes have been designed internationally even though North America may still be the largest provider of teaching consultation to university teachers (Hicks, 1999b).

Teaching consultation, also known as instructional consultation, and commonly referred to simply as consultation, is a process through which an individual teacher talks with a teaching consultant—typically a staff developer or trained peer—to receive advice and assistance in addressing a teaching concern. Generally, consultation takes place over three issues: general teaching issues, student ratings data, and classroom observation and/or videotape of teaching. Menges (1997) asserted, "no other service provided by teaching centres has greater potential for producing deep and enduring effects on academics and teaching" (p. v). Marincovich (1999) concurred, adding that one of the most important steps universities can take to increase the effectiveness of their teaching evaluation system is to provide a teaching consultation service.

In practice, teaching consultation takes the form of dialogue and feedback between a teacher and consultant, working to better understand the nature of a teaching problem and to reach an understanding of how to resolve the problem and bring about change in teaching by developing improvement strategies (Boud & McDonald, 1981). It has the following characteristic features: First, consultation is a problem solving activity, in that teachers typically seek consultation for expert advice when there is a specific teaching concern, a problem in the classroom, or on receipt of low student ratings, the need being to improve teaching practice (Knapper & Piccinin, 1999).

Second, consultation is seen as a voluntary, confidential, and individualised process. It is generally initiated by teachers who seek assistance to improve an aspect of

teaching, although department heads do make referrals. This is probably why consultation seems to be stigmatised as a remedial programme. Knapper and Piccinin (1999) drew the similarity with some aspects of counselling and the notion of person-centred therapy but with feedback as a catalyst for change.

Third, consultation is a formative process in which teachers receive feedback on their teaching and information on strategies that could be adopted to improve practice in a supportive, educative, collaborative, and investigative manner, without making judgements and demands for change (Lenze, 1996). In concise terms, consultation is a structured collaborative problem-solving process that uses information about teaching performance as a basis for reflection and discussion about change and improving practice.

Although there is a tendency to associate consultation with the less effective teachers, it also benefits those teachers who might already be teaching well but wishing to develop certain skills to become more effective. A dilemma for staff developers is that many teachers who should probably seek help and advice on their teaching often do not seek out consultation. Instead, it is teachers who are already committed to continuing professional development that tend to use the service to become even better.

And yet, any attempt to improve teaching by coercing teachers to seek consultation is almost certainly doomed to failure. Coercion, as Knight and Trowler (2000) warned, runs the risk of compounding negative feelings about teaching. When this happens, teaching improvement is likely to remain superficial, creating a culture of compliance in which there is 'change without change'. Teaching does not improve and student learning does not improve.

It should be understood that the term 'consultation' is sometimes loosely used to refer to just about any kind of guidance and assistance activity aimed at improving teaching and learning. In this generic sense 'consultation' may be associated with activities such as; mentoring, peer coaching, formative peer review, action learning, peer support groups, peer consultation, individual consultation, and other types of advice on teaching and learning activities. In the tradition of ratings feedback research, this present study is concerned with consultation as a structured intervention that facilitates learning from student ratings feedback through informational and emotional support.

Much of the literature on staff development programmes emphasises the need for a process that is voluntary, confidential, and reflective. Few would disagree that these are indeed desirable features to increase teacher participation and engagement in such activities. Yet, the idea that consultative feedback should be based on voluntarism seems worrisome nowadays, given the presumption that quality in student learning is dependent on better teaching that can be informed and supported by feedback from student ratings.

Drawing a parallel with the research role of academics, Handal (1999) wondered, whether it is not time to see consultation as mandatory professional development for all teachers. Handal's (1999) point is that university teachers are accountable to their institution, the profession, and to their students to teach well and promote quality student learning. Therefore, if they are to be successful, teachers need to learn about their teaching and to develop better ways of teaching, through such means as consultative feedback. Consultation should probably, then, be carefully integrated

into staff development programmes, where it will better serve the needs of the majority of teachers, rather than be an optional activity that serves only the few committed ones.

Critical to effective consultation is the knowledge and skills used by the consultant in helping teachers to improve their instructional effectiveness. Interestingly, faculty developers working as consultants generally receive little or no formal training to perform their role effectively and there is as yet not enough information as to what counts as effective consulting. Instead, faculty developers practice consultation "by the seat of their pants" (Knapper & Piccinin, 1999; Brinko, 1990). Although this area has not been examined in much detail it is recognised that teaching consultants need knowledge of both teaching and learning styles; of how to assist teachers with their teaching; of different consultative styles and to use the style appropriate for different situations; of a range of instructional methods and strategies; and knowledge of new developments in both faculty development and in student learning.

Best use of this knowledge, however, requires essential skills which involve being able to establish trust between the consultant and teacher; to give effective feedback to teachers; to identify, diagnose, and problem solve; willingness to help teachers work through issues relating to their teaching; and being able to draw upon research concerning teaching and student learning (Brinko, 1997; Boud & McDonald, 1981).

Effective consultation also entails the consultant being able to make teachers feel comfortable about seeking help to improve their teaching; and being able to help teachers set specific goals for improvement, monitor progress, and maintain continuous improvement. Because feedback on teaching is central to consultation it is also important for the consultant to have a good knowledge of the various methods to obtain

information on teaching effectiveness; knowledge of the literature on teaching evaluation; involvement in the evaluation of teaching; and being able to use the data from student ratings to determine teachers' specific needs and guide improvement (Border, 1997; Theall & Franklin, 1997; Sorcinelli, 1997; Morrison 1997).

The Consultation Process

Although approaches to consultation vary somewhat they all seem to follow a cyclical process involving four main phases (Brinko, 1997). The first phrase, initial contact, occurs when the teacher recognises the need for advice or help with a teaching issue and contacts the staff developer, either in person, by telephone, or by electronic mail. The cycle then moves into the conference phase, or as Berquist and Phillips (1975) preferred 'contracting'.

This phase is characterised by an extensive discussion between the consultant and the teacher and could be thought of as a needs assessment, necessary to be able to design appropriate strategies to bring about change. It is at this stage that the teacher decides whether to commit to the process. It also gives the consultant the opportunity to better understand the particular teaching context and negotiate the collection of additional information.

The third phrase in the cycle is that of data collection in which appropriate media are used to provide evidence on teaching behaviours in order to diagnose the problem. Student ratings are an important source of information but sources such as classroom observation and/or videotape recording of teaching and self-evaluation of teaching strengths and weaknesses are usually considered.

In the fourth phase of the consulting cycle, the information collected is reviewed and analysed with the teacher to diagnose problems and explore solutions. A noticeable omission from the cycle is an explicit step for follow-up procedures. Without the support of follow-up, teachers are left on their own to implement improvement strategies. Failure to do so in a manner that students find effectual will almost certainly result in additional negative feedback which could prove rather damaging to the self-esteem and self-confidence of these teachers.

Models of Consultation

Drawing on research from the fields of education, psychology, and organisational behaviour, Brinko (1990, 1997) proposed five models that describes the interaction that is likely to take place between a consultant and a teacher—1) product, 2) prescriptive, 3) collaborative/process, 4) affiliative, and 5) confrontational. The behavioural pattern that emerges in the consultation process will vary according to influential factors such as the number of meetings, whether consultation includes student ratings, the model adopted by the consultant, expectations, personality traits, synergy between consultant and teacher, and even the institutional climate.

In the product model the teacher approaches the consultant for assistance on a problem already identified. The consultant's role is therefore to offer solution in the form of advice, or assistance on, for example, the construction of a test, video, or some other form of "product". In the prescriptive mode, the consultant is more likely to be the one to identify, diagnose, and suggest possible solutions to the problem. What is problematic here is that, this may create dependency on the consultant as *the expert* with

the answers. This might actually prevent the teacher from becoming actively involved in the learning process instead remaining a passive recipient of advice.

The collaborative model places the consultant in the role of 'facilitator of change', working in partnership with the teacher to resolve problems to effect improvement in teaching. The use of this model therefore is likely to see the teacher and consultant working together as a team. The importance of adopting a collaborative approach was reinforced by teachers in a recent study. These teachers pointed out that the consultant/teacher relationship is not one of "doctor/patient" relationship but should be organised as a team effort (Stanley, Porter, and Szabó, 1997). The consultative interaction is expected to be teacher-centred and collaborative if it is to be useful.

Although not a common approach with instructional consultation, in the affiliative model the consultant focuses on empowering the client and solving personal problems that may be negatively affecting teaching performance. For the confrontational model, the consultant adopts the role of 'challenger' or 'devil's advocate'. The aim is to somehow force the client into a position of either defending or accepting the problem, where it is recognised that the real problem is different from the stated problem. Brinko (1997) was careful to note that the confrontational model is not recognised as suited to instructional consultant but that research results seem to suggest that it holds promise to bring about change with some teachers.

The models offer a framework for examining the interactive process of instructional consultation, but it may be difficult to clearly delineate a model in a consultant-teacher meeting. Any model of interaction can emerge in consultation sessions but Brinko (1990) found that the prescriptive and collaborative models are the most likely ones to emerge in practice.

As consultation is expected to meet the specific needs of teachers it is not expected that the process is standardised. The consultant may therefore need to draw on the different models in order to work flexibly and respond to the unique demands of each teacher-client (Boud & McDonald, 1981). With growing interest in the use of peer- and group-based approaches, research is now needed to determine whether similar interaction patterns exist, and basically how the relationship evolves in these models.

A particularly important aspect of the consultation process is the quality of the interaction between the consultant and the teacher. Brinko's (1993) point that protecting the self-esteem of the teacher is crucial in giving feedback was echoed by Stanley, Porter, and Szabó (1997) for consultation. Teachers in their study stressed that it is already embarrassing for teachers to know that their teaching is not going well. Importantly, it is believed that consultants should aim to make teachers feel comfortable in sharing their experiences and seeking help, rather than destroy the confidence in their ability to teach.

Types of Consultation Programmes

A characteristic feature of teaching consultation is variation in approaches and practices. Morrison (1997) suggested a very useful framework that identifies six different types of consultation programmes that might be organised as shown in Figure 4.1.

Method	*Role Relationship*		
	Developer as Consultant	Peer as Consultant	Peer as Partner
Individual	Traditional	Peer Consultant	Peer Partner
Group	Developer-led Workshop	Peer-led Workshop	Support Group

Figure 4.1 Instructional Consultation: A Typology of Programmes (Morrison, 1997)

Individual Consultation Programmes

The focus here is to assist the individual staff member to improve teaching:

- Traditional programme: staff developers provide one-to-one consultation to individual teachers.
- Peer consultant programme: a trained colleague provides consultation to another colleague to improve an aspect of teaching.
- Peer partner programme: two colleagues who chose to work together as partners engage in mutual inquiry processes into teaching and learning.

Group Consultation Programmes

In group-based programmes, colleagues serve in consultative roles for one another:

- Developer-led workshops: staff developers lead workshops that allow colleagues to learn from each other.
- Peer led workshops: a colleague or team of colleagues act as facilitator(s) for a workshop. The intent being to draw on participants for feedback and consultation.
- Support groups: colleagues work together to support their individual efforts to solve teaching problems and stimulate learning for teaching improvement.

Use of Support Groups

Given the traditional norm of privacy in teaching in university cultures, the dominant model of consultation over student ratings has been individual, one-to-one, consultation. However, with the increasing emphasis on collaborative professional development as a strategy to improve quality in teaching, use of support group initiatives is increasing. These support groups go by many names: teaching communities of practice, critical friendship groups, teacher learning communities, teaching teams, peer support groups, teacher discussion groups, quality circles, teaching circles, and teacher focus groups, to name a few. It is observed, however that these groups seem more likely to be established in schools than in universities.

The act of professionals meeting in groups as a means to improve performance is not new and is prevalent in diverse fields. For example, group (team) work has long been promoted in business (cf. Deming 1986; Senge 1993) as a means to improving

quality and effectiveness in organisations. Senge (1993), for example, argued "Teams, not individuals, are the fundamental learning unit in modern organisations" (p. 10). In education support groups are conceptualised as small groups of teachers informally organised not only for social interaction but also for meaningful knowing about the practice of teaching in order to bring about change and improvement.

The use of support groups, though, for consultation over student ratings is not prevalent in the literature. In the model being tested in the present investigation, peer support groups were formed for the purpose of supporting teachers to learn from student ratings feedback. In this instance, individual group members bring to the attention of the group specific areas of practice they have targeted for improvement, having received low ratings feedback.

Through dialogue, exchange of ideas, sharing of experience, and mutual support teachers should become aware of possible improvement strategies that might be relevant to improve the area of teaching targeted. Peer support groups, as used in this context is grounded in the ideas relating to communities of practice and action learning. These concepts recognise that individual learning is important but also realise that knowing about and improving practice is also dependent on the opportunity to construct knowledge through dialogue in social interactions with colleagues.

Similar with individual consultation, the use of peer support groups as a consultation approach provide mainly three kinds of support (McKeachie, et al. 1980; Cohen, 1991; Brinko, 1997; Morrison 1995):

- Informational support – interpreting ratings, sharing knowledge and experience, developing improvement strategies;
- Emotional support – motivation and social support; and

- Reflective support – self-reflection and reflection on teaching to improve practice, engaging in problem solving.

The use of peer support groups to improve practice is analogous to the idea of 'critical friends'. Critical friendships are seen as practical and collaborative partnerships among colleagues. Such 'friendships' are voluntarily formed to support reflection on practice, share ideas or problem-solve about an aspect of practice. These are characterised by principles of: relationship between equals, mutual support, collaboration, reciprocity, reflection, dialogue, social interaction, constructive criticism, and a relationship of confidence and basic trust in the good intentions of the critical friend (Handal, 1999; Fullan & Hargreaves, 1992; Day, 1999).

Just to mention a few, the important benefits of such critical friendships would include: enhancing professional dialogue, breaking the isolation of practice while promoting further collaborative development, facilitating reflection on practice through group discussion, questioning and even confronting the other 'friends' to facilitate deep reflection, while increasing awareness of aspects of practice that may not have been considered before. As noted by Sachs (2000), "the tension that sometimes emerges through the observations and interventions of a critical friend can be productive and lead to new insights and opportunities not previously apparent to other parties" (p. 89). For Hatton and Smith (1995) critical friendships is an opportunity to give voice to a teacher's thinking, while allowing the 'voice' to be heard in a sympathetic but constructively critical way.

Critical friendships are also based on the recognition that collaboration between teachers is necessary to promote professional development:

> ... individual teachers cannot significantly improve their practices in isolation without opportunities for discussion with professional peers and others operating in a significant role-relationship to them (Elliot, 1992, p. 25, cited in Handal, 1999, p. 65).

Critical friendships formed against a common background such as disciplinary areas can indeed allow for critical dialogue about the nature of practice and how it can be improved as the notion of community of develops. In reality, the 'critical' element of the friendship may not be present in many such relationships. The relationship may be reduced to 'chit chat' and trading tips, possibly spreading poor practice. This actually avoids the questioning and confrontation that may be necessary for reflection and evaluation of teaching on a deep level.

This may result because the 'friends' want to avoid tensions to the social aspect of the relationship. It may also be that the level of trust between the 'friends' has not yet reached a comfortable level from not having spent enough time in the process, or even because ground rules were not established at the beginning of the relationship, so there is uncertainty on how far to go (e.g. Hatton & Smith, 1995; Day, 1999).

Despite these limitations support groups are generally regarded as a useful way of helping teachers to work with and learn from each other. Therefore, used properly it can help teachers to understand their practice and at the same time promote a sense of collaboration and collegiality among teachers for teaching improvement.

Chapter 5

Consultation over Student Ratings: What Works? A Meta-analysis

There is, as yet, little evidence on 'what works' in consultation over student ratings. This chapter reports on an exploratory meta-analysis that sought to identify the strategies and practices that may be important for effective consultative feedback. This meta-analysis adds information to the question of how best to organise consultative support. The meta-analysis is a complete study on its own, and the findings were influential in shaping aspects of the intervention experimentally tested in the present investigation.

Introduction

Consultation over student ratings, consultative feedback, is now widely recognised as an effective support strategy that is most beneficial to teachers to use

ratings feedback to improve their teaching. Even so, the practice is complicated by little evidence on the role of different consultation practices and strategies on the effects of consultative feedback. This noticeable absence is tantamount to a denial of the importance of these elements in enhancing the effectiveness of the process.

The traditional notion is that, it is the consultant that makes the process effective (e.g. McKeachie et al. 1980; Wilson, 1986; P. Cohen, 1991). The effectiveness of consultative feedback has been repeatedly linked to the support that the consultant provides as the significant factor in process. While not denying the importance of the consultant's role in supporting teachers, they do use different practices and strategies in the consultation process. What is the impact of these practices and strategies on the outcome effects of the consultation process?

Research results show variable effects of consultative feedback on teaching effectiveness. For example, from the comparisons made by Menges and Brinko (1986) in their meta-analysis, studies emerged with effect sizes ranging from 0 to 2.50, and considerable variation in between, implying the presence of other important factors apart from the consultant. Arguably then, all forms of consultative feedback activities might not work equally well when reviewing student ratings feedback. Unfortunately, as with Cohen's (1980) synthesis, the factors associated with the strongest effects compared with those of the weakest were not sufficiently explored to explain the variations in effects. It must be said though, that the scant descriptions of processes and interactions in study reports do make this task rather difficult.

The question of the most successful way to combine consultation with student ratings has been raised on many occasions (e.g. McKeachie et al. 1980; Wilson, 1986; Murray, 1997) but has remained unanswered. So, despite the enthusiasm about the

potentials of consultative feedback to improve teaching effectiveness, there is a dearth of research on what strategies really work, and how to make consultative feedback work even better.

In the absence of empirical evidence there has been much speculation. Wilson (1986) believed that the effectiveness of his approach rested on the availability of practical teaching ideas that are relevant to the behaviours teachers might wish to change. For Piccinin, Cristi and McCoy (1999) what works is a custom-designed process, basic or intensive, according to the specific needs of teachers. Aleamoni (1978) believed it is the opportunity to discuss the ratings with a resource person that is the key factor in the effectiveness of consultative feedback.

Similarly, McKeachie et al. (1980) held that it is the motivational, emotional, and informational support from the consultant, in terms of interpreting ratings and providing suggestions for improvement that make consultation works. Marsh and Roche (1993) linked success of the strategy to use of a valid and reliable ratings form and setting specific improvement targets. So, it is obvious that a number of different strategies may be used effectively. But should the practice remain fragmented with researchers and practitioners operating on a trial and error basis?

Indeed, it is surprising that for such an important intervention for university teaching there is a paucity of evidence on how best to structure consultative feedback to maximise its benefits. This lack of evidence on the practices of effective consultation is seen as a serious deficit in student feedback research (Marsh & Roche, 1993). The need to secure evidence on which to base decisions and practice takes on a special appeal when viewed in the light of current reform efforts in higher education. Emphasis on

and use of student ratings in quality monitoring, by both institutions and governments, have increased tremendously in recent times.

Aspects of the consultation process also vary in important ways. For instance, the process might constitute a brief 15-minute meeting (Aleamoni, 1978) or 10 hours (Erickson & Erickson, 1979), augmented with activities such as workshops/seminars and class observation and/or videotaping (Hoyt & Howard, 1978), or resource materials on recommended teaching ideas that 'work' (Wilson, 1986; Marsh & Roche, 1993). In addition, consultants vary in their level of expertise. In some cases professional staff developers are utilised (Atchison, 1987) while in others trained or untrained peers (McKeachie, et al. 1980; Rozeman & Kerwin, 1991), or even graduate students (Erickson & Sheehan, 1976) serve as consultants.

What is more, any one of the five models of consultative styles proposed by Brinko (1997) might be adopted in the consultant/teacher interaction, although not often explicitly stated in many research reports. For example, a consultant might choose to work in a prescriptive manner to analyse, interpret, and recommend improvement strategies for the teacher. Alternatively, a consultant might choose to adopt a collaborative approach, where both the consultant and the teacher see the process as a team effort to engage in creative problem solving to facilitate teaching improvement.

Surely, the success of the consultation process may hinge on a number of factors other than the support that teachers might receive from the consultant. Research has focused on demonstrating that consultation works presumably under different conditions, yet failing to provide adequate information as to the components that worked and those that did not work for the most effective consultation. The problem is

further compounded by a lack of replications of the different types of approaches employed in different sites.

This is not to imply that the multiplicity of approaches is an indication of weakness in the practice of consultation. After all, improving teaching is a complex task that no one approach would be appropriate. But we need to be clear about what is important for successful consultative feedback. An inherent risk in an imprecise picture of the strategies that should be used to enhance the effectiveness of consultative feedback, is that its true value could be ignored, and the practice replaced by other programmes that might not place emphasis on the use of student ratings feedback. It is therefore important to understand the conditions under which teaching improvement through consultative feedback can be maximised. A meta-analysis can help to address this problem quite effectively by cumulating the findings of a number of individual studies.

META-ANALYTIC REVIEWS

A meta-analysis is a statistical procedure for combining research findings from many studies. It has become a well-accepted method of conducting a quantitative review and has developed in conceptual and methodological sophistication over the years (Glass, McGaw & Smith, 1981; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Cooper & Hedges, 1994; Lipsey & Wilson, 2001). The apparent attractiveness of meta-analysis comes from the understanding that cumulative evidence provide a more accurate and credible form of evidence on the efficacy of an intervention, offering a better guide to policy and practice, than the results from an individual study.

Corresponding with the work of Glass, McGaw and Smith (1981), Hedges and Olkin (1985), and Hunter and Schmidt (1990), Rosenthal (1995) defined a meta-analytic review as a quantitative summary of finding across a body of research that describes:

> the typical strength of the effect or phenomenon, its variability, its statistical significance, and the nature of the moderator variables from which one can predict the relative strength of the effect or phenomenon (p. 183).

Historically, the first statistical synthesis of results from independent studies is reported as being conducted by statistician Karl Pearson in 1904 but the term *meta-analysis* was coined in 1976 by psychologist Gene V. Glass. Thereafter, Cooper and Rosenthal (1980) made the empirical case for meta-analysis by showing that narrative reviews lead to inaccurate or imprecise characterisations of the cumulative research results (Egger & Smith, 1997; Cooper, Valentine, & Charlton, 2000). As a method for synthesising research findings it provides valuable information for subsequent use in research, policymaking, and practice.

Merits of Meta-analysis

Narrative Reviews

The standard approach to summarising research results is the use of narrative or qualitative reviews that may look for patterns in the studies and draw conclusions. There are two problems associated with narrative reviews. First, narrative reviews suffer from a subjective, imprecise, and often inaccurate means of accumulating research results. Individual reviewers are known to rely on informal, biased, and unspecified procedures in selecting, assessing, and reporting on studies. For instance,

the selection of studies may be based on the reviewer's own view of the quality of the study, or those that support their particular views, or even those that represent a convenience sample of studies.

Reviewers might also rely on vote counting of statistically significant results in which the hypothesis of interest is or is not supported and choose the view receiving the most votes. With this level of subjectivity and use of unsystematic procedures, it becomes difficult to replicate the review. This may lead to erroneous interpretation of the body of knowledge and conflicting conclusions as often demonstrated in the literature (Lipsey & Wilson, 2001; Wolf, 1984; Johnson & Eagly, 2000).

The second problem associated with narrative reviews relates to the task of reviewing a large number of studies on a particular issue. As an example, Cashin (1999) found over 2,000 citations in the Educational Resources Information Centre (ERIC) database alone on the term "student evaluation of teaching performance". Even though over one-half could be opinion pieces, it remains that there is still a large number of research studies to be reviewed. Besides, research into the use of student ratings feedback is increasing at a rapid rate as the debate into validity continues. The dilemma here, as Glass, McGaw and Smith (1981) observed, the information processing capacity of the human mind is limited to process the data reliably and validly.

Although ratings feedback research continues to be riddled with controversies, it is probably now easy to see that much of the inconsistency in the early review of evidence on the usefulness of student ratings stemmed from the use of mainly narrative reviews. To illustrate, Dowell and Neal (1982) reviewed only six studies that attempted to validate student ratings from a possible pool of 41 studies using a quantitative but not

meta-analytic technique. This selective review led to their conclusion that student ratings were not valid measures of teaching effectiveness.

Using meta-analytic procedures, with the same criteria Dowell and Neal (1982) established, Cohen (1983) retrieved 22 studies and found a correlation of $r = 0.38$, for a variance of 14.4%, in contrast to Dowell and Neal's 3.9% reported between-teacher variance. Cohen's (1983) findings concluded that there was a moderate positive correlation between overall teacher ratings and student achievement, supporting the validity of student ratings. Meta-analysis not only offers a more efficient and objective means of summarising research evidence but also allows researchers to arrive at conclusions that are more accurate and credible than can other type of reviews (Rosenthal & DiMatteo, 2001).

Meta-analysis

As a quantitative review, meta-analysis aggregates the results of several individual studies that can be meaningfully compared. The underlying principle of meta-analysis is that no single study can provide complete answer on an issue of interest. Rather, knowledge base develops through the accumulation and integration of findings from a number of research studies. The use of this technique is seen as a more objective assessment of the evidence than traditional narrative reviews, a means to provide more precise estimate of treatment effect, and an efficient way to explain heterogeneity between the results of individual studies (Egger & Smith, 1997).

There are two key elements in meta-analysis. First, conducting a meta-analysis follows a research process that recognises transparency. As such, it uses rigorous and explicit procedures to ensure thoroughness in finding and selecting studies; accuracy in coding information from research reports; accuracy in computing estimates of effect

sizes and analysing the results; and exploration of sources responsible for variability among effect sizes. This is necessary to facilitate replication and critique of the review.

Second, the results of individual studies are converted into a common metric to allow for comparisons and interpretation of research results from studies on different outcome measures, showing the pattern of results that are not observable in individual studies. For this reason, it uses a measure called effect size statistic, a standardised measure of change associated with the treatment. More precisely, Cohen (1977) defined effect size as "the degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false" (p. 9).

The two most common metrics are the *d* and *r* indices, for experimental and correlational studies respectively. The *d* index represents the measure of the difference between the means of the experimental and control group expressed in terms of their standard deviation. The *r* index represents the standardised regression coefficient, including correlation coefficient, of the relationship between the independent and dependent variable. In a meta-analysis, effect sizes are, (a) calculated from the outcomes of comparisons, (b) averaged across comparisons to estimate the general magnitude of effect, and (c) compared between comparisons to discover if variations in outcomes exist and if so what features in the comparisons might account for them.

The measure of magnitude of effect is particularly important in evaluating educational research findings as it shifts attention to the more important question of how much difference an intervention makes, rather than if the difference was statistically significant. Significance test is a function of the size of effect and sample size of a study and in educational research samples sizes are usually quite small. This produces

low statistical power to reject the null hypothesis and detect statistical significance (Fitz-Gibbon, 1984).

For meta-analysis to be appropriate, however, the synthesis must: (a) focus on empirical studies, (b) have the goal of integrating the results of studies so as to create generalisations and set the limiting conditions of the generalisations, (c) employ a neutral perspective, that is, the research is not being mustered to support a particular point of view determined prior to the review, (d) cover a near exhaustive selection of relevant studies, and (e) use the same standards of rigour required of primary researchers (Cooper & Dorr, 1995).

The benefits of meta-analysis in clarifying contradictory conclusions, and in generating knowledge is clearly illustrated in student feedback research by Cohen's (1980) meta-analysis followed by an update by Menges and Brinko (1986) and L'Hommedieu, Menges and Brinko (1990). In contrast to the inconclusive evidence from narrative reviews and individual studies on the usefulness of student ratings these syntheses demonstrated quite easily that a positive relationship exists between student ratings and teaching improvement.

What's more, these meta-analytic reviews uncovered that the association increased when ratings were augmented with individual consultation. The limitation of these reviews, however, was that the variation in effect sizes found were not explored and explained. Performing moderator analyses to explain variation in study outcomes is regarded as a very useful and informative procedure in conducting a meta-analysis (Rosenthal, 1995; Cooper, Valentine & Charlton, 2000).

Limits of Meta-analysis

Although use of meta-analytic techniques offers many benefits, it does not offer a perfect solution to problems in research reviews. There are limitations to its use. Cooper and Dorr (1995) explained that a meta-analysis, (a) is unable to establish that a causal relationship exists between the variables, (b) cannot overcome the problem of confounding study-level variables, and (c) perhaps most importantly, a meta-analysis is no substitute for wisdom. A statistical method cannot generate theories that do not already exist. And even less ambitiously, a statistical method cannot point out to its users what variables should be moderators of relationships. Only the human intellect can do these things.

Equally, there are criticisms of the technique. For instance, Eysenck (1994) is critical of the practice that many meta-analyses are being conducted by non-experts in a particular field. The concern here is that, these meta-analysts may not be able to appropriately integrate treatments that are comparable, thereby resulting in meaningless and misleading estimates of effects, especially when study findings are aggregated for an average mean effect. Over generalisation can therefore occur. This criticism is often referred to as the "apples and oranges" problem in meta-analysis.

There is also the criticism that studies with discrepant findings are often combined. On this point, Eysenck (1994) argued it would be more prudent to clarify discrepancies than average estimates of effect sizes over discrepant data.

Graham (1995) is also unconvinced that there is much "richness and rigour" to be gained by mixing good and poor studies together in a meta-analysis. Variation in the quality of studies included in a meta-analysis can make ambiguous the real story the data is trying to tell. This criticism is termed "garbage in and garbage out".

Another concern about meta-analysis relates of its validity. The validity of meta-analytic conclusions is threatened by failure to conform to the rigorous and systematic procedures associated with the proper use of the technique. For instance, during the problem formulation stage, bias could be introduced if the meta-analyst does not pay careful attention to conceptual distinctions in definition and hypotheses that were viewed as important by others in the field.

The validity of the literature search could be compromised by the use of a few selective sources of research reports, probably those that support a conclusion, or by publication bias, bias toward publishing reports that indicate significant results. Studies that show some kind of positive effect tend to be published more often than those with negative effects (Rosenthal, 1979). This means that if the meta-analysis is restricted to published evidence it will very likely distort the results owing to this publication bias.

The results of the meta-analysis are also likely to be flawed if the information from the individual study reports is incorrectly extracted and coded, dependent effect sizes are treated as though they are independent, and failing to weight effect sizes by their degree of precision before combining them for an overall estimate of effect (Cooper, Valentine, & Charlton, 2000; Wolf, 1984; Johnson & Eagly, 2000; Dunkin, 1996).

However, it appears that much of the criticism levelled at meta-analysis is in line with the inappropriate use of meta-analytic procedures rather than the technique itself as a method of review. Glass, McGaw and Smith (1981) summed up the concerns about meta-analysis by suggesting that the problem with meta-analysis result from uncritical use, and lack of care in conducting the research rather than problems inherent in the procedure itself.

THE STUDY

With two previous meta-analytic reviews providing evidence of the effects of consultation over student ratings on teaching effectiveness, the purpose of this exploratory meta-analysis was not intended to provide an update of the overall efficacy of consultative feedback per se, as there are few new studies in the field. Instead, the main purpose of this review was to identify the consultation practices and strategies that are most likely to maximise the effects of consultative feedback on teaching effectiveness. The findings should represent a valuable contribution to the literature on the strategies that are likely to work best. This is useful information to consider in the design and implementation of feedback interventions, such as the one being examined through this thesis.

This exploratory meta-analysis sought to address the following questions:

1. What is the effect of consultation over student ratings feedback on teaching effectiveness?
2. How effective are the different consultation practices and strategies on maximising the effects of consultative feedback on teaching effectiveness?
3. What are the conditions under which consultative feedback is likely to be maximally beneficial to university teachers?

Criteria for Inclusion

This review focused on the use of consultation with student ratings feedback as an intervention for teaching improvement. All studies from the 1970s, the period recognised as the beginning of formal teaching improvement programmes

(Centra, 1978) to the present were considered. For this review consultative feedback is used to refer to a feedback intervention that incorporates dialogue with a consultant for the analysis and interpretation of teaching behaviours as evidenced by student ratings. As such, studies in which teachers received ratings results on their teaching with explanations on how to read the reports but were not given the opportunity to discuss the ratings or develop improvement strategies with the help of a consultant were not included.

This meta-analytic review was also delimited to randomised controlled trials, in which participants were randomly assigned to either a control or treatment group. Randomised controlled trials are widely accepted as the gold standard in assessing intervention efficacy. It is important to point out that because this review was concerned with only consultative feedback not many studies were available for review in the first place. Delimiting the review to studies with a true experimental design meant that non-experimental studies were excluded but there were not enough of these to form a different category for comparative purposes.

For studies to be included, the treatment had to be consultation in conjunction with student ratings feedback administered in a higher education institution. Studies also had to involve regular staff members or teaching assistants with full responsibility for the classes that formed the context for the study. For studies with multiple treatments only the treatment of interest, consultation, was selected across studies. The treatment, rather than the study was the unit of analysis for determining effect sizes in this review.

Where the treatment group receiving consultative feedback was compared to a group receiving no feedback ratings and a group receiving feedback ratings without

consultation in a study, such reports contributed two effect sizes, one for each type of comparison. Therefore, the control could receive no feedback ratings results or ratings only without consultation. Another criterion was that study results had to be presented in quantitative form to permit computation of an effect size.

For the most part ratings feedback studies are conducted within one semester, with a comparison between mid-term and end of term ratings collected within one semester to assess the treatment effects. A few studies compared the effects across semesters. To keep all measures similar only outcome measures within semesters rather than across semesters were extracted.

Identification of Studies

The studies that met the eligibility criteria were identified through a thorough search of the literature. Three sources were used to identify potentially eligible study reports. First, the bibliographies of the prior meta-analyses on the use of student feedback and other reviews of the literature were examined. Second, a comprehensive computerised search of electronic databases including, Dissertation Abstracts International, Educational Resources Information Centre (ERIC), MEDLINE, PsycFIRST, Social Science Citation Index, and British Education Index, were searched to identify possible studies.

Finally, a manual search of the reference lists of journal and unpublished conference reports retrieved were examined for related studies. Of the 25 related study reports examined only 12 interventions met the inclusion criteria. This sample of study is probably small because this present review is concerned with only consultative feedback investigations. This differs from the earlier meta-analyses that included all

types of ratings feedback studies. Menges and Brinko (1986) reviewed 30 studies but only five combined ratings feedback with consultation. Cohen (1980) reviewed 17 feedback studies but his examination of augmented feedback lumped all types of augmentation not just individual consultation.

As part of the inclusiveness criterion of meta-analysis a thorough search for both published and unpublished studies was undertaken. Attempts to locate unpublished reports proved costly and difficult even with the assistance of the British Library Document Supply Centre, as the majority of the studies were conducted in the United States before the 1990s. Of the unpublished reports retrieved only three met the inclusion criteria for the present review. No attempt was made to contact experts in the field for reports that might be tucked away in file drawers.

Coding Procedures

All eligible study reports were read and coded by the researcher on two separate occasions with an interval of one month to ensure that coding was consistent. Differences were reconciled by a further review of the reports. A coding sheet was designed to facilitate the extraction and recording of information from the studies. This information was coded along five dimensions: (a) study characteristics (e.g. year of publication, author, location, form of publication); (b) participant characteristics (e.g., staff or teaching assistants, ranks, disciplines); (c) form of consultation (e.g. individual or group consultation, consultation design, treatment to control group, the consultant); (d) consultation components (e.g. length of session, use of other sources of evidence on teaching behaviours, development of improvement strategies; augments to consultation); and (e) study design (instrument, sample size, consultation process). This

detailed coding allowed for the identification of the characteristics of effective consultation.

Noteworthy, is that, the common practice in meta-analytic procedures is for at least two coders to independently extract information from the studies to coding forms. Alternatively, if one person codes the studies, in the first instance, a second person may independently code a random sample. The codes assigned by each coder are then compared to see how much they agree. Disagreements are usually resolved through discussion and possible recoding. The use of this procedure is seen as a way of demonstrating reliability and allowing for the computation of an "intercoder" agreement reliability estimate.

A limitation of this review, therefore, is that only one person coded the research reports. Being cognizant of the seriousness of the Type 3 errors, erroneous detailing, (Dunkin, 1996), that might be made in a research synthesis, and which threatens the validity of the meta-analysis, due care was exercised in the coding process to minimise errors.

Outcome Measures

Teaching effectiveness as measured by student ratings feedback was the primary outcome considered in this review. As teaching is considered to be a multidimensional activity, ratings questionnaires usually contain a number of components used to measure the construct effective teaching. It is therefore common in student feedback research for results to be reported on different components of effective teaching as measured by ratings forms. Where multiple effect sizes were calculated for individual studies these were averaged to ensure that only one estimate of treatment effect from

each report contributed to the review (Hedges & Olkin, 1985; Gleser & Olkin, 1994; Rosenthal 1994).

To maintain the assumption of independence of effect sizes that underlies the validity of meta-analysis, each unit of analysis should contribute just one effect size estimate to the overall effect estimate. Multiple effect sizes as indicators of the same construct by the same subject per sample are seen as non-independent as they are derived from the same group and are conceptually and statistically correlated.

Sophisticated statistical models have been suggested to deal with the problem of dependent effect sizes but due to their complexity are rarely practiced (Gleser & Olkin, 1994; Lipsey & Wilson, 2001; Cooper, Valentine, & Charlton, 2000). While very important as measures of teaching effectiveness, outcomes such as student achievement, attitude towards subject and/or progress were not considered. Data for these outcomes were not fully reported on or consistently examined in many of the studies.

Moderator Variables

Given that the primary objective of this review was to identify the strategies and practices for most effective consultative feedback, the influences of potential moderator variables were of primary interest. A moderator variable is a variable that influences the relationship between the independent and the dependent variable. The practice of examining moderator variables provides an explanation for the variability in effect sizes from the different studies. As Rosenthal and DiMatteo (2001) pointed out, this examination also allows for further testing of details of theories, and a better understanding of the research literature.

Although the common practice in meta-analytic reviews is to search for moderator variables when there is an indication of significant heterogeneity of a sample of effect sizes, Rosenthal (Rosenthal, 1995; Rosenthal & DiMatteo, 2001) noted that even if the test of heterogeneity is not statistically significant, planned searches for moderators can and should be conducted among the obtained effect sizes. It is reasoned that the distribution of homogenous effect sizes could contain one or more contrasts that are substantially and statistically significant. In line with the objective of the present review a series of contrasts were planned *a priori*.

Analysis

The index of standardised mean difference effect size calculated was the unbiased estimator *d* (Hedges & Olkin, 1985; Johnson & Eagly, 2000). This index of effect size is derived from the difference between the means of the experimental (X_E) and control ($\overline{X}_C$) groups divided by the pooled standard deviation (S_P) of the sample and corrected for small and unequal sample bias (*J*). This is defined as

$$d = \frac{\overline{X}_E - \overline{X}_C}{S_P} J$$

where the pooled standard deviation, S_P, is found from,

$$S_P = \sqrt{\frac{(N_E - 1)(S_E)^2 + (N_C - 1)(S_C)^2}{N_E + N_C - 2}}$$

and *J*, the correction factor from

$$J = 1 - \frac{3}{4(N_C + N_E - 2) - 1}$$

where *N* is the sample size, and *S* is the standard deviation of the different groups.

Where results were reported as *F* values estimates of effect sizes were computed according to the procedures described by Glass, McGaw, and Smith (1981).

For purposes of interpretation, effect size is an estimate of the magnitude of the effect of a treatment indicating whether and by how much performance of the group receiving the intervention exceeds the performance of the control group. Effect sizes are expressed in standard deviation units. For instance, a standardised mean difference effect size of +1.00 indicates that, on average, the performance of the group receiving the treatment exceeds the performance of the control group by 1 standard deviation. A negative effect size of –1.00 means that there is a 1 standard deviation advantage for the control group over the group receiving the treatment. Reporting effect sizes in terms of standard deviations is, however, not readily interpretable and is therefore not often used.

More common, is use of the interpretation guidelines established by Cohen (1977, 1988, 1992) based on the concept of statistical power, where an effect size of .20 is defined as small, .50 as medium, and .80 or above as representing a large effect. Effect sizes may also be interpreted as the percentile standing of the average person in the treatment group relative to the average person in the control group who received no treatment. In this case, a standardised mean difference effect size of 0.0 indicates that the mean of the treatment group is at the 50th percentile of the control group. An effect size of .80 would therefore move the mean of the average treated person to the 79th percentile relative to the average person in the control group.

Another approach to interpreting effect sizes, on the assumption of normal distribution of scores with equal variability for the treatment and control group, is in terms of non-overlapping distributions of scores for the treated group with those of the control group, referred to as *U*3 by Cohen (1977, 1988). In this way, a standardised

mean difference effect size of .60 in a meta-analysis of treatment effects is depicted as a success rate of 73% for the treatment group compared with 50% in the control group.

It is important to recognise, however, that an effect size takes on meaning only within a particular context (Rosenthal, 1994). In other words, the practical significance of an effect size estimate depends on the nature of the outcome and its importance in the particular field in which the research was conducted. So, an effect size of .20 might be seen as "poor" in one context but "good" in another context.

The information required to compute individual effect sizes was entered into a Microsoft EXCEL spreadsheet prepared by Robert Coe a senior researcher in the School of Education, University of Durham. The mean effect size is calculated as a weighted average with each effect size, *d*, being weighted by the inverse of its variance, a procedure that gives proportionally greater weight to effect sizes based on larger samples under the assumption that larger samples provide a more precise estimate of the population value (Shadish & Haddock, 1994; Rosenthal, 1991; Hedges & Olkin, 1985).

Both unweighted and weighted effect sizes can be used to calculate the mean effect. In the unweighted procedure each effect size is given equal weight in calculating the average effect. With the weighted procedure each independent effect is weighted by its sample size. This ensures that the influence of each study on the overall results of the meta-analysis is determined by the precision of its estimate. The weighting option is generally preferred because it is more precise.

As effect sizes are imprecise, they will vary somewhat even if they all estimate the same underlying population. To test whether observed differences in effect sizes are statistically significant the 'homogeneity test' also known as the 'heterogeneity test' was conducted to assess the null hypothesis that all effect sizes are homogeneous. If the

effect sizes are homogeneous this means they are from a single population and that they can be averaged to represent the population value.

More specifically, the purpose of the homogeneity test is to determine whether sampling error alone accounts for the variation in effect sizes or whether features of studies, samples, treatment designs or outcome measures also contributed to the variation. The homogeneity test relies on the *Q*–statistic that has an approximate chi-square distribution with $k - 1$ degrees of freedom, where k is the number of independent effect sizes. Computations for the mean weighted effect sizes and *Q*–statistic were carried out using a fixed effect model of meta-analysis with the statistical software *MetaWin* 2.0 (Rosenberg, Adams & Gurevitch, 2000).

At present there is debate about threats to the validity of meta-analysis conclusions in terms of the use of either the fixed versus random effects model of meta-analysis (Cooper, Valentine & Charlton, 2000; Lipsey & Wilson, 2001). The fixed effects model is based on the assumption that there is one true effect size shared by all studies or group of studies. Analysis with the fixed effects model therefore only accounts for the variability in effect sizes due to sampling error in findings between studies. Whereas, the random effects model assumes variability between effect sizes due to sampling error and random variability from other sources in the population of effects, that is, unique differences in the set of effect sizes.

The disadvantage of the fixed effect model is that it does not permit generalizations to other studies other than those in the sample. The random effects approach, though a less powerful test of the null hypothesis, permit generalization to studies not in the same group (Rosenthal & DiMatteo, 2001). A significant '*Q*',

according to Lipsey and Wilson (2001), challenges the assumption of the fixed effect model.

At the same time, a non-significant '*Q*' does not provide justification for the use of a fixed effects model, as a small number of effect sizes, especially if they are based on small samples will not provide enough statistical power to reject homogeneity even when there is much variability among the effect sizes. Some meta-analysts argue, however, that the fixed effects model can be applied, if a thorough and appropriate search for influences on effect sizes is part of the analytical strategy.

If the value of '*Q*' is below the critical chi-square value then the distribution of effect sizes about their mean is no greater than that expected from sampling error alone. In this case the null hypothesis is not rejected as '*Q*' is not statistically significant. As previously mentioned, a non-significant *Q* does not preclude the testing of moderators. A statistically significant '*Q*' results when the *Q*–statistic exceeds the critical chi-square value. This implies that there are differences among the estimates of effects that are probably due to systematic differences among the design of the intervention, participants or some other feature of the studies, and is not due just to sampling error. Moderator variables should be investigated to explain this variation.

Results

Study Characteristics

The studies included in this meta-analysis fell into two main categories based on whether or not teachers in the control group were given ratings feedback. For nine studies the feedback intervention group was compared with a control group that

received no ratings feedback at all. Three studies compared the feedback intervention group against a control group that received ratings feedback but no consultation. Table 5.1 presents a summary of the characteristics of the studies included in this review together with the effect sizes computed for each study. The typical study was published during the period 1975-1986, comprised staff members, and used an approach that lasted for one hour on average. Eleven of the twelve studies were carried out in North America, which raises questions about generalisations of findings.

Table 5.1 also shows the effect sizes vary in magnitude but all were positive indicating support for the effects consultative feedback on teaching effectiveness. Further examination reveals that four studies reported large effect sizes, that is $d > 0.80$, five studies had effect sizes in the medium category, $d > 0.50$, and only two studies reported small or trivial effects. Eleven studies engaged teachers in individual consultation with only one using a group-based model of consultation, which itself reported a moderate effect size of $d = 0.68$.

Table 5.1
Summary of Studies included in the Meta-Analysis ($N = 12$)

Study	Source	Subjects	Discipline	Consultant	Consultation Form	Time	Supplements	Data collection	Improvement Target	Instrument	*n*	*d*
Atchison, 1987 (A), USA	Dissertation	Staff	Various	Expert	Advisory (individual)	2 hour	No	Mid-term ratings	No	Local	20	0.51
Atchison, 1987 (B), USA	Dissertation	Staff	Various	Expert	Advisory (individual)	2 hour	No	Mid-term ratings	No	Local	20	0.66
Bray & Howard, 1980, USA	Journal	Teaching assistants	Various	Expert	Advisory (individual)	2 hours	Self-ratings	Mid-term ratings	No	Standardised	19	0.86
Erickson & Erickson, 1979 USA	Journal	Staff	Various	Expert	Educational (individual)	2+ hours	Self-ratings	Mid-term ratings, observation, interview, videotaping	Yes	Standardised	31	1.14
Erickson & Sheehan, 1976 USA	Report	Staff	Various	Graduate student	Educational (individual)	2+ hours	Self-ratings	Mid-term ratings, interview, observation, videotape	Yes	Standardised	27	0.68
Hampton, 2001, USA	Dissertation	Teaching assistants	Computer Science, Chemistry	Expert	Advisory (individual)	2 hours	Norms data Printed material	Mid-term ratings, observation	No	Local	37	0.72
Hoyt & Howard, 1978, USA	Journal	Staff	Various	Peers	Educational (group)	2+ hours	No	Mid-term ratings, observation, videotaping, discussion on teaching	No	Local	31	0.68
McKeachie, et al. 1980 (A) USA	Journal	Staff, TAs	Psychology	Peer	Advisory (individual)	1 hour	Self-ratings Norms data	Mid-term ratings, interview	Yes	Standardised	27	1.05
McKeachie, et al. 1980 (B) USA	Journal	Staff, TAs	Psychology	Peer	Advisory (individual)	1 hour	Self-ratings Norms data	Mid-term ratings, interview	Yes	Standardised	26	0.85
Marsh & Roche, 1993 Australia	Journal	Staff	Various	Expert	Advisory (Individual)	2 hours	Self-ratings, Norms data, printed material	Mid-term ratings, interview	Yes	Standardised	53	0.01
Overall & Marsh, 1979, USA	Journal	Teaching assistants	Computing	Expert	Diagnostic (individual)	1 hour	Norms data	Mid-term ratings	No	Local	15	0.18
Payne & Hobbs, 1979 USA	Journal	Staff	Education	Expert	Diagnostic (individual)	1 hour	Self-ratings, Norms data	Mid-term ratings	No	Local	78	0.46

Note. d = effect size

Distribution of Effect Sizes

Figure 5.1 displays a stem and leaf plot of the distribution of effect sizes determined by the meta-analysis. The stem and leaf plot, from Tukey's (1977) exploratory data analysis, is a graphic display of the shape of the distribution while giving detail concerning individual values.

Stem	Leaf
1.1	4
1.0	5
0.9	
0.8	5, 6
0.7	2
0.6	6, 8, 8
0.5	1
0.4	6
0.3	
0.2	
0.1	8
0.0	1

Figure 5.1 Stem-and-Leaf Plot Showing Distribution of Mean Effect Sizes (N=12)

It can be seen from the stem and leaf diagram that the effect sizes ranged from 1.14 to 0.01 and centred approximately at 0.68. The distribution is approximately symmetrical but reveals a wide gap in the effect sizes between study number 10 (d = 0.46) and the next two studies, study number 11 (d =0.18) and study 12 (d = 0.01), representing potential outliers, that is, extreme values that are not consistent with the other values and which have the potential to distort the analysis.

The effect sizes and their associated 95% confidence intervals along with the overall effect size, before outlier procedures, are displayed in Figure 5.2. It can be seen that the confidence interval for 8 studies included zero, and only 4 studies showed 95%

confidence intervals that might include the observed magnitude of effect and therefore considered to be statistically significant.

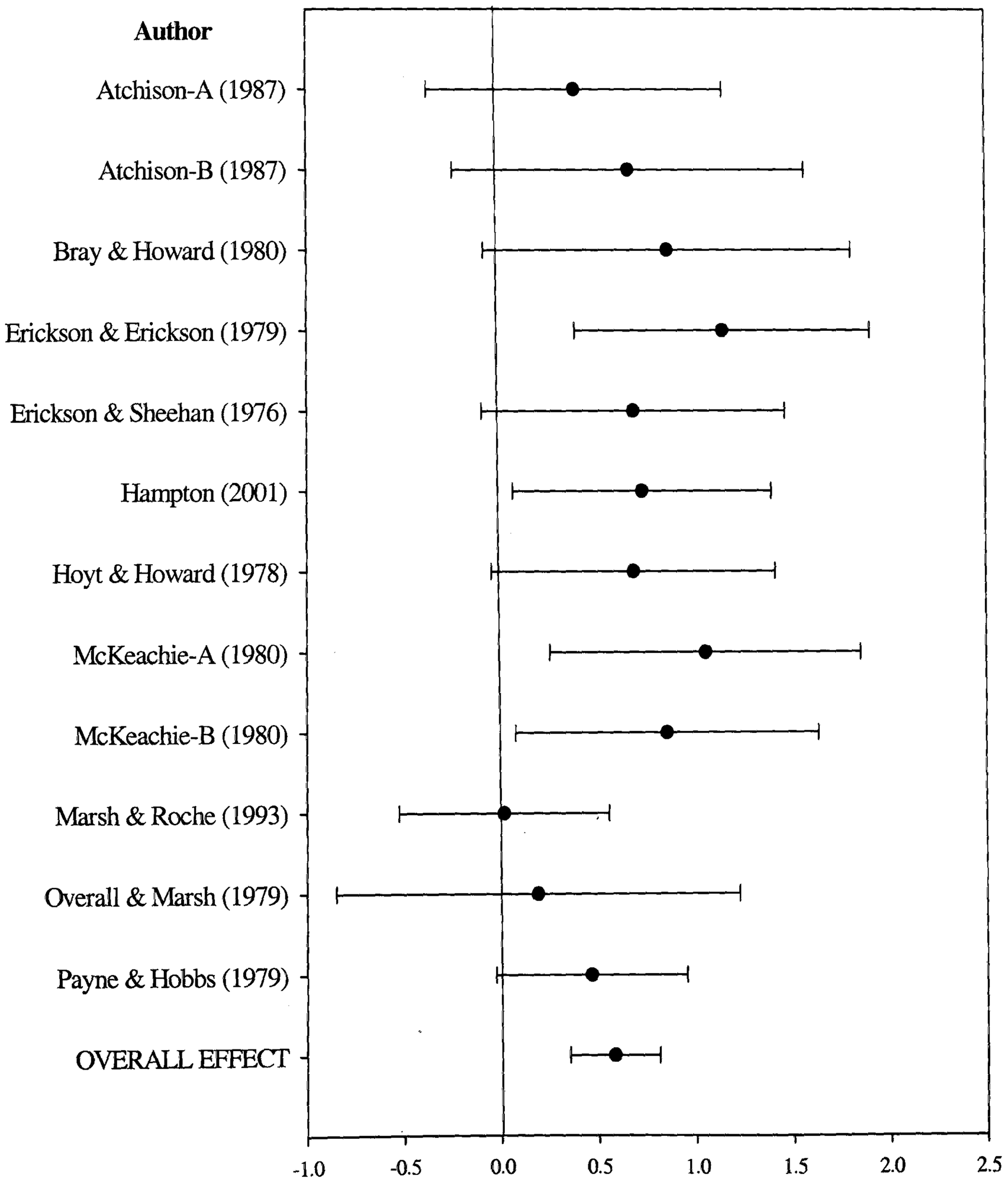


Figure 5.2 Effect Sizes, Overall Effect Size and their 95% Confidence Intervals by Study (n = 12), (Before outlier procedures)

Outlier Procedures

The overall mean weighted effect before outlier procedures was $d_+ = 0.58$ with the 95% confidence interval being 0.35 to 0.82. The test statistic for heterogeneity, $Q = 9.94$, was not larger than the chi-square value ($\chi^2(11) = 19.68$) . This means that the null hypothesis of homogeneity could not be rejected, suggesting instead that variation in effect sizes could be explained by sampling error. Yet it must be recognised that the sample was probably too small to generate enough statistical power to reject the null hypothesis in the first place.

The potential outliers, as identified from Figure 5.1, were checked for their influence on the overall mean effect (Greenhouse & Iyengar, 1994; Hedges & Olkin, 1985). When the study with effect size $d = 0.18$ (Overall & Marsh, 1979) was held out the overall mean was $d_+ = 0.60$, indicating that it was not unduly influencing the mean effect. However, when the study with effect size $d = 0.01$ (Marsh & Roche, 1993) was held out the mean effect rose to $d_+ = 0.69$, indicating that this particular effect size was artificially deflating the mean effect.

This indicates that the mean effect size, $d_+ = 0.58$, may not be an accurate representation of the outcome of the studies in the sample. This deviation was considered large enough to distort the analysis and interpretation of the overall effect of consultative feedback. The outlying value was removed. The overall mean effect along with the Q–statistic was recalculated to give a better representation of the findings about consultative feedback.

The elimination or adjustment of outlying values before proceeding with the analysis of results is a common practice in meta-analysis. It is not unusual to set aside up to 20% of the data at any one time, to provide a better fit to a model

(Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Lipsey & Wilson (2001) added that as the purpose of meta-analysis is to provide an adequate representation of treatment effects no purpose is served by including notably discrepant values from those found. These values are recognised as not only unrepresentative of the results of the research but may even be spurious.

Consequently, outliers are usually removed or adjusted to make the distribution of effects more normal and to mitigate the distorting effect of these values on the mean effect, and on moderator analyses (Cooper, Valentine & Charlton, 2000). The process of adjusting and recoding outliers to, say, the next nearest value, so that they more closely conform to the overall distribution is called "winsorising" the data points.

An examination of the study removed, Marsh and Roche (1993), revealed that it was in fact characteristically different from the other studies. On one level, it was the only one conducted outside of North America. Up to the time of that study, the use of student ratings feedback was more commonplace in North American universities than in European and Australian universities. The use of student ratings questionnaires was a novelty in the institution.

On another level, the effect was essentially zero ($d = 0.01$) despite the use of a particularly strong intervention when compared to the others. This particular intervention provided participants with printed teaching improvement suggestions keyed to the dimensions of the ratings form before the consultation session, used of a reliable instrument, set improvement goals, used self-ratings, and normative data, and engaged participants in a relatively long consultation session.

The researchers themselves attributed the reduction of experimental/control comparisons to the novelty, for both students and teachers, of using student feedback

ratings forms; the *John Henry* effect, as control teachers also administered self-ratings that might have caused them to scrutinise their teaching more than before because it was their first opportunity to examine their teaching. Besides, many teachers indicated they had volunteered to obtain positive ratings to support applications for promotion.

Publication Bias

An influential source of upward bias in the mean effect size of a meta-analysis is publication bias, from an under-representation of unpublished studies or studies with negative results in the meta-analysis. Studies with statistically significant or 'favourable' results are more likely to be published and available for review than studies with non-significant or 'negative' results. Rosenthal (1979) termed this the "file drawer problem" for, at worst, journals are filled with 5% of the studies with Type I errors while 95% of studies that show non-significant results are buried in file drawers. In fact, Lipsey and Wilson (1993) showed that on average published studies have larger mean effect sizes than unpublished ones.

One method used to detect the presence or absence of publication bias is inspection of a 'funnel plot' (e.g. Light, Singer, & Willett, 1994; Ferrer, 1998). A funnel plot is a scatterplot of effect sizes plotted against the sample size of the studies. It is so called because studies with smaller sample sizes will display greater variability in effect sizes at the bottom of the graph with the variability decreasing among larger studies. In the absence of publication bias the graph should take the shape of an inverted funnel. Figure 5.3 represents a funnel plot for the sample of studies in the present meta-analysis. Inspection of the plot shows no clear indication of bias from an omission of unpublished studies or studies with non-significant results. It is difficult,

however, to detect publication bias with a funnel plot when the number of studies is small.

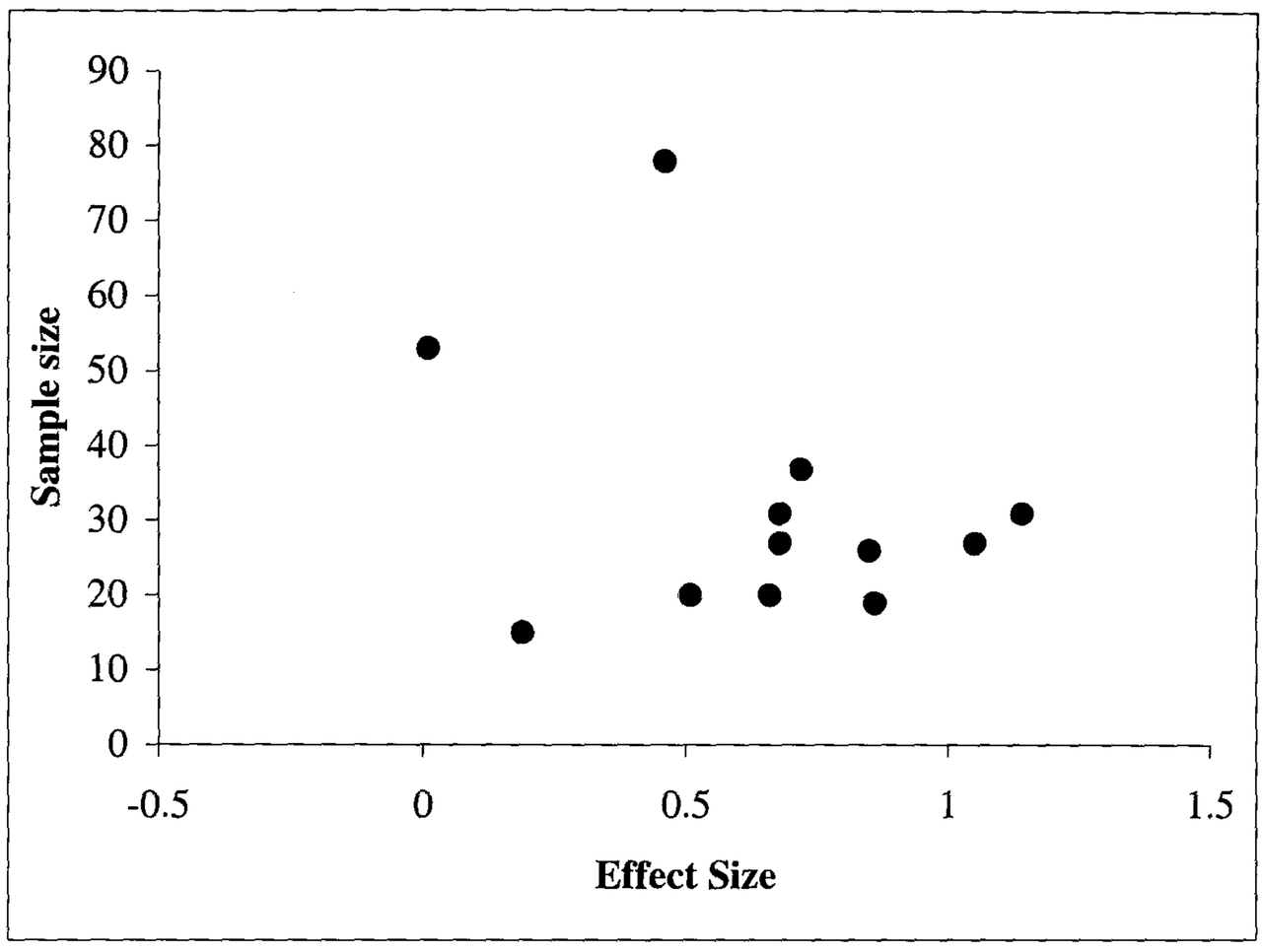


Figure 5.3 Funnel Plot of Effect Sizes ($N = 12$)

Another method used to determine publication bias is the calculation of a *fail-safe* number. The fail safe *N* (Orwin, 1983, in Lipsey & Wilson, 2001) is an estimation of the number of non-significant, unpublished, unretrieved, or studies with zero effect that would be needed to change the results found in a meta-analysis and reduce the overall effect size estimate to a mean effect size of $d = 0.20$, which represents a small effect size.

If the number is large relative to the number of observed studies, one can feel fairly confident that the observed results, even with some publication bias, can be treated as a reliable estimate of the true effect. However, the level of confidence attached to the obtained fail-safe number is dependent on the field and how likely it is

for unpublished data to exists but there are no hard and fast rules. Computation for the *fail-safe* number used the formula

$$k_0 = k\left[\frac{ES_k}{ES_c} - 1\right]$$

where k_0 is the number of effect sizes with a value of zero needed to reduce the mean effect size to ES_c, the criterion effect size, k is the number of studies in the meta-analysis with weighted mean effect ES_k, to determine how many unretrieved studies, for example, would be needed to change the results found.

The results of the fail-safe N analysis indicated that a total of 23 independent studies not retrieved for this review would be needed to change the results of this meta-analysis. It seems highly improbable that this many studies, almost twice as many of the studies found for this present meta-analysis, would be filed away because of zero effect or null results.

Effect on Teaching Effectiveness

Table 5.2 reports the unweighted and weighted mean effect sizes, d_+, and standard deviations computed for studies reviewed after outlier procedures.

Table 5.2
Overall Effect of Consultation on Teaching Effectiveness (n = 11)

Category	n	Effect Size (d)		95% confidence interval for weighted effects		Homogeneity (Q)
		Unweighted	Weighted	Lower	Upper	
No ratings feedback control	8	0.70	0.68	0.37	1.00	4.21
Ratings feedback only control	3	0.73	0.73	-0.30	1.76	0.12
Overall Effect	11	0.71	0.69	0.43	0.95	4.36

Note: Q = homogeneity statistic for mean effect size

As evidenced in Table 5.2 the overall mean weighted effect of consultative feedback on teaching effectiveness is $d_+ = 0.69$ ($Q = 4.36$, 95% CI = 0.43 to 0.95), reflecting a moderate and positive association between consultative feedback and gains in teaching effectiveness. On Cohen's (1977, 1988, 1992) proposed guidelines, an effect of this magnitude is large enough to be visible to the naked eyes of a careful observer. Moreover, as the 95% confidence interval, does not include zero it might be reasonably concluded that on average, consultative feedback is statistically significantly associated with improvement in teaching effectiveness.

Figure 5.4 illustrates the distribution for the overall effect size, $d_+ = 0.69$. If this effect size is used to compare the teaching effectiveness of teachers who received consultative feedback with those who did not—corresponding of Cohen's (1977, 1988) *U*3 measure—it means that 75% of teachers who received support through consultative feedback improved in their teaching effectiveness beyond that of the average teacher who did not receive consultative support.

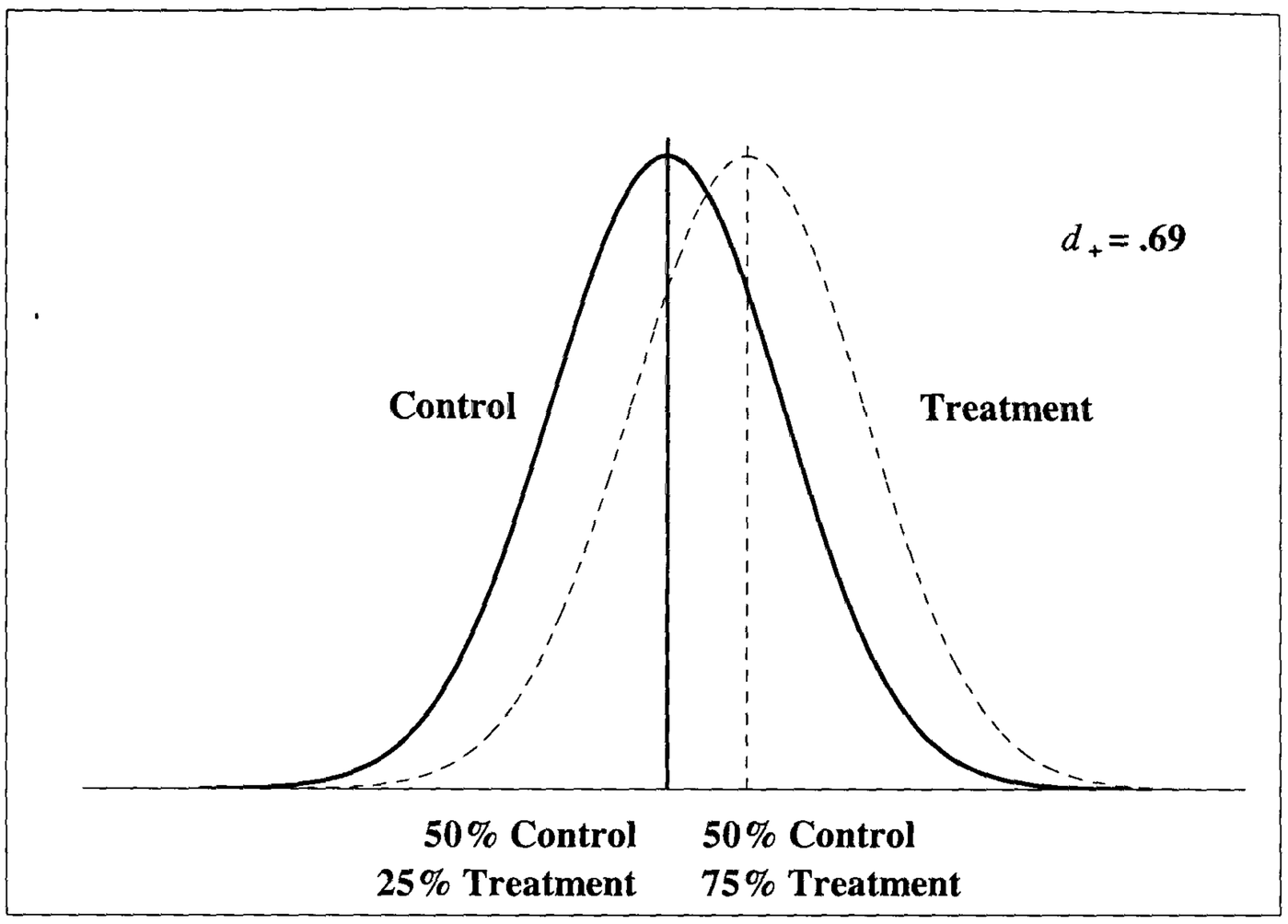


Figure 5.4 Distribution for the Overall Mean Effect of Consultative Feedback

The results in Table 5.2 also report a weighted mean effect size of $d_+ = 0.68$ ($Q = 4.21$, 95% CI = 0.37 to 1.00) for the eight studies that compared consultative feedback with no ratings feedback to teachers. For the 'ratings only' studies the mean effect size is $d_+ = 0.73$ (95% CI = –0.30 to +1.75). This effect size approached Cohen's (1977, 1988) threshold of .80 for a large effect but this result is unreliable as the confidence interval included zero, indicating that the true effect could easily be zero. As this category included only three studies, it is suspected that the statistical power was too low to detect a significant effect at the 95% confidence level. A more reliable effect was obtained at the 90% confidence level (CI = .03 to 1.43).

The test for heterogeneity of effect sizes showed no significant heterogeneity ($Q = 4.36$, $\chi^2(10) = 18.31$). This non-significant 'Q' indicates that any observed variability in the effect sizes probably resulted from sampling error alone. Following Wilson and Lipsey (2000), even though the results found no indication of heterogeneity,

planned moderator analyses were carried out on the conceptual ground that the practices and strategies consultants adopt are associated with the effects and goes beyond sampling error. In any case, the small sample of effect sizes probably yielded low statistical power to reject the null hypothesis.

Moderator Analyses: Consultation Practices and Strategies

Exploration of the influence of different consultation practices on teaching performance was limited by the sparse descriptions in the research reports. On the information that could be extracted 30 effect sizes were computed from the 11 studies, so 14 categorical models were fitted to assess the influence of the different strategies and practices adopted by consultants on the effectiveness of the process. It is important to recognise that although studies used the experimental design, the moderating factors do not provide evidence of causation, only an association.

The search for moderator variables of effect sizes is a correlational investigation because the moderators examined were not randomly assigned to examine interaction (Miller & Pollock, 1994). On this basis, even if differential effects are found this does not establish that the variables of interest caused or mediated the effect, as there is the familiar concern about the presence of confounding variables. Further, subdivision of the sample of studies will considerably reduce statistical power leading to unreliable effect sizes. That noted, several variables were examined as possible moderators of teaching effectiveness. Table 5.3 presents the weighted mean effect size for each category of moderating variables along with the test for homogeneity statistic and the 95% confidence interval.

Overall, the moderator analyses suggest that consultative feedback is effective across a variety of conditions and contexts. But there is an even stronger suggestion about the possibility of certain strategies to lead to greater gains in teaching effectiveness. Nonetheless, the findings should be interpreted with care and conclusions are tentative as the sample of studies used for the analyses was small and no statistically significant differences were found.

Table 5.3
Mean Effect Sizes as a Function of Moderator Variables

Variables	*n*	d_+	95 % CI	Q_B	Q_W
		Process Variable			
Approach				2.26	
Educational	3	0.83	-0.13 to 1.79		0.94
Advisory	6	0.78	+0.34 to 1.22		0.93
Diagnostic	2	0.41	-2.41 to 3.24		0.23
Classroom Observation				0.53	
With	4	0.80	+0.20 to 1.40		1.02
Without	7	0.63	+0.26 to 0.99		2.81
Videotape Recording				0.54	
With	3	0.83	-0.13 to 1.79		0.94
Without	8	0.64	+0.32 to 0.96		2.88
Interview				2.11	
With	4	0.93	+0.29 to 1.58		0.80
Without	7	0.57	+0.23 to 0.93		1.44
		Context Variable			
Duration of Consultation				0.63	
1 hour	4	0.61	+0.04 to 1.17		2.49
2 hours	4	0.69	+0.01 to 1.36		0.29
Over 2 hours	3	0.83	-0.13 to 1.79		0.94
Consultant				0.61	
Expert	7	0.64	+0.29 to 0.98		3.31
Peers	3	0.85	-0.15 to 1.84		0.44
Nature of Consultation				0.34	
Discipline-based	5	0.63	+0.20 to 1.07		2.58
Generic	6	0.77	+0.33 to 1.21		1.44
Participants				0.92	
Staff	6	0.65	+0.27 to 1.03		2.31
Teaching Assistants	3	0.64	-0.43 to 1.71		1.02
		Content Variable			
Normative Data				0.34	
With	5	0.63	+0.19 to 1.07		2.58
Without	6	0.77	+0.33 to 1.21		1.44
Self-Ratings				0.42	
With	6	0.75	+0.37 to 1.13		0.85
Without	5	0.60	+0.08 to 1.12		3.08
Improvement Goals				0.24	
With	4	0.78	+0.10 to 1.46		2.24
Without	7	0.66	+0.32 to 1.00		1.87
		Study Variable			
Ratings Forms				2.48	
Standardised	5	0.92	+0.40 to 1.44		0.82
Local	6	0.55	+0.16 to 0.93		1.05
Publication Type				0.06	
Journal	7	0.71	+0.37 to 1.06		4.18
Dissertation/Report	4	0.66	+0.01 to 1.30		0.14
Publication Year				0.49	
1975-1979	5	0.62	+0.18 to 1.06		2.97
1980-1987	5	0.80	-0.25 to 1.35		0.89

Note: d_+ = Weighted mean effect size; CI = Confidence interval; Q_B = Q-statistic of between group differences; Q_W = Q-statistic of within group differences

Effect of Process Variables

Approach to Consultation

The search for moderator variables started with how researchers operationally defined consultation. From the details given in research reports the approach to consultation was labelled diagnostic, advisory, or educational, on the basis of their components. The diagnostic approach was defined as the consultation process that simply involved interpretation of ratings with some discussion and recommendations for improvement. The 'advisory' model built on the diagnostic approach and included discussion about the class, use of self-ratings, use of at least one additional source of evidence on teaching, as well as the development of improvement strategies. Meanwhile, the 'educational' model would contain all the elements of the diagnostic and advisory models in addition to the use of educational activities, such as seminars and workshops on teaching issues, as part of the consultation process and would normally extend beyond two hours.

As reported in Table 5.3 the use of the advisory approach to consultation was positive with a moderate to large effect on teaching performance ($d_+ = 0.78$, CI = 0.34 to 1.22). Use of the educational approach yielded only a slightly larger effect on teaching effectiveness but this association was not reliable as the confidence interval contained zero ($d_+ = 0.83$, CI = -0.13 to 1.79). By contrast, the use of the diagnostic approach to consultation reported only a modest effect on teaching effectiveness ($d_+ = 0.41$, CI = -2.41 to 3.24).

Figure 5.5 provides a graphical view of the effects of using additional information on teaching behaviours collected through sources such as classroom observation, videotape recording, and interviews.

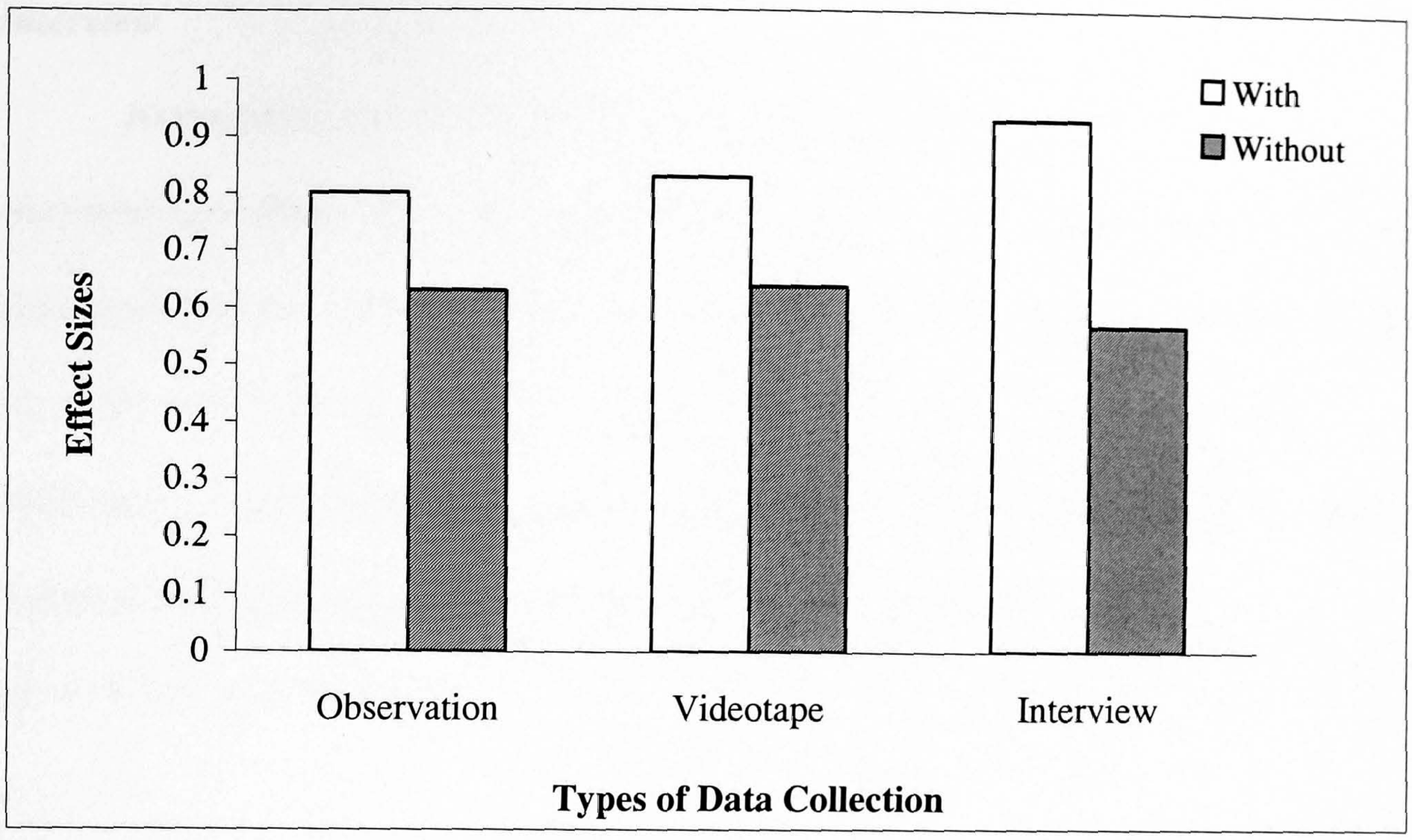


Figure 5.5 Effect Sizes as a Function of Additional Information on Teaching

Observation

Use of classroom observation alone, as an additional source of information to supplement the data from student ratings, resulted in a mean effect of $d_+ = 0.80$ (CI = 0.20 to 1.40) suggesting slightly more benefits for teachers when this method was used compared with consultation sessions that did not incorporate it as part of the process ($d_+ = 0.63$, CI = 0.26 to 0.99).

Videotape Recording

Videotaping is almost always paired with observation to gather additional information on teaching. Use of this method is shown to be very beneficial to increasing teaching effectiveness, ($d_+ = .83$, CI = -0.13 to 1.79), but this effect is not reliably different from zero. Consultation sessions that did not gather additional information with videotape and observation reported a slightly smaller mean effect ($d_+ = .64$, CI = 0.32 to 0.96).

Interview

As could be established from the research reports, use of the strategy of interviewing teachers about the class and their progress with the class before discussing the ratings and developing improvement strategies resulted in a large estimate of effect (d_+ = .93, CI = 0.29 to 1.58). When the consultant did not give the teacher the opportunity to talk about teaching in the particular class before assisting the teacher to interpret the ratings feedback, the effect was barely moderate but positive (d_+ = .57, CI = 0.23 to 0.93).

Effect of Context Variables

Results from the analysis of the relationship between the context of consultation and teaching effectiveness are also reported in Table 5.3. Figure 5.6 presents a graphical summary of the results.

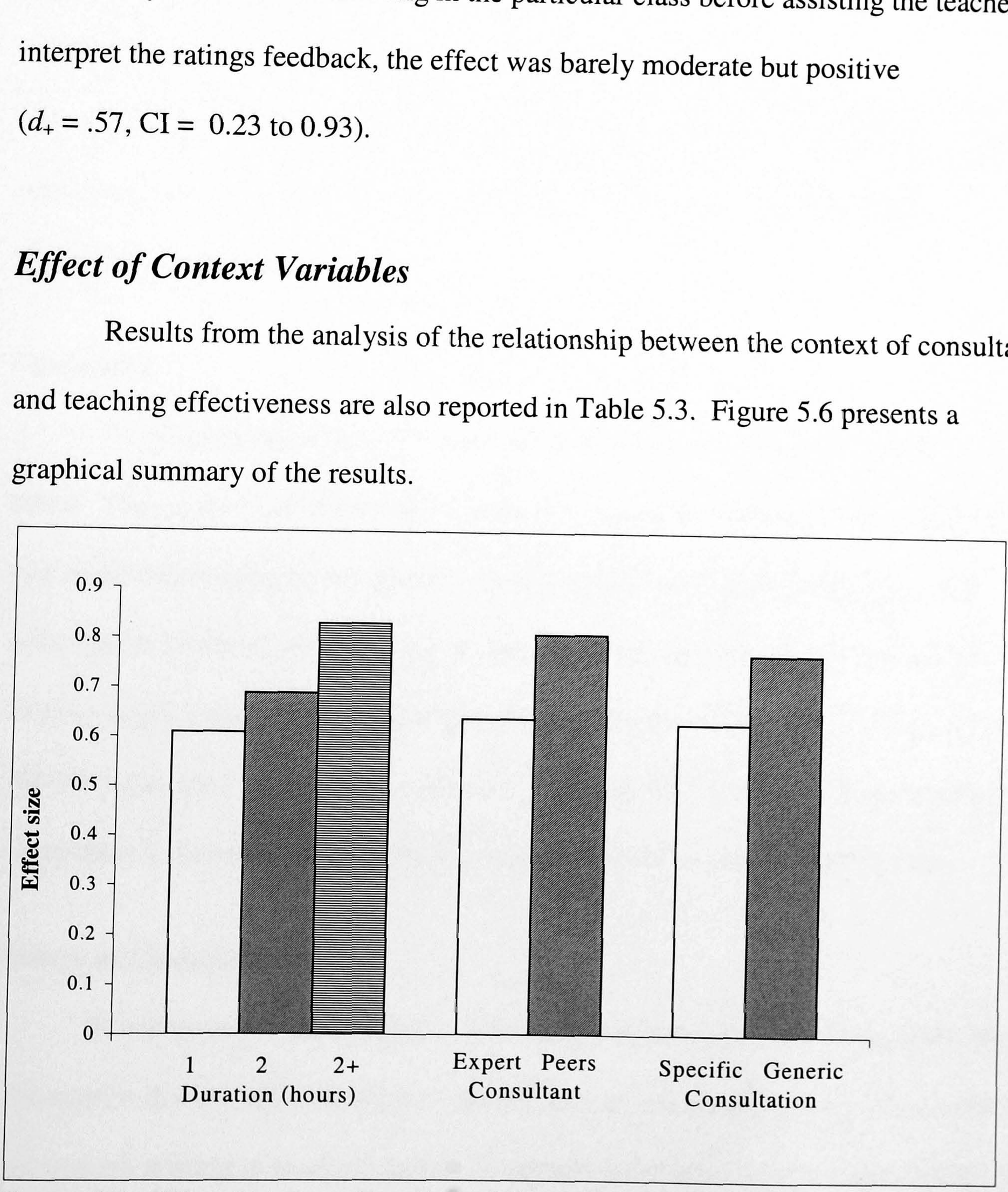


Figure 5.6 Effect Sizes as a Function of Consultation Context Variables

Duration of Consultation

Comparisons of the effects of the duration of consultation were based on assumptions made from descriptions of activities in the studies where this was not explicitly stated. As Figure 5.6 illustrates, teachers received some benefits when the consultation process lasted for up to two hours (d_+ = .69, CI = 0.01 to 1.36) or even one hour (d_+ = .61, CI = 0.04 to 1.17). However, consultation activities that engaged teachers for over two hours resulted in slightly more benefits (d_+ = .83, CI = -0.13 to 1.79) but this association was not reliable with the confidence interval containing zero. Sessions for over two hours corresponded with the educational approach to consultation.

Consultant

To examine the effects of the status of the consultant the default position was taken. That is, the consultant was an 'expert' when reports did not explicitly identify this or the information on the authors was not available. Figure 5.6 displays that consultative feedback activities that involved peers as consultants was associated with a slightly larger mean effect on teaching effectiveness (d_+ = .85, CI = -0.15 to 1.84), than did those that used 'experts' (d_+ = .64, CI = 0.29 to 0.98). The effect size for peer as consultant is somewhat doubtful because the confidence interval included zero.

Nature of Consultation

Investigation of the influence of the nature of the consultation was based on the assumption that studies conducted in specific disciplines are likely to involve pointed discussions relating to teaching in that discipline than studies in which the consultant had to deal with teachers from a variety of disciplines. The findings seem to suggest

that the effect of the assumed discipline-specific consultation was positive and moderate (d_+ = .63, CI = 0.20 to 1.07). This effect was slightly lower than that for consultation sessions where generic type teaching issues might have been discussed (d_+ = .77, CI = 0.33 to 1.21).

Participants

Consultation was conducted with either full time staff or teaching assistants with major responsibility for the class. As shown in Table 5.3 there is actually no difference in the estimated effect sizes for studies that assessed the intervention with staff members (d_+ = .65, CI = 0.27 to 1.03) and those that used teaching assistants (d_+ = .64, CI = -0.43 to 1.71).

Effect of Content Variables

To explore whether the inclusion of teacher self-ratings, normative data, and improvement targets influenced outcome, studies reporting such measures were analysed. Figure 5.7 graphically summarises the results.

Normative Data

From the results in Table 5.3 it can be seen that the practice of discussing a staff member's teaching in relation to normative data yielded a slightly lower effect (d_+ = .63, CI = 0.19 to 1.07), compared to when normative comparisons were not made (d_+ = .77, CI = 0.33 to 1.21).

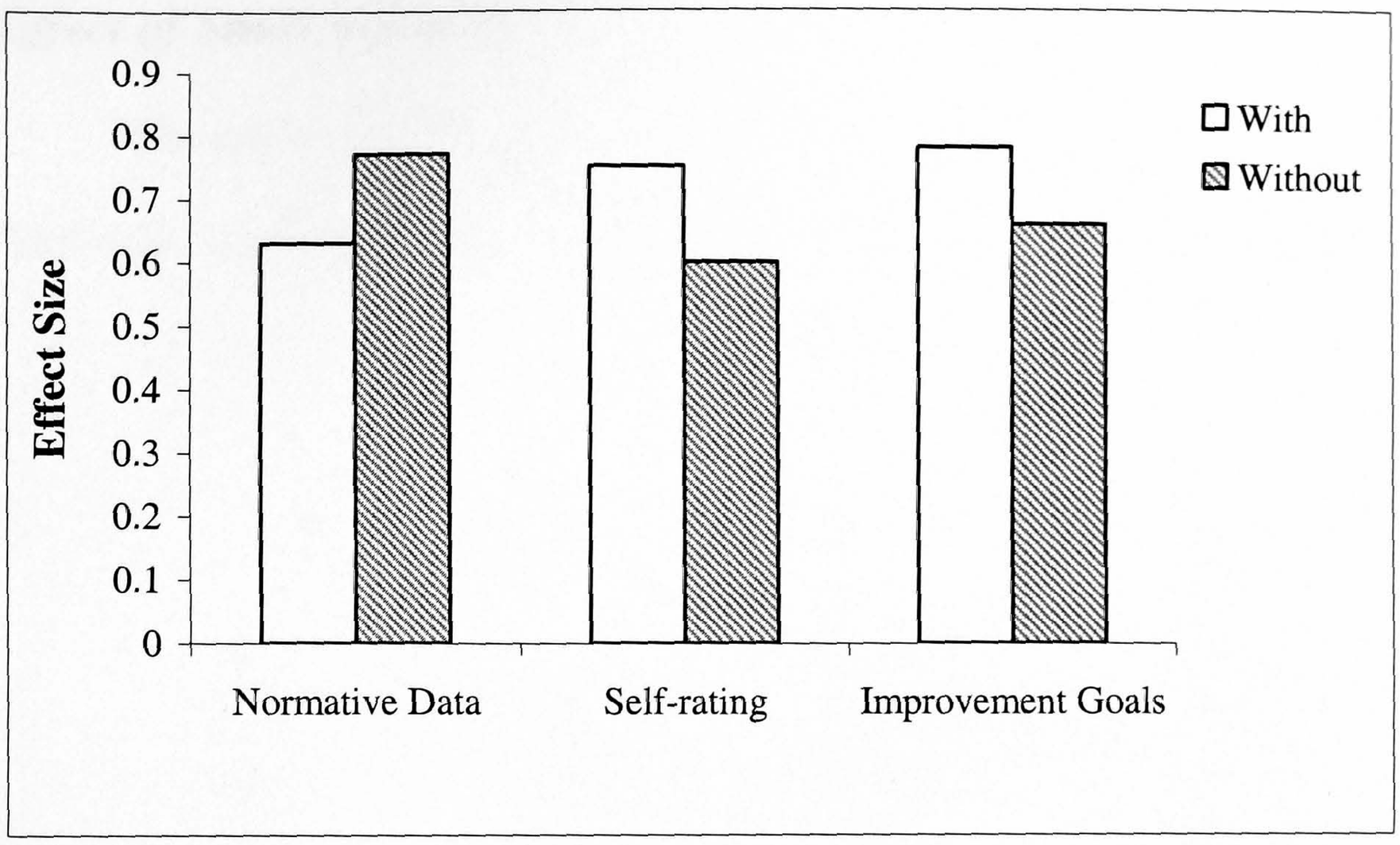


Figure 5.7 Effect Sizes as a Function of Consultation Content Variables

Self-Ratings

The effect for use of teacher self-ratings in the consultation process was only slightly larger (d_+ = .75, CI = 0.37 to 1.13) than that for consultation sessions that did not incorporate the use of teacher self-ratings (d_+ = .60, CI = 0.08 to 1.12).

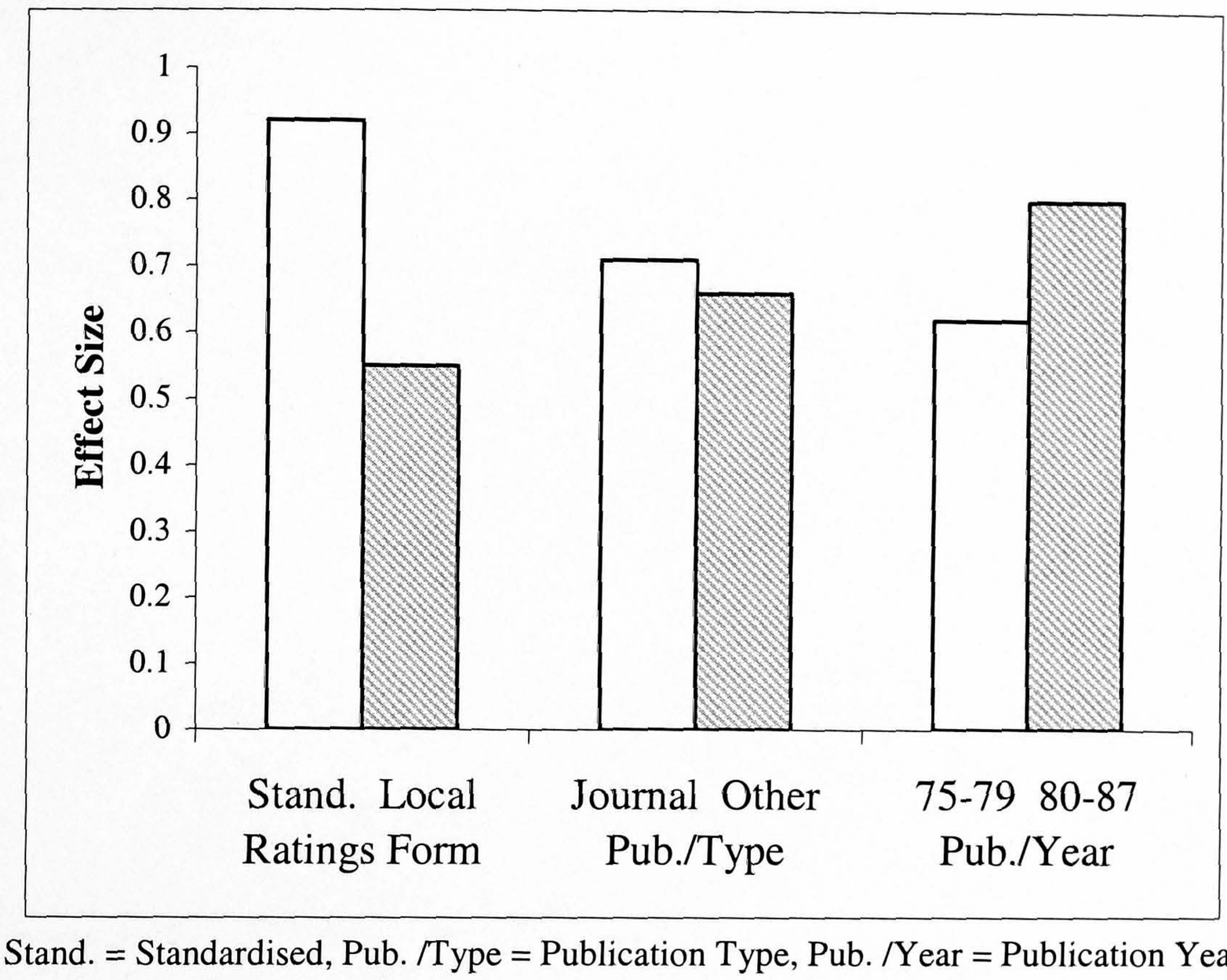
Improvement Goals

The targeting of areas for improvement and setting goals to improve in these targeted areas in the consultation process resulted in a positive moderate association between consultative feedback and teaching effectiveness

(d_+ = .78, CI = 0.10 to 1.46). Consultation sessions in which no targets were reported yielded a slightly lower effect, (d_+ = .66, CI = 0.32 to 1.00).

Effect of Study Variables

The sparse details contained in study reports only allowed for the examination of three study characteristics. Figure 5.8 summarises the results for these study variables.



Note: Stand. = Standardised, Pub. /Type = Publication Type, Pub. /Year = Publication Year

Figure 5.8 Effect Sizes as a Function of Study Variables

Ratings Forms

The ratings forms used in the studies were categorised as either standardised or locally developed following L'Hommedieu, et al. (1990) descriptions of known standardised instruments, and information from the research reports themselves. Locally developed instruments were identified as those whose content was determined by the researchers conducting the study or those that used a modified version of a standardised instrument for the research.

As can be seen from Figure 5.7 the use of standardised ratings forms significantly moderated the relationship between consultative feedback and teaching

effectiveness (d_+ = .92, CI = 0.40 to 1.44). Use of locally developed ratings forms returned a much smaller effect (d_+ = .55, CI = 0.16 to 0.93). It should be noted that no correction was made for the unreliability in the effect size estimates for local ratings forms.

Publication Type

As might be expected the mean effect size of studies published in journals (d_+ = .71, CI = 0.37 to 1.06) was higher, but only very slightly than that obtained for unpublished studies (d_+ = .66, CI = 0.01 to 1.30).

Publication Year

On average, studies published during the period 1980–1987 (d_+ = .80, CI = –0.25 to 1.35) had slightly larger effect sizes than studies published during the period 1975–1979 (d_+ = .62, CI = 0.18 to 1.06).

DISCUSSION

The use of consultative feedback as an effective teaching improvement strategy has received much empirical support. Yet there have been persistent questions regarding what works and does not work. The general objective of this review was to identify the strategies and practices that are potentially important to make consultative feedback most successful. This review represents an important contribution to the literature because in addition to providing support for the robust effects of consultation over student ratings to improve teaching effectiveness, it provides important information for academic researchers and practitioners to consider in designing interventions to support teachers in learning from student ratings feedback.

These are, however, limitations to this study. Importantly, only a small sample of studies was located to be included in the review. This means that the statistical power was very low and so estimates of effect sizes might not be reliable. In this way, the results are inconclusive and should be interpreted with care.

Overall, the results of this meta-analysis found, not surprisingly, that a positive moderate relationship exists between consultative feedback and teaching effectiveness. The mean effect size = 0.69, is comparable to the results of Cohen's (1980) earlier meta-analysis that reported mean effect size of 0.64 for augmented feedback but is smaller than the estimate (*ES* = 1.10) reported by Menges and Brinko (1986) in their update of Cohen's meta-analysis.

Although the result of this present meta-analysis has shown strong support for one-to-one consultation it also reflects the potential of group-based consultation in which teachers work together in groups supporting each other and learning together how to improve their teaching. On-to-one consultation, although important, is considered a luxury most universities cannot afford in the current climate of reduced budgets. Given the current emphasis on the need for collaboration and collegiality among teachers to improve teaching and learning this finding is welcomed.

Maximising the Effects of Consultative Feedback

First and foremost, the findings of this review make it clear that the intervention consultative feedback may be practiced in a variety of forms. In other words, there is no single approach that could be applied in every context and not all approaches will be equally effective. Although the findings appear to broadly agree with the view in the literature that there is no 'correct way' to provide consultative feedback, it disagrees with Brinko's (1997) argument that "no one kind of consultation is more effective than others" (p. 3). In common with Cohen (1980), use of meta-analytic procedures made it possible to provide a more objective and accurate conclusions about the effectiveness of different approaches to consultative feedback.

The results of this review suggest that a diagnostic or 'bare bone' type of consultation, characterised by mere analysis and interpretation of ratings with some discussion on how teachers might improve their practice may be less effective than the kind that collects information about teaching performance from other sources, and actively engages teachers in the improvement process. The results also show a clear association between the use of certain strategies and gains in teaching effectiveness.

However, some caution is required in the interpretation of the findings from this meta-analysis. Given the small number of studies available, one would not have expected to find clearly statistically significant differences among the different consultation approaches and strategies. Even where a particular strategy appears to be more effective than another, the confidence intervals overlap and it is not clear that any differences go beyond what might have been expected from sampling variation. Thus

the most robust finding may be that more research is needed. Nevertheless, some differences were found and these are suggestive, even if not conclusive.

Process Variables

To evaluate the impact of consultative feedback on teaching performance researchers adopted a variety of approaches. The search for moderator variables started with how consultation was operationally defined, and was coded as diagnostic, advisory, or educational. The results suggest that the use of either the advisory or educational approach, rather than the diagnostic approach to consultation, provided more benefits to teachers in terms of greater gains in teaching effectiveness. The general components of the advisory and educational approaches include:

- Data from multiple sources: evidence on teaching behaviours was collected with more than one source of information and the more sources used the greater the effects.
- Extended discussion: duration of consultation was on average 2 hours.
- Training: through workshops/seminar on teaching issues.

The moderator analyses indicated that teaching effectiveness improved more when the consultation process involved use of information from sources such as interviews, classroom observation, and videotaping. This finding is consistent with the general consensus in the literature that student ratings represent only one source of information, that is, student views about teaching, which by itself is insufficient to explain teaching. Here, the recommendation is that at least two additional sources of information should be used to supplement student ratings feedback to make a fair

assessment about teaching performance (Kulik, 2001; Braskamp & Ory, 1994; Seldin 1999a).

Interview

The use of interviews in the consultation process corresponds with the conference phase in the consultation cycle. The act of interviewing the teacher at the beginning of consultation is seen as a means of generating useful information for the consultant to better understand the teaching context of the individual teacher, and the teacher's experience. Perhaps more important, it provides an opportunity for teachers to reflect on their experience as they talk about their instructional practices, which may also be insightful to both the teacher and the consultant.

Further, it is during the interview that decision on the need to gather additional evidence on teaching behaviours is agreed on and the methods for the data collection negotiated. Bergquist and Phillips (1975) pointed out that without this activity, or too little time in it, too much to the left to the imagination forcing the consultant to make assumptions about the needs of the teacher, which may prove incorrect and result in a waste of time and money.

From the research reports, it seems that there is a tendency to overlook this conference phase to rely instead on the information from student ratings and other data sources that may have been employed. One reason for this is that researchers tend to standardise rather than individualise the consultation process. Above all else, teachers expect the consultation process to focus on their specific needs to improve their teaching (Stanley, Porter & Szabó, 1997).

Observation

The use of observation as a source of information on teaching is regarded as one of the core practices in teaching consultation. It is seen as an important method to offer insights on the climate, rapport, interaction, and functioning of a particular classroom on a given day. This in turn helps to provide a fuller picture of teaching behaviours and the consequences for student learning that would be discussed in the consultation session.

One concern about the use of classroom observation is that a one-shot visit may not provide a representative sample of teaching behaviours. On this basis teachers are often critical of its use for summative evaluation where it might be given undue weight. They are, however, more approving when it is used for formative evaluation. In this case, its use is likely to be less intimidating. When peers undertake classroom observation it adds another dimension to the process, as colleagues are in a good position to assess course goals, content, and organization, as well as the quality of methods and materials used in delivery (Braskamp & Ory, 1994; DeZure, 1999).

Videotape Recording

Probably a more powerful observation method is the use of videotape recording. Viewing videotape recordings on one's classroom performance can help teachers to become aware of teaching behaviours in the actual classroom setting. This offers an opportunity for self-assessment. Supporting its use Carroll (1981) said "one of the potentially most powerful forms of self-assessment is the opportunity to 'see ourselves as other see us' through video recording" (p. 193).

Following an extensive review of the literature, Fuller and Manning (1973) agreed that videotaping offers the opportunity for self-confrontation. The warning,

though, is that videotape replays have the potential to be more harmful than helpful in that it might produce fear, anxiety, and embarrassment and in turn hinder learning. The crucial element in the use of videotaping, it is recognised, is to focus on strengths and weaknesses that are within the control of the teacher to address.

In summary, the sole use of student ratings feedback as a source of information on teaching performance is simply not practical. Teaching is too much a complex activity for one source to give an accurate picture of effectiveness. The results of this meta-analysis underscore the point that using interviews, classroom observation, and videotaping to gather additional sources of information are important in assessing teaching. If used properly, they can stimulate the desire to alter teaching behaviours, and acceptance of the responsibility for teaching improvement.

Context Variables

Analyses of moderator variables indicated a tendency for context related variables—duration of consultation, use of experts vs. peers as consultant, and generic vs. discipline-specific consultation—to moderate the relationship between consultative feedback and teaching effectiveness.

Duration of Consultation

The results of this meta-analysis suggest that as the duration of consultation period increases, the effects of consultative feedback on teaching effectiveness may increase, albeit marginally. This finding corroborates research by Piccinin (1999), which showed that even a 'brief' consultation improved teaching but the effect was modest. With longer consultation sessions the effect was more dramatic in Piccinin's (1999) study.

One explanation for this finding is that the duration of the consultation process is related to it contents. There is the tendency for longer consultation sessions to be linked to discussions on information from other sources of information rather than from student ratings alone. This allows for more sustained dialogue and explorations of instructional practices. At the same time it extends the opportunity for teachers to become actively involved in the process from which a collaborative approach is likely to emerge, which sees the consultant and teacher working together. A relatively long consultation session also provides teachers with psychological space for reflection on their experience as teachers.

By contrast, a consultation session for only one hour might be limited to interpreting ratings, identifying strengths and weaknesses, and does not offer much in the way of sufficient time to facilitate reflection on experience for learning to take place. Here, the teacher might simply be a passive recipient of advice. Active engagement in the consultation process makes it more likely that learning and change will inevitably occur. There is no reason to expect, though, that there is a direct relationship between the length of the session and the quality of consultation.

The need to improve teaching is probably a greater concern in higher education today than it was when the majority of studies used in this review were conducted. So, there is justification for an extended consultation period that is educational in nature and which allows for reflection and experimentation. However, university teachers now work in a more complex environment, where there are for example, accountability and quality audits, increased academic and administrative workload, and pressure to 'publish or perish'. One cannot help but wonder therefore about the extent to which

teachers would give up two or more hours of their time to engage in an activity that is seen as voluntary and an addition to already full work schedules.

The implication of this is that as long as consultative feedback remains voluntary teachers are more likely to see engagement in the process as an additional activity, rather than an extension of their professional responsibility. Though they should, with so many activities competing for their time, it seems probable that teachers will put off seeking consulting until a major 'crisis', such as low student ratings in the face of an upcoming promotion.

Consultant

Not only is the role of the consultant a crucial factor in consultative feedback, but it seems that who serves as consultant is equally important to the effectiveness of the process. The results of this meta-analysis indicate that while teachers might benefit from consultation with either a staff developer or peer, they will generally derive slightly more benefits from being supported and assisted by their peers.

One interpretation of this finding is that teachers appreciate the opportunity to interact and draw on the knowledge and experience of colleagues recognised as having appropriate skills and experience. It could also be seen as teachers' willingness to collaborate for teaching improvement if given the appropriate learning and collaboration space.

An explanation for this finding is that academics who seek consulting are more likely to be untenured, tenure-track, assistant professors (Stanley, et al. 1997; Piccinin 1999) who recognise the role of teaching for promotion and tenure. Unfortunately, adequate demographic information on volunteers was not reported in many studies for further exploration.

At the same time, academics who serve as consultants to their peers tend to have a reputation as an outstanding teacher and have many years of teaching experience, as well as an interest in helping their colleagues. These individuals would normally receive training to use the system devised by professional staff developers and may be paired with colleagues from either the same or a different discipline.

So, while many peer consultants might not have the specialised knowledge in teaching development as perhaps staff developers do, they have the advantage of knowing the minutiae of teaching in the disciplines or have expertise from their experience. North (1999) argued that it is only peers that are suited to; assess the appropriateness of course goals, help define types of evidence for student learning, assess subject matter mastery, and make judgements on discipline-specific aspects of teaching. North (1999) added, "only colleagues can detect the scent of staleness, and only colleagues can provide the spark that can lead to revitalization" (p. 188).

Nature of Consultation

By tradition consultation is driven by the generic issues of teaching rather than by discipline-specific concerns. The findings of this meta-analysis suggest that consultation may be approached quite appropriately from either a generic or discipline-based perspective. Consultation of a generic nature, however, emerged as having a slightly stronger influence on teaching performance than did the assumed discipline-specific consultation. This finding is probably not surprising and is in line with the practices of the time period in which the majority of studies were conducted where the emphasis was on developing teaching competencies.

During the 1970s and early 1980s the thinking was that poor teaching behaviours could be corrected by prescribing, as it were, 'teaching tips' rooted in

universal and generic principles of good teaching and learning, independent of the disciplines. Levinson-Rose and Menges (1981), for example, criticised these activities for being mechanical in nature, which resulted in superficial changes, if any at all. By contrast, the emphasis in present day higher education has shifted from developing technical competency to knowledge about teaching and learning, and at the same time taking account of the differences in the disciplines.

Understandably, there are effective teaching behaviours that are common to the disciplines. Nonetheless, there is near unanimity that there are clear differences in teaching and learning in the disciplines that should not, or cannot be ignored (Jenkins, 1996; Neumann, 2001). Rowland (2000) contended that continual emphasis on the generic and technical issues of teaching serves only to reinforce the notion that teaching is an 'amateurship' activity that requires no intellectual capability.

Participants

There was also no real difference in effect sizes with the use of consultative feedback for regular academic staff or teaching assistants. For the purpose of this discussion, if it is assumed that teaching assistants represent junior staff members and the regular staff members are the more experienced teachers, then, consultative feedback generalises well across teachers of all academic levels. That means consultation is a relevant activity for the professional development of all teachers. This is important in helping to remove the stigma of 'remedial' associated with seeking consultation.

Content Variables

Normative Data

The results of moderator analyses also suggest that actually not reporting ratings feedback in relation to normative data or making reference to such data in consultation may hold slightly more benefits for improving teaching than when normative data is used. The use of norms allow a teacher's performance to be assessed and ranked against their colleagues, in comparable contexts, such as similar subjects, course levels or disciplines, to determine their relative standing.

The finding of this meta-analysis lends some support to McKeachie's (1996) observation, after 45 years of researching student ratings and disseminating norms, that providing norms data might do more harm than good in improving the quality of teaching. Norms, McKeachie explained, have a negative motivational effect because even though teachers may have received favourable ratings from students, they may find that their averages place them below the mean or median for the group they were being compared with. In this respect, teachers may experience frustration rather than a desire to improve their performance.

McKeachie (1996) further noted that norms may actually create a competitive environment, which is likely to give rise to secretiveness and unwillingness on the part of teachers to support their colleagues. This of course runs counter to the present focus of higher education to foster collegiality and collaboration. Here, it is believed that a culture of collaboration will, in turn, improve the quality of teaching and student learning.

Other researchers and practitioners share the concerns of McKeachie, but believe there is need for a basis of comparison. For example, Gillmore (1996) argued

that without normative information the results would display the Lake Woebegone effect, a situation where everyone is above average. An unintended consequence is that a teacher might not only overestimate his/her teaching competency but this self-deception holds damaging effects for the quality of student learning.

For Cashin (1996), norms actually help the teacher to interpret the ratings as student ratings tend to display negative skewness with scores clustering at the high end of the scale. Cashin (1996) also believes that with guidance on how to use the normative data teachers can monitor their own performance.

With similar views, Aleamoni (1996) proposed that norms could be created by gathering data on a longitudinal basis on different rating conditions, for example course, department, or discipline, to establish a 'standard' for each given condition. In this way, Aleamoni reasoned, teachers would be compared against this 'standard' rather than with individuals, to overcome McKeachie's concern.

Self-Ratings

The use of teacher self-ratings in the consultation process has an apparent influence on the effects of consultative feedback. The evidence from this meta-analysis suggests that incorporating teacher self-ratings in the consultation process may be beneficial in helping teachers to bring about substantive improvement than when such ratings are not considered.

One explanation for this is that, the use of teacher self-ratings will not only help teachers to diagnose their own strengths and weaknesses. They will also help teachers become aware of any discrepancies between their ratings and that of students'. In this sense, they may critically reflect on their practice and take steps to reduce that discrepancy (Seldin, 1999b). This represents a particularly strong argument if student

ratings are markedly lower than the self-ratings and attention is drawn to the discrepancy by the consultant.

Reflection from this perspective, however, is best undertaken with the assistance of a consultant. Because self-ratings can introduce dissonance and reveal weaknesses that may prove stressful, a supportive climate to facilitate change is very important. If undertaken alone reflection may be descriptive rather than critical. Or, teachers might simply become defensive to the perceived threat to their view of their own teaching effectiveness. The consultant, with the right skills, can help teachers to reflect at a deep level that leads the individual to question their assumptions about teaching and learning, opening the way for change and improvement.

Use of teacher self-ratings in the different research reports required teachers to make judgements about their teaching by completing the same ratings form that was completed by their students but independent of them. Teacher self-ratings are therefore most useful when used in conjunction with at least one other data source. At best, if there is any discrepancy between the self-ratings and student feedback, for example, this will create a state of dissonance to the extent that the discrepant areas will be targeted and necessary steps taken to improve teaching.

Despite the potential benefits of teacher self-ratings in stimulating a desire to improve and setting goals to deal with any perceived discrepancy between the actual and desirable performance, improvement will not necessarily follow. This is because, performance, to a large extent, is dependent on the individual teacher's view of his/her ability to successfully create the conditions that will bring about improvement.

In other words, teacher self-efficacy, which has powerful implications for improving teaching is an important determinant of performance, probably more than

any dissonance that might result from the use of self-ratings. This is consistent with Bandura's (1982, 1993) self-efficacy theory that relates to one's self-confidence regarding performing a specific task. In the context of student ratings, Roche and Marsh (2000) used the notion of teacher self-concept.

In a unique study, Roche and Marsh (2000) found evidence for the influence of student ratings on teachers' self-concept. Teachers, the researchers found, adjust their self-perception upwards or downwards when they receive their student ratings feedback. Teacher self-concept, the perception teachers hold about their teaching effectiveness, therefore has powerful implications for improving teaching, motivation to teach, engagement in teaching improvement activities, and satisfaction with one's teaching efforts. This is an issue that consultants may now need to take account of in the consultation process.

Improvement Goals

Motivation to improve teaching is one of four preconditions Centra (1993) indicated must be met if the use of student ratings feedback is to lead to teaching improvement. One source of this motivation is setting improvement targets and the desire to achieve those goals. The results of this meta-analysis suggest that although it is possible for teaching performance to benefit from consultative feedback without setting improvement targets, the effects of consultative feedback is likely to be somewhat stronger when targets are established.

This is consistent with goal setting theory (Locke & Latham, 1984, 1990) which elaborates that goals enhance performance through their direct effect on an individual's thought and actions thereby increasing motivation. The theory assumes that goals are a

major determinant of behaviour regulation. Goals help to focus a person's attention on goal related factors, which may lead to performance improvement.

The goal setting theory also explains that goals can improve performance on a task if they are specific, proximal, accepted, and adequately challenging (Locke & Latham, 1990). The importance of these elements is demonstrated by Marsh and Roche (1993) who devised a system whereby study participants targeted for improvement two or three components of effective teaching for which they received low student ratings but had themselves rated as very important on their self-ratings form.

The researchers found that subsequent ratings on these targeted components increased substantially relative to non-targeted areas. The long-term implication of this goal setting strategy is that as the teacher is rewarded with higher ratings, then this may lead to the selection of additional areas to target, ultimately improving overall teaching effectiveness.

Yet, even when clear and specific goals are established teachers might still not know how to improve their teaching. Guided support from a consultant is still important not only in helping teachers develop their skills in targeting areas for improvement but also in developing and implementing strategies to achieve their goals. Here, the provision of resource materials on teaching ideas should be seen as an important part of the process (Wilson, 1986; Marsh & Roche, 1993; Hampton, 2001).

Study Variables

Ratings Forms

A variety of ratings forms were used in the studies reviewed for this meta-analysis. Forms were either standardised or 'home made', multidimensional with many

items or contained only a few global type items. The findings of this review suggest that it is the use of standardised student ratings questionnaires that can be expected to have greater effects on increasing teaching effectiveness than the use of non-standardised instruments.

The explanation for this finding is that standardised instruments are more valid and reliable as the items and subscales of such forms would have been supported by theory, the judgment of experts, and subjected to rigorous psychometric procedures to ensure that they measure what they purport to, teaching effectiveness. Reliability that examines the agreement among students is usually quite high for these instruments. For validity, the extent to which ratings measure what they are intended to measure, these well-developed forms are moderately correlated with the important criterion of student learning (Centra, 1993; Koon & Murray, 1995; Marsh & Dunkin, 1992).

By contrast, 'home made' student ratings questionnaires, non-standardised forms, might simply contain a mere list of items that not only makes it difficult to interpret what is being measured but also leaves teachers questioning the validity of student ratings. The use of poorly developed instruments affects the quality of information available to teachers for teaching improvement.

Publication Type and Year

Effect sizes did not differ much for studies that were published in peer-reviewed journals and those that were unpublished, that is, dissertations and conference reports. Published studies showed just a slightly larger degree of association between consultative feedback and teaching effectiveness.

From the moderator analyses it would also appear that the consultative feedback intervention used in the period 1980 – 1987 may have been slightly more effective in

influencing improvements in teaching than that used in the period 1975-1979. Examination of the approach to consultation adopted in these two time periods and study features revealed no clear differences in the characteristics of the intervention.

Beyond study characteristics, however, is the context of the studies. During the period 1975-1979, the time when approximately half of the studies were conducted there was much debate on the usefulness of student ratings to improving teaching. As mentioned earlier, this resulted from inconsistent and inconclusive findings from individual studies and narrative reviews of the literature. This debate was settled by Cohen's (1980) meta-analysis.

It could be assumed that teachers' attitude towards the value of student ratings might have been more favourable on the results of that meta-analysis, so that teachers were probably more willing to actually attend to the data, with the help of a consultant, to inform changes in teaching behaviours. The implication here is that, acceptance of the value of student views, as a source of important information on teaching strengths and weaknesses, is important to consider changes in teaching behaviours from ratings feedback in the first place.

Teachers should therefore be given the opportunity to re-examine their assumptions and beliefs concerning teaching and learning in the consultation process. Otherwise consultation could be seen as a way to obtain higher student ratings rather than for teaching improvement of the sort Meeth (2000b) described as, "the distance a teacher moves in understanding how to make his/her teaching more appropriate, meaningful, and effective" (p. 350).

With the notion that teaching behaviours are underpinned by the teacher's conception of teaching and learning (Kember & Gow, 1994; Trigwell & Prosser, 1996),

a conceptual change approach is now the preferred paradigm. It is believed teaching improvement activities will have a more lasting effect. Ho, et al. (2001) demonstrated use of this approach through a short-course staff development programme. The researchers reported that the results were "encouraging" to suggest that changes in teachers' conceptions can bring about positive changes in teaching practices. Assumptions about teaching and attitudes towards the value of student ratings should now be checked in the consultation process.

Implications for Practice

The results of this meta-analysis suggest that use of consultative feedback to improve university teaching can be practised in a variety of forms. However, with the use of certain strategies, consultative feedback can be expected to lead to large and positive effects on teaching effectiveness. It is obvious, then, that the essential components of consultative feedback is not just in terms of 'talking' over student ratings and devising improvement strategies with the assistance of a teaching consultant.

From the findings of this meta-analysis nine practices and strategies have been identified as being associated with effective consultation:

1. active involvement of teachers in the learning process
2. use of multiple sources of information
3. sufficient time for dialogue and interaction in the process
4. interaction with peers
5. limited use of normative data
6. use of teacher self-ratings
7. setting improvement goals
8. use of standardised student ratings questionnaires
9. adoption of a conceptual change approach

It is important to note that while the evidence for these strategies may not be very strong, it represents the best knowledge to date. Also, it should be recognised that however beneficial these practices and strategies might seem to be, the impact of their use is greatly dependent on the quality of the interaction between the consultant and the teacher.

This meta-analysis therefore provides evidence that may warrant a shift in focus from the consultant as *the* factor in consultative feedback to placing emphasis on the content, strategies and practices that can maximise the benefits of consultative feedback for teachers. The implications for practice presented here incorporated the results of the meta-analysis, other literature, and theoretical arguments to provide the best available knowledge to help guide policy and practice. It is therefore suggested that student ratings feedback may be effectively used to improve teaching effectiveness when it is structured within a consultative feedback framework that has, in addition to support from a consultant, the following features:

1. *Use of Multiple Sources of Information*

The benefits of consultative feedback may be expected to improve when feedback on teaching performance is examined from multiple perspectives as provided by the information collected from other sources in addition to student ratings. Full use should be made of interviews with the teacher, classroom observation, videotape recording, and teacher self-ratings. This should provide a more accurate picture of teaching behaviours to both the teacher and consultant. In this way, the information can facilitate goal setting and the development of appropriate improvement strategies, tailored to the specific needs of the individual teacher.

2. *Opportunities for Peer Interaction*

The staff developer and one-to-one consultation are still important features of a teaching consultation programme. However, successful consultation may result when teachers have the opportunity to interact with and draw on the knowledge and experiences of their colleagues in learning from student ratings feedback and how to be more effective as a teacher. One approach is to use appropriately trained staff members as consultants.

Another approach is to provide opportunities for teachers to work with their peers in groups from their disciplinary perspectives. Not only is the use of peers a cost-effective and time-efficient alternative to professional staff developers, it facilitates the development of a collaborative learning culture in which there is sharing and openness about teaching, which is closely aligned with the present reform efforts in higher education.

3. *Provision of Adequate Learning Time and Space*

The provision of sufficient time and space to learn from student ratings will provide an opportunity for teachers to reflect on their teaching experience in a substantive way, thereby increasing the effects of consultation. To reflect on one's teaching implies that sufficient time is available for individuals to engage in purposeful dialogue to encourage learning from experience as captured by information from student ratings and other sources of data.

Learning from student feedback is not relatively immediate. It requires sufficient time and space to analyse the feedback from the various sources, examine assumptions, rethink practice, and plan for improvements. With sufficient time and space the result is likely to be that teachers are more willing to engage in experimentation and development of improvement strategies. In effect, what this means is that consultation may have to be seen as an on-going process rather than a one-shot exercise.

4. *Opportunity for Self-evaluation*

The opportunity for teachers to engage in self-evaluation through completion of the same ratings instrument as students can provide enormous benefits for teachers and teaching improvement. The use of self-ratings allows teachers to assess teaching strengths and weaknesses, and overall performance as perceived by their students. Teacher self-ratings, especially when coupled with reflection, form an integral component of learning from student ratings.

In fact, consultation offers an ideal condition for the comparison of self-ratings and ratings from students as the process is non-evaluative and supportive in helping teachers to attend to any discrepancy between their desired performance and

performance as perceived by students. The use of teacher self-ratings therefore has the potential to lead to meaningful change and long-term improvement.

5. *Provision of Meaningful Feedback Information*

To be most successful, consultative feedback should be based on timely, accurate, and relevant feedback information that truly reflects a teacher's performance as perceived by students, to begin with. This calls for the use of well-developed ratings questionnaires that are suited to the particular purpose and needs of the context. Ratings feedback from standardised ratings forms provides information that is more likely to be viewed as more credible, thereby increasing use of the data.

Where the use of psychometrically sound ratings forms might not already exist, practitioners should pressure university administrators to commit the resources to allow for the validation of ratings forms. At the same time, the feedback should preferably be diagnostic in nature to provide good leads to the particular areas that might need attention, and perhaps more importantly, the information should be used solely for formative evaluation purposes. Just as important is the need for detailed and user-friendly reports that can be easily interpreted and which provide enough information to assess and facilitate teaching improvement.

6. *Examination of Conception of Teaching*

Improved teaching depends on developing an understanding of one's role in teaching. The consultation process can be effectively used to help teachers to understand and shape their conceptions of teaching and learning in higher education. The opportunity for teachers to articulate their values and beliefs about teaching, where it could be examined and challenged should provide substantial benefits from the

consultation process. This might be so as teachers may develop an awareness of themselves as teachers and their taken for granted assumptions. Reflection on teaching performance and taking steps towards improvement without the opportunity to subject beliefs and values to change, lead only to superficial learning, which is certainly not sufficient for long-term teaching improvement to follow it might be argued.

7. *Establishment of Improvement Goals*

The most effective consultation involves targeting certain teaching areas and establishing improvement goals. The information learned about one's teaching should be used for goal setting and the targeting of specific areas for teaching improvement. To begin with, the recommendation is for teachers to select two or three areas of teaching that will have an immediate impact on student learning, and with new goals being set as existing goals are achieved. It is these goals that will provide a sense of purpose and direction to improvement efforts, and which in turn act as standards against which to measure performance improvement.

Progress toward achieving these goals would be monitored by the feedback from subsequent student ratings. Success in these targeted areas is likely to see more areas being targeted for improvement. Equally important, is the need for teachers to have ownership of the goal setting process to increase motivation and commitment. These goals as agreed on, are best documented to serve as reference for follow-up on progress and performance.

In summary, achieving the most benefits from consultative feedback is important to its selection and use as a teaching improvement strategy. To this end the consultation process should incorporate many of these strategies and practices identified in this review. This is may be a necessary condition to maximise the effects of

consultative feedback on improving teaching effectiveness. The ultimate aim, of course, being to improve the quality of students' learning experience.

Directions for Future Research

The results of this meta-analytic review provide a number of leads to guide future consultative feedback research. First, inadequate reporting of information in the research reports made it almost impossible to examine, for example, the relationship between demographic variables, such as gender, teaching experience or academic status, and the effects of consultative feedback. For instance, even though there was some evidence of gender differences, it was not possible to explore this in any detail. Future research need to pay attention to the reporting of data and statistical information for the purpose of increasing utility of findings and facilitating advancement in the field.

Second, considerably more research on the effects of consultative feedback in settings other than North America is sorely needed. Only one study found was conducted outside North America. This makes it difficult to generalise about the effectiveness of consultation to other contexts.

Third, although this review has uncovered some strategies that might be important for consultative feedback there is need for research that directly assesses the effects of these strategies to clarify the factors that really promote higher levels of teaching improvement in consultation over student ratings. Fourth, the sample of studies found for this meta-analysis was too small to provide adequate statistical power to detect differences efficiently.

Finally, the use of individual, one-to-one, consultation may no longer be feasible nor in line with the current educational reforms that urge teacher collaboration. Instead,

peer consultation and group-based consultation may be an effective alternative but this area has not received much attention from researchers. More research into the use of peers as consultants and group-based consultation is greatly needed.

In conclusion, this meta-analysis provides support for evidence of the benefits of consultation over student ratings for improving the quality of university teaching. At the same time it represents an important contribution to the literature on what may be expected to work for the most successful consultative feedback. The findings might not be a revolutionary panacea for consultation over student ratings, or even resolve all questions concerning the most effective way of providing consultation over student ratings for that matter. Yet it provides the best available knowledge that may be quite useful for practitioners to consider in planning, designing, and implementing consultative feedback interventions.

Chapter 6

Methodology

This chapter gives a detailed description of the processes used to investigate the effects of feedback and group-based peer consultation on teaching effectiveness. The chapter opens with reasons for use of a mixed method research design, a brief discussion on the use of randomised controlled trials in educational research, and details on the context of study. This is followed by information on data collection procedures and how the intervention was actually administered. The chapter closes with a discussion on the statistical procedures used in the data analysis.

Introduction

The present investigation into the use of student ratings adopted a multimethod research design, commonly referred to as mixed methods (Tashakkori & Teddlie, 1998). In this way, a combination of quantitative and qualitative research methods was used to examine the effects of augmenting student ratings feedback with consultation in peer

support groups on teaching effectiveness. A randomised controlled trial, true experimental design, was the dominant quantitative approach with a less dominant qualitative interview. Qualitative interviews were used to gain insights into how participants experienced different aspects of the feedback intervention and to determine their perception of the efficacy of consulting over ratings feedback in peer support groups, the group-based peer consultation model.

The mixing of research methods in a single study is termed methodological triangulation (Denzin, 1978). Denzin identified three basic types of triangulation in collecting research information: (a) using multiple methods, (b) using multiple data sources, and (c) using more than one investigator in the research process. The advantage of triangulation is that it adds rigour, breadth, richness, insight, and depth to an inquiry (Denzin & Lincoln, 2000). It also seeks convergence, corroboration, and correspondence of results across the different methods (Greene, 2000). It is this convergence between qualitative and quantitative research, Fitz-Gibbon (1985) contended, that makes us hopeful that social science is proving cumulative and informative. Not that convergence will always happen. Further, as Greene (2000) pointed out, the acceptance of difference in the methods is the starting point for more meaningful social science.

A complete evaluation requires both quantitative and qualitative data. Each supports and is complementary to the other (Fitz-Gibbon & Morris, 1987). For the most part, ratings feedback studies have tended to use a true experimental design to show that under certain conditions ratings feedback can be effectively used to improve teaching. There is, however, less illumination on how teachers responded to the intervention, the

components that worked and did not work well, teachers' views of applicability in practice, and general satisfaction with the intervention to meet their needs.

Understandably, claims of satisfaction with an intervention are less trustworthy than evidence from an experiment. But assessing participants' experience with an intervention does provide vital information to further develop the intervention. At the same time, getting their views on its merit and worth is crucial to its acceptance and adoption to achieve desired outcomes. Presently, there are calls for more qualitative information on how teachers use ratings feedback and the impact of interventions on their practice (Theall & Franklin, 2000; Menges & Austin, 2001).

By convention also, ratings feedback studies with a true experiment design use simple randomisation, where teachers are independently sampled and randomly assigned to a feedback or control group. The present investigation departs from this standard to use a cluster randomisation design where groups of teachers are randomised to treatment groups. This sampling design resulted in a clustered or nested structure with teacher participants nested within groups, defined by department units, in turn nested within universities. This nested structure is graphically illustrated in Figure 6.1.

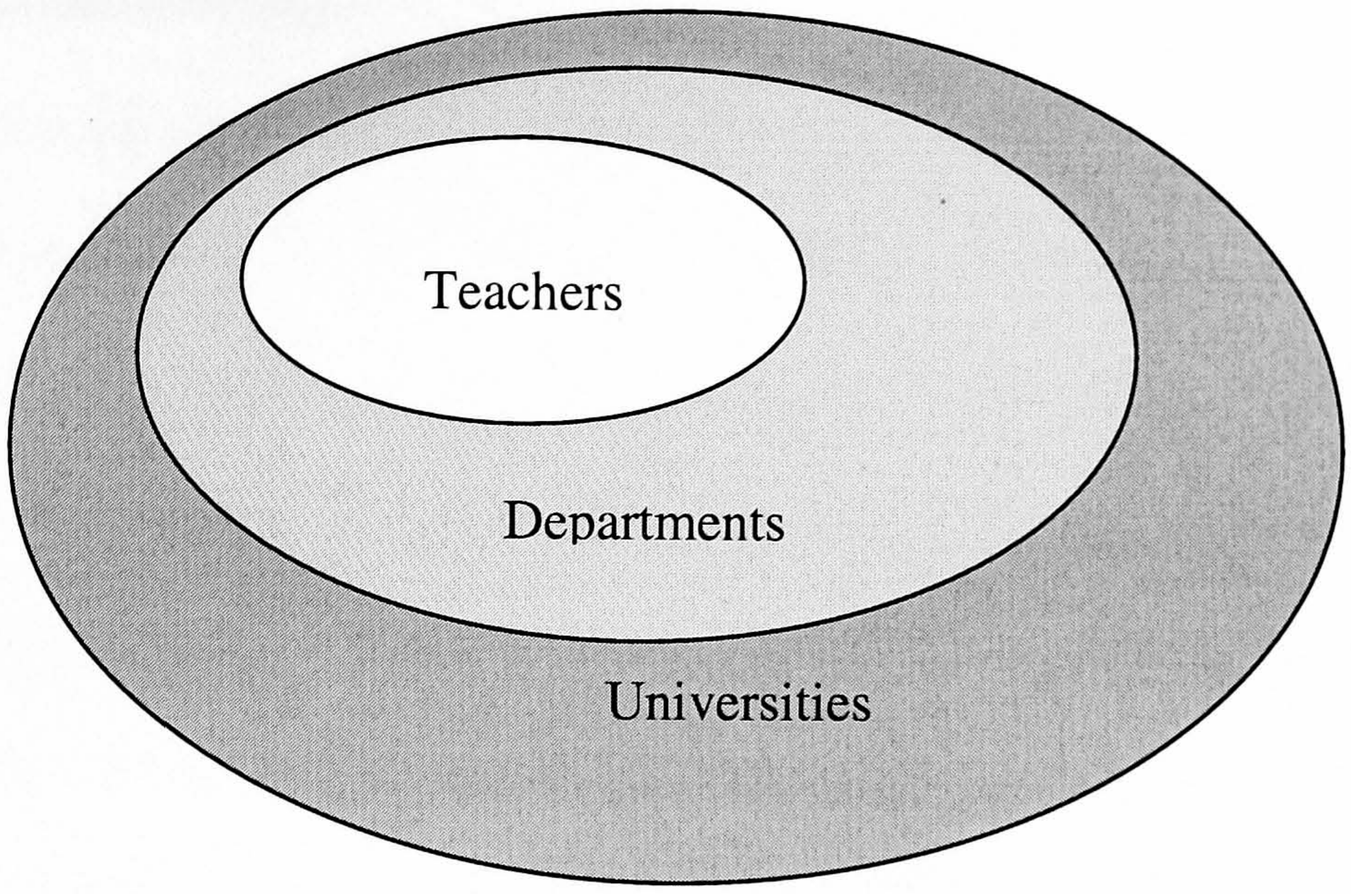


Figure 6.1 Model of Nested Sample Design for this Study

Randomised Controlled Trials in Education Research

The calls are now more frequent for more experiments, namely randomised controlled trials to evaluate educational programmes. Alongside the debate about the quality and relevance of education research is the recognition that experiments represent a primary vehicle through which to provide evidence of what works and what does not work in order to improve the quality of education (Davies, 1999; D. Hargreaves, 2000; Hammersley, 1997). The following quotation from the seminal work of Campbell and Stanley (1963) seems to capture the essence of the value of experiments in education research:

> Experiments …[are] the only way of settling disputes regarding educational practice … the only way of verifying educational improvements, and … the only way of establishing cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favo[u]r of inferior novelties (p. 2).

Randomised controlled trials (RCTs) are widely accepted as providing the 'gold standard' of evidence in research. In simple terms, a randomised controlled trial is an experimental design in which participants have been randomly assigned to treatment groups, that is, one group receives the intervention while the other does not and serve as controls. Thus, allocation to treatment is regarded as left purely to chance.

Design-wise, RCTs are seen as the best way of eliminating selection bias and offers the best safeguard to minimise bias from other factors that could easily distort the outcome. In other words, a RCT offers the most rigorous way of determining whether observed differences in outcomes can be assumed to be due to the intervention and not from extraneous variables, validating a conclusion of causal relationship (Cook, 2002; Tymms, 1999; Boruch, et al. 2002).

Although some of the earliest examples of experiments may be found in educational research (Oakley, 1998), they are now more firmly established in medicine than in education. The context of medicine and education may well be different in some respects but it appears that educational researchers are not very willing to conduct experiments in schools. The argument, similar to that of Kember (2003), is that the complexity of the education system makes it almost impossible to design and conduct effective experiments. In the same way, Cook (2002) went to great lengths to show educational evaluators that their reasons for not doing experiments were not fully justified.

Educational researchers, it seems, are more content to rely on methods such as case control studies that involve comparisons of matched schools and classrooms, correlation studies that look for relationships, and quasi-experimental methods (Torgerson & Torgerson, 2001). This is not to suggest that these research methods are

of little importance and we should rely only on RCTs. Despite their worth and usefulness, these other methods are not able to evaluate the effectiveness of programmes and provide the evidence required to influence decisions about practice and policy.

Yet, it would be misleading to suggest that all RCTs are high quality studies with reliable results. There are the problems of poorly designed and conducted experiments, which can produce misleading results and many are small-scale so findings may not be reliable and generalisable. Also, they are appropriate to provide answers to only certain types of questions, there are ethical concerns associated with them, and the results may be over relied on against the better judgement of professional knowledge and experience (Cook, 2002; Morrison, 2001; Humes & Bryce, 2001).

Nonetheless, randomised controlled trials are justified when there is uncertainty about the effectiveness of different interventions and decisions hold consequences for millions of people in light of limited resources. Besides, they generate cumulative evidence as results from individual studies may be synthesised through meta-analysis to provide evidence as to the effectiveness of a particular intervention (Fitz-Gibbon, 1996; Tymms, 1999).

It is the accumulation of evidence from randomised controlled trials that will help to promote evidence-based practice and policy. Educational practice, it is observed, has remained for too long a matter of political ideology, personal preferences, conventional wisdom, and wishful thinking:

> The ease with which politicians, policy-makers—and even teachers—have been able to get away with implementing their prejudices without even token consideration of the evidence…is a disgrace (Curriculum Evaluation and Management Centre, 2000, p. 1).

Context of the Study

'Quality' is now in Jamaica. As with many other countries, Jamaica, a small developing country, is involved in restructuring and reforming its tertiary education system following the fashion of emphasising quality and using the jargon; quality student learning, quality in teaching, quality monitoring, quality assurance, quality improvement, and quality enhancement. For that reason, local universities have started to establish formal systems of quality assurance with increasing emphasis on efficiency and in response to the overarching concerns of stakeholders—government, employers, and students in particular—about the quality of teaching and accountability for public money.

Added to this is the political agenda to increase access to tertiary education where less than 20% of the 18-24 cohort is enrolled, compared to 67% in the United Kingdom (UNESCO, 2002). There is, however, growing awareness in the Caribbean region—the English speaking Caribbean Community (CARICOM) states—of the central role of higher education and the implications for economic and social development of lagging behind the more developed countries in terms of access to and quality in higher education:

> The imperative is not only to enrol, but … to realise an improved quality of output closely aligned to current and anticipated need and the realities of the society. Quality assurance mechanisms … will be critical to this development (Caribbean Community Secretariat, 1997).

This is echoed by UNESCO (1998):

> Without adequate higher education and research institutions providing a critical mass of skilled and educated people, … developing countries … cannot reduce the gap separating them from the industrially developed ones.

Yet, the longstanding problem for developing countries like Jamaica, is the severe shortage of financial resources to spend on developing university teaching in the face of increasing student numbers, overcrowding, rising costs, deteriorating infrastructure, rapidly changing technology, and substantial brain drain.

Up to 1994, Jamaica had only the Mona Campus of the regional University of the West Indies, which began as a college of the University of London in 1948, serving its populace. This position changed in 1995 when the University of Technology, Jamaica, formerly College of Arts, Science and Technology, modelled on the former British polytechnic system, achieved university status and became Jamaica's first national university. In 1999 a private institution modelled on the American system, Northern Caribbean University, formerly West Indies College, was also granted university status. Presently, these local institutions have to keep a keen eye on the competition from American universities operating offshore centres in Jamaica and who often rely on local academics for part-time staff.

For the three local institutions, efforts towards assuring quality and improving internal efficiency have resulted in the borrowing of quality models from North America and the United Kingdom to ensure that standards are comparable to those applied in the developed world. So too, much emphasis is placed on the use of student ratings, espousing the same rhetoric of needing to improve the quality of teaching and learning in higher education. But unlike many universities in these regions, there is little support through staff development opportunities to assist teachers to learn how best to use the ratings feedback to improve their teaching effectiveness. Given the already inadequate resources for teacher development due to financial constraints in Jamaica, it seems highly unlikely that institutions will be able to fund such activities.

The difficult position for Jamaica is also reflected in terms of low remuneration for university teachers, heavy teaching loads, and limited involvement and funding for educational research. To supplement their income, for example, many teachers tend to take on additional hours for overtime, accept part-time contracts in other institutions, or get involved in outside consultancies. This is in addition to pressures to improve teaching performance and accept non-academic responsibilities within their own institutions. As can be expected, under these circumstances, teachers find themselves with full work schedules that leave many with little discretionary time to engage in professional development activities that might be required let alone are voluntary, although this might be important to improving the quality of teaching.

At this point, it is important to mention two developments during the semester in which the present research was carried out, with implications for the ecological validity of the study. As it happened, classes were interrupted for approximately one week in the third week of the semester due to torrential rains associated with two tropical storms, Isidore and Lili, September 2002. Classes were once again interrupted from the second week of October, 2002, when a general election was announced to be held in the same month. Sadly, Jamaica's elections are associated with violent political disturbances before and after the elections.

These events affected the attendance of both students and teachers. This then placed teachers under pressure to 'catch up' and cover course outlines to enable students to be able to write final examinations. Final examinations question papers are usually set by at least the third week in the semester to facilitate external moderation. For sure, these events may attenuate the effects of the intervention and at the same time bring into question the internal validity and generalisability of findings.

STUDY DESIGN

Sample

Following approval from the ethics committee for research involving human subjects at the University of Durham, England, the present investigation was conducted among teachers in two universities in Jamaica. Entry to the universities was gained following a letter to the vice-presidents in both institutions (see Appendix A). The letter outlined the purpose of the study, the nature of the intervention, the benefits to individual teachers, and the benefits to the institution. The original intent was to conduct the study in only one university. The decision to include a second institution was motivated by interest in the potential usefulness of the project from individuals outside these institutions and who thought it would be interesting to include another institution.

The proposal for the study along with a copy of the questionnaire was also submitted to the respective research and ethics committees in the two universities. In both cases approval was granted on the condition that the institution would not be identified in the report. With all three universities on the island characteristically different in terms of size, location, and structure specific description of the study institutions would easily identify the universities. For reasons of confidentiality and anonymity of findings detailed descriptions cannot be reported here and the institutions are denoted as University A and University B.

The sample for the study comprised full-time university teachers who were voluntary participants. In University A, teachers were recruited through a series of lunch hour discussion sessions on the usefulness of student ratings in the different

departments, a letter inviting teachers to volunteer (Appendix B), and personal invitations through visits by the researcher to the different departments. In University B participants were recruited through a presentation by the researcher in a staff meeting and personal appeal from the vice-president. No invitation to participate was made beyond this presentation due to time constraint. The research project itself was promoted as a means of using feedback from student ratings for professional development.

Teacher participants were told in writing and in the various presentations that mid-semester and end-of-semester ratings would be collected from their students, and that they would be asked to evaluate their own teaching effectiveness. All prospective participants were informed that they would be randomly assigned to one of two groups in the trial. Participants in the intervention group would receive mid-semester ratings feedback results and the opportunity to discuss these ratings with colleagues in a group setting, with mutual support in learning from the ratings and making changes to improve teaching. Participants in the control group would receive no ratings feedback until the end of the study.

On every occasion the aims and objectives, the procedures, the requirements of teachers, and benefits of the study were clearly outlined. These approaches resulted in a final sample of 79 teachers volunteering to participate in the study and represented a variety of disciplinary areas in the two universities. The volunteers were asked to nominate the course and class from which the student ratings would be collected. Volunteers were not offered incentives or given release time from their institutions to participate in the study.

All participating teachers were assured of confidentiality and anonymity and advised of the voluntary nature of their participation. Teacher participants were also assured that the data collected would not be shared with their institutions but used only for research purposes. These teacher participants were also asked to sign a consent form (Appendix C) that repeated the aims, objectives, procedures, and assurances as outlined in the letter.

Ratings Collection

The research was carried out during the Christmas semester of the 2002-2003 academic year that runs for 16 weeks from September to December.

Mid-semester Ratings

Mid-semester ratings, or pre-feedback ratings, were collected from students in the different classes during the sixth week for University A and the seventh week for University B, of the semester during regular class periods. The original plan was to collect the ratings in the fifth week of the semester, that is, after four weeks of instruction, in line with the common practice in ratings feedback research. In the case of this present study the collection of mid-semester ratings was delayed because of the interruption caused by the rain as previously mentioned.

Differences in the collection time between the two universities arose because of differences in the time when the two institutions were approached and volunteers secured. A packet was prepared for each class from which ratings were to be collected. The packets contained the ratings forms and an information sheet (Appendix D) with guidelines for administration of the forms. Simultaneously, teachers were asked to complete an identical ratings form for their self-evaluation.

With the logistical constraints of a single researcher and limited financial resources to hire a full-time assistant, classes being dispersed across the campuses, and similar class times for the courses nominated by teachers it was not possible to maintain a standard procedure to collect the ratings. This meant that both the researcher and teachers administered the ratings forms. While the option would have been to allow teachers to collect the ratings for uniformity, some teachers expressed the desire not to become involved in this way.

Other teachers who were involved in the collection of ratings were given written guidelines on the administration procedures and were encouraged to nominate a student to administer the forms and leave the room. Research has shown that the manner in which ratings are collected may have an effect on the ratings, albeit a negligible one (Feldman, 1979). Even so, the recommendation is for standardised procedures in the administration of student ratings forms.

End of Semester Ratings

End of semester ratings, or post feedback ratings, were collected during the last two weeks of scheduled classes in the semester, prior to final examinations using similar procedures to that for collecting mid-semester ratings. As is the common practice in ratings feedback research, these ratings were used to assess the immediate effects of the intervention. What was problematical in attempting to collect end of semester ratings was that the institutions also started to administer their own ratings forms. This resulted in students in one institution complaining of 'questionnaire fatigue' with some students refusing to complete the ratings forms.

Further, some teachers felt that in light of the events of the semester they needed "every minute" of their class time to complete course outlines so they were unable to

allocate time for the administration of the follow-up questionnaire. Meanwhile some teachers simply forgot to administer the forms. The influences of these factors meant that complete sets of mid-semester and end of semester ratings were secured for only 71 of 79 teachers. Table 6.1 gives information on the distribution of the sample and attrition rate.

Table 6.1
Teacher Participation by University and Intervention Group

University	Group			
	Control		Feedback	
	Mid-semester	End of semester	Mid-semester	End of semester
University A	12	11	12	8
University B	27	26	28	26
Total	39	37	40	34

One consequence of attrition is that teachers for whom no follow-up ratings were collected might be very different from those for whom both sets of ratings were secured. This introduces bias and renders the sample less representative of the target population. Probably more damaging is the differential loss between the treatment groups that threatens internal validity (Gall, Borg, & Gall, 1996). Preliminary analysis found no difference in the ratings of teachers with complete ratings and those with ratings for only one portion of the semester for any of the nine scales on the questionnaire or overall ratings. It can therefore be concluded that teachers for whom no end-of-semester ratings were collected are very similar to those with both mid- and end-of-semester ratings in terms of the ratings given by students.

Identification of Peer Support Groups

Traditionally, consultation as a feedback intervention is based on the one-to-one model which sees the individual teacher working privately with a teaching consultant—usually a staff developer or trained peer—to interpret the ratings feedback and develop improvement strategies. Given the objective of the present study, consultation was offered as a group process in that the intervention was administered to teachers in peer support groups, defined by department units in the two universities. From this emerged hierarchically structured student ratings data with three levels, as illustrated in Figure 6.2.

The group-based peer consultation model therefore involved teachers in dialogue and discussion with their peers in their own working space to solve teaching problems as identified from the ratings feedback. That means, that the groups were discipline-based and the trial closely mirrored the real working context, as opposed to being contrived for the experiment. What's more, this approach would help to minimise the problem of contamination from participants receiving different aspects of the treatment sharing the same workspace. All participants in the same department unit therefore received the same treatment.

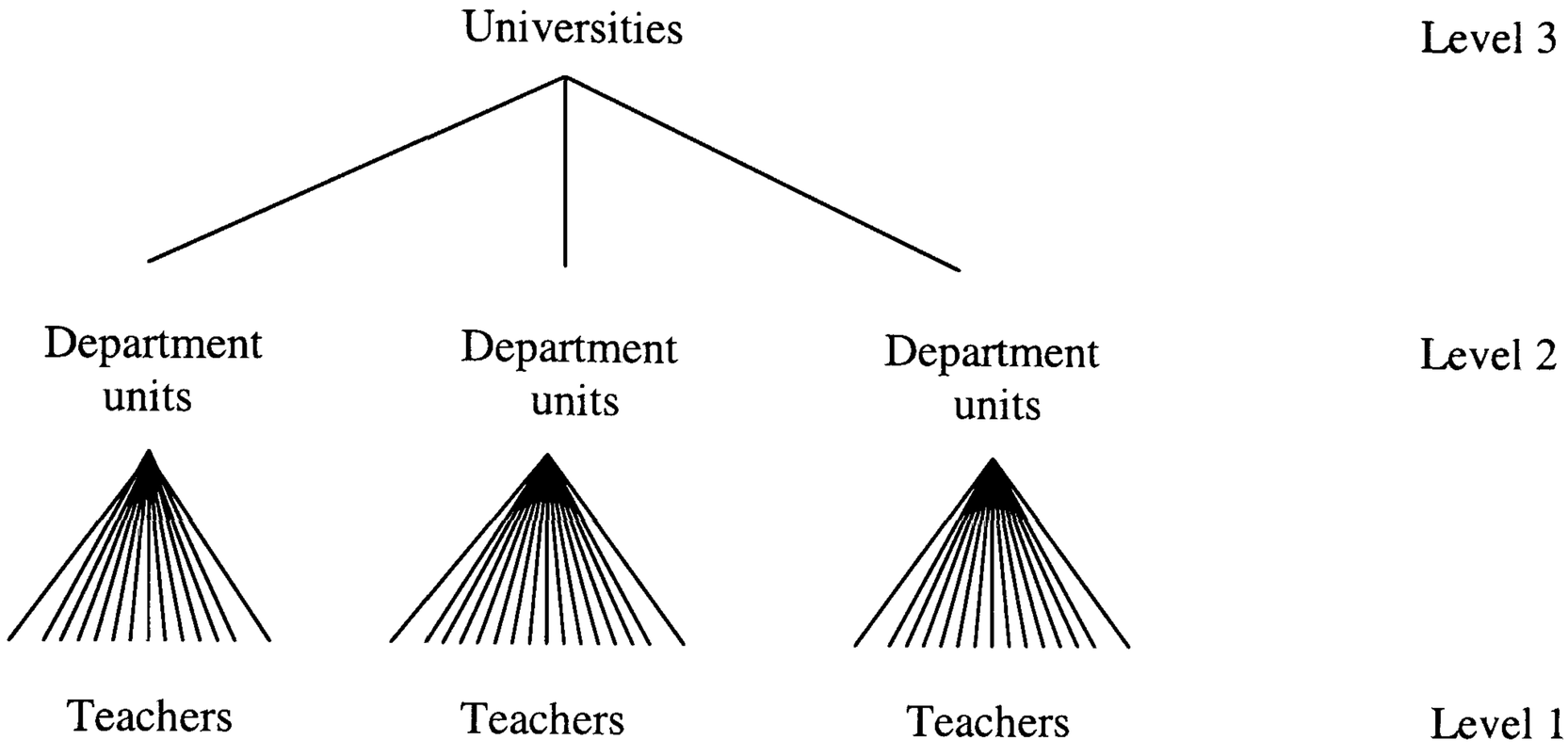


Figure 6.2 Hierarchical Structure of Ratings Feedback Data in this Study.

Following the collection of mid-semester ratings and without knowledge of the feedback ratings, teachers were made to remain in their original department groupings that were conveniently used as the peer groups. It was felt this would increase the chances of group members meeting to talk and being supportive as required by the intervention. Fourteen peer group clusters emerged in University A and nine in University B. The names of these department units and the number of teacher volunteers in each were written on individual slips of paper that were then folded and placed in a large container. Following the lottery procedure the container was vigorously shaken and a slip of paper blindly chosen. The process was repeated each time to draw the sample for the experimental group in the first instance.

Initial use of this simple randomisation procedure to assign peer groups clusters to the intervention groups yielded uneven group sizes. The replacement randomisation procedure was therefore adopted to obtain group sizes with disparities that did not exceed two. Replacement randomisation repeats the simple randomisation procedure

until a desired balance is achieved. Schulz and Grimes (2002) stated that although this method may seem arbitrary, it is quite adequate as long as it is implemented before the experiment begins. Randomisation procedures were unrestricted, in that assignment to treatment groups was not based on pre-stratification or matching according to baseline characteristics.

A total of eleven peer group clusters were allocated to the intervention group and twelve to the control group. The sizes of peer groups depended on the number of volunteers in a particular unit and ranged from two to seven members. The twenty-three clusters could be grouped into the four disciplinary areas according to Biglan's (1973a, b) classification that places disciplines into the categories, hard-pure, hard-applied, soft-pure, and soft-applied, and labelled as natural sciences, applied sciences, humanities, and business accordingly. Six groups were formed in business, seven in humanities, four in applied sciences, and six in the natural sciences.

Instrumentation

The dependent variable for this study was teaching effectiveness as measured by student ratings on the Students' Evaluation of Educational Quality (SEEQ) instrument (Appendix E). When L'Hommedieu, Menges and Brinko (1990) critically reviewed the studies used in a follow-up meta-analysis to Cohen's (1980) synthesis on the impact of student ratings feedback, they suggested that the overall magnitude of effect was almost certainly attenuated by methodological problems in the various studies. Among the recommendations was that future research should use standardised instruments such as the SEEQ. The thinking is that, the test of the utility of feedback from student ratings is more easily interpreted if results are not confounded with the reliability and validity

threats associated with non-standardised instruments. This does not imply that ratings forms are easily transferable across cultural boundaries.

There is almost full agreement in the ratings research field that because teaching is a multidimensional process, ratings forms should reflect this multidimensionality, especially when the purpose of the evaluation is for teaching improvement purposes, (Marsh, 1991a, b; d'Apollonia & Abrami, 1997a; Cashin, Downey & Sixbury, 1994; Feldman, 1989b). The SEEQ is recognised as valid and seems to capture the multidimensional nature of teaching quite well.

The SEEQ is a well-established multidimensional instrument and was developed by Herbert W. Marsh and has been thoroughly researched over the years. It is reported by Marsh and Dunkin (1992) that the items on the SEEQ were originally derived from a review of the literature and other ratings forms; interviews with university teachers and students about characteristics of effective teaching; teacher judgements about the potential usefulness of different items; and an examination of open-ended student comments on pilot versions.

The validity of the SEEQ is based on over 30 factor analyses in different settings, multitrait-multimethod analyses, logical analysis of the qualities of effective teaching, as well as being supported by principles of adult learning. As noted in Chapter 2 the estimated reliability for the SEEQ is very high although much of the research was in the North American context (Marsh & Hocevar, 1991a; Marsh, 1982; Marsh, 1987).

Despite being developed in North America, the SEEQ has been used in numerous and diverse cultural settings including, the United Kingdom, Australia, India, China, Philippines, New Zealand, Hong Kong, and Spain, where its factor structure has

been maintained (Coffey & Gibbs, 2001; Marsh & Roche, 1993; Watkins, 1994). In all these applications the SEEQ was demonstrated as being appropriate to measure teaching effectiveness. The SEEQ is available for free use with permission. Approval for use of the questionnaire was duly obtained from Herbert W. Marsh (personal communication, November 24, 2001).

The SEEQ questionnaire contains 31 items grouped into nine scales of four to two items and which measure different dimensions of teaching effectiveness. The questionnaire also included questions for demographic/background information and asked students to provide written comments. In this way, a criticism of the SEEQ is that it is much longer than the average form that could easily lead to questionnaire fatigue. Although this might affect the quality of responses that students provide, good measures are typically long as reliability tends to increase when length increases. A summary of the nine SEEQ scales follows:

Learning/Academic Value: Whether students found the class intellectually stimulating; interest in the subject increased as a result of the class.

Instructor Enthusiasm: Staff member was enthusiastic, gave presentations that held students' interest and made the subject more understandable.

Organisation/Clarity: Course materials and objectives were clearly outlined; class presentations were well prepared.

Group Interaction: Students were encouraged to participate in class discussions, to seek help, and to express their own ideas.

Individual Rapport: Opportunities were provided that took account of individual differences in capability; staff members are accessible to individual students.

Breadth of Coverage: Various points of view were discussed; implications of various theories were contrasted.

Assessment/Grading: Feedback on graded material was valuable; graded materials fairly measured knowledge of the course as emphasised by the staff member.

Assignments/Readings: Texts and supplementary readings were valuable; assignments contributed to the appreciation and understanding of the unit.

Overall Rating: Two global ratings items, one for the overall rating of the teacher and one for the overall rating of the course.

The first part of the questionnaire asked students to respond incognito to the ratings items using a five-point Likert scale, 1 = strongly disagree to 5 = strongly agree. The second part collected demographic and background information on the student, the course, and invited students to comment on aspects of teaching they felt may need improvement.

Teachers were also asked to rate their own teaching effectiveness on the same SEEQ questionnaire but with the referent "I" rather than "lecturer" only at mid-semester (Appendix F). Teachers were specifically instructed to rate their teaching in the class nominated for use in the study. The second part on teachers' version of the SEEQ collected demographic information and responses to two questions. One question asked about the frequency with which they talked about their teaching with colleagues, with responses ranging from "always" to "never".

It emerged through informal conversation that for many teachers 'talked about teaching' was interpreted as talking about their students rather than the act of teaching itself. In hindsight, the question should have asked whether they talked with colleagues about how they teach. The response to this question was not considered in the main

analyses but is shown in Appendix G. The other question represented a global item relating to teachers' attitude towards the value of student ratings to provide useful information to improve their teaching. Responses ranged from "extremely useful" to "rarely useful". The conceptual basis for this question was that effective use of student ratings data requires teachers to value student views of their teaching (Centra, 1993).

Following procedures used by Marsh and Roche (1993), participants were also asked to rate the relative importance of the SEEQ factors for teaching in their courses on a four-point scale, 1 = very important to 4 = not important. Teachers were encouraged to use this importance rating in conjunction with the ratings results to reflect on their teaching and determine areas for improvement.

Use of the SEEQ was not straightforward. In one university, near the point of administration, there was an unexpected objection to the wording of some items, even though it was previously approved. In particular, it was felt that the items focused too much on the teacher rather than how far teaching was stimulating independent learning. As an example, item no. 12 stated, "Lecturer gives lectures/tutorials that facilitated taking notes." The objection raised was that steps were being taken to change the culture of teacher-centred instruction to that which was more student-centred and at the same time encourage students to take more responsibility for their learning. It was therefore believed that administering the questionnaire could undermine such efforts. With the understanding that the instrument cannot be modified without prior approval from the developer, and with the researcher already in the field, its use had to be negotiated.

The Intervention

As already noted the focus of the present investigation was on the use of a structured group-based peer model instead of the traditional individual model of consultation with an expert. The focus of the intervention was on: (a) disciplinary peers interacting in support groups discussing ratings feedback and (b) group members providing mutual support in interpreting and generating improvement strategies. In short, the goal was to get teachers to talk with their peers about the ratings feedback received, share knowledge and experiences, and at the same time provide mutual support to make changes to improve teaching effectiveness. This aspect of the intervention took place over a four-week period but this may have not been equal across the peer groups.

Following the collection and processing of the pre-feedback ratings teachers in the intervention group was given a ratings results packet two weeks after the ratings were collected in University A and after one week in University B. Each packet contained a letter in which participants were reminded of the nature of the intervention and proposing a date for the first group meeting (Appendix H). The letter also gave the names of the other peer partners with whom the teacher would be working in the group setting.

This packet also contained the ratings report, an interpretation guide, a copy of the SEEQ, and eight double-sided sheets with teaching ideas. An information sheet (Appendix I) on the procedures for the functioning of the peer support groups was also provided. More details on these materials are given below. The drawback of providing this amount of written material is that teachers complained that there was just too much to read, which probably resulted in non-use. Two weeks later teachers in the control

group were informed in a letter that they were not selected for the immediate receipt of the intervention but would receive training with the intervention materials on completion of the study at the end of the semester.

Ratings Report

An individualised three-page report, (Appendix J), was prepared for each teacher participant. The reports were computer generated with Microsoft Excel programmed with Visual Basics for Applications. The report provided frequencies and percentages of the ratings from students for each item in each response category for each SEEQ scale. The overall mean, standard deviation, and a bar graph that summarised the response for each SEEQ scale were also provided. No normative referenced information was included that would allow for comparison between teachers.

An important feature of the report was to highlight the discrepancy between teacher self-ratings and student ratings on each SEEQ scale with two different shades on a bar graph. A similar bar graph was used to compare teacher self-ratings and the class average for the items pertaining to the difficulty, workload, and pace of the course. Students' written comments were typed and included in the report. Long comments were summarised, otherwise comments were reported as given by students. For similar comments one statement was provided but marked with an asterisk to indicate a number of students expressed similar views.

Interpretation Guide

A one-page double-sided interpretation guide, (Appendix K), was developed from information contained in the literature and websites for American and Australian universities with teaching and learning centres. This guide contained information on

how to interpret the mean and standard deviation, taking into account the number of students who responded, how to look for patterns in student responses, and how to analyse the scores on the SEEQ scales. The guide also proposed a model for using ratings data to improve teaching.

Teaching Ideas (TIPS sheets)

One of four pre-conditions that should be met before it may be expected that the use of ratings feedback will improve teaching, is that teachers know what to do to bring about change and improvement (Centra, 1993). It was considered important to follow Wilson (1986) and Marsh and Roche (1993) to provide teachers with a set of teaching tips they could use on their own and in the peer groups. With this in mind, eight double-sided single sheets with teaching ideas, (Appendix L), corresponding to the scales of the SEEQ were prepared and labelled "Teaching Ideas and Practical Strategies" (TIPS). On average each sheet contained 12 teaching tips and participants were encouraged to adapt them to their needs.

Most of the teaching ideas were adapted from the work of Marsh and Roche (1994) who in turn drew on the work of Wilson (1986). The original development of the teaching ideas was based on suggestions from teachers who had received Distinguished Teaching awards or multiple "best teacher" nominations by graduating seniors. Permission to use the ideas was received from Professor Herbert Marsh. Additional tips came from the work of Davis, Wood, and Wilson (1983), obtained from the website of the University of California, Berkeley, where they are compiled as *A Berkeley Compendium of Suggestions for Teaching with Excellence*. Permission to use the tips was received from Barbara Davis at the University of California, Berkeley.

Peer Group Meetings

The first meeting with the different peer groups was deliberately scheduled one week after the ratings results were delivered. This was done in the hope that teachers would use the time to engage in personal reflection and become acquainted with the intervention materials before the first group meeting. The researcher acted as participant observer and played the role of guiding the discussion in the groups.

It was expected that at least two group meetings would be held lasting for about one hour on each occasion, to allow each member to receive meaningful support in learning from the feedback information. It was not expected, however, that group members would be able to discuss all their concerns in the meetings but that their interest and commitment to the process would lead to learning outside of the group setting.

Overall, most meetings were short in duration and were held during the lunch hour or squeezed in between classes. On average, meetings were held for between 40 and 60 minutes. Only 6 of the 11 groups actually met and only 2 actually had follow-up sessions.

The Interviews

Individual, face-to-face, semi-structured interviews were conducted on completion of the collection of post feedback ratings to understand how teachers experienced the intervention and get their views on the feasibility of the feedback intervention. The use of interviews as a data collection method is essentially an oral, face-to-face or telephone administration of a questionnaire involving an individual or a

group. Interviews are commonly categorised by the degree of freedom given to both interviewer and interviewee.

Although interviews are often seen as structured or unstructured, based on the extent to which the questions are standardised, they are best thought of as on a continuum leading from structured to unstructured. This research used a semi-structured format. Semi-structured interviews are guided by a set of pre-defined open-ended questions and all respondents are asked the same questions in the same order with limited probing (Fontana & Frey, 2000; Patton, 1990). The semi-structured interview format was used because of time constraint.

Semi-structured interviews have an advantage over structured interviews in terms of providing more freedom in responses from respondents as well as more flexibility for the interviewer to follow through on responses. This format also ensures that all relevant questions are covered with each interviewee. Unlike unstructured interviews, however, it does not probe deeply into respondents' opinions nor does it allow respondents to talk freely at length about issues that might be important to them (Freebody, 2003; Patton, 1990; Lincoln & Guba, 1985).

A particular strength of interviews as a method of data collection research technique is that, when well conducted it can produce more in-depth and richer information than any other data collection method, especially when conducted in the semi-structured or unstructured formats. Use of interviews also allows for follow through on incomplete or unclear responses by asking additional questions. A basic drawback of using interviews is that responses are more susceptible to biases mainly because they are not made incognito. Responses may also be influenced by the

characteristics of the interviewer and the relationship between the interviewer and the respondent (Gay & Airasian, 2000).

For example, bias might be introduced into the process from the tendency for the interviewer to seek answers that support preconceived notions, as well as the tendency of respondents to be unduly helpful by attempting to respond to what they perceive the interviewer wants to hear. Another source of bias is the misperception on the part of the interviewer of what respondents are saying especially if notes are made rather than audio taped. There is also the possibility of misunderstanding on the part of respondents of what is being asked (Cohen & Manion, 1994).

All 16 teachers in the intervention group, from both universities, who did not meet with their peer partners and 19 of 24 teachers who met with their peer partners in group meetings, representing a convenience sample, were interviewed individually. It should be noted that teachers were interviewed whether or not end of semester ratings were collected. The interview guide (Appendix M) contained 10 open-ended questions and covered views on different aspects of the intervention, benefits from discussing ratings feedback with peers, strengths and weaknesses of the feedback intervention, and recommendations for use in staff development programmes. The interviews lasted between 15 and 30 minutes and did not provide much scope for deep probing of views. Notes were made during interviews following initial objections to being audio taped. These notes were later reconstructed and supplemented with observation and field notes for the qualitative analysis.

Treatment for Control Group

Teachers in the control group did not meet in groups nor did they receive ratings results until the end of the study. On collection of the end of semester ratings, results packets identical to those given to the intervention group were sent to participants in the control group. The enclosed letter invited these teachers to one of two meetings for the purpose of demonstrating how they might use the ratings feedback and the resource materials (Appendix N). The meetings were poorly attended. But the researcher did meet with the majority of these teachers on an individual basis from going around to the different departments. Teachers were generally keen on the results, the materials received, and the discussion on the use of student ratings.

Letter of Appreciation

With the announcement of an award for instructional excellence in one university many teachers requested a certificate of participation or a letter as evidence for their teaching portfolios. Subsequently, all participants in the two universities were sent a letter of appreciation for their participation in the study (see Appendix O). Noteworthy too, is that some teachers indicated in the interviews that they had referred to and used their involvement in the research to support their applications for promotion, showing that they were taking steps to improve their teaching.

STATISTICAL PROCEDURES

Statistical analyses were based on the scale scores of the SEEQ for the 71 teachers with complete sets of pre- and post feedback ratings and on the convenient strategy of intention to treat analysis. This means that the effects of the feedback intervention were examined according to the treatment groups study participants were initially assigned to, regardless of whether the full treatment was received.

Data Screening

As recommended by Tabachnick and Fidell (1996), prior to analysis of the data, an exploratory data analysis was conducted to deal with missing values, determine if the data met certain statistical assumptions regarding the shapes of their distributions and to detect the presence of outliers. First, missing values were replaced using SPSS, with the overall mean for an item if a ratings form was at least 75% completed, following Overall and Marsh (1979).

Second, stem-and-leaf plots, histograms, and boxplots, were generated to identify outliers—extreme values, high or low—in the data set that could seriously distort the results and interpretation. Where outliers were identified the raw data were checked and the value adjusted to the next value in the distribution of scores where necessary. Apart from that, there were no indications of serious distortion and distributions were approximately normally distributed as determined from the skewness and kurtosis values in relation to their standard errors.

Conventional statistical analyses were carried out with the software package Statistical Package for the Social Sciences (SPSS) for Windows, version 11.0.

Multilevel models were estimated with *MLwiN* version 1.02 (Goldstein, et al. 1998) using the iterative generalised least squares estimation. Statistical significance was tested with a two-tailed test to allow for the possibility that the result may occur in either direction.

Given the nature of the present investigation, the criterion, $p \leq .10$ level of statistical significance was set as the standard. By convention, the alpha level of $p \leq .05$ is used to decide whether an intervention effect is statistically significant, unlikely to have occurred by chance, sampling error. Setting this low alpha level lessens the chance of committing a Type I error of wrongfully rejecting the null hypothesis. At the same time, however, this increases the chance of committing a Type II error, inferring that no difference exists between the means of the treatment groups when in fact there is one.

For this study committing a Type II error may be more serious than committing a Type I error. Committing a Type I error that the intervention is beneficial when in fact it is not, will result in no harm to either teachers or students. Neither would implementation of the strategy result in a substantial wastage of funds for university administrations. The only consequence will be that further research will not support use of this particular model of consultation.

However, if a Type II error is committed that the intervention has no beneficial effects when in fact it does, this denies universities the opportunity to implement a relatively low-cost and potentially effective strategy that may help teachers improve their teaching effectiveness and quality in student learning. Also, this may prevent uptake and widespread use of the intervention, when all that might be needed is refinement through further research. In this sense, it is more reasonable to take the

greater risk of making a Type I error in order to decrease the chance of making a Type II error by setting the alpha level to be as high as $p \leq .10$ (Gay, & Airasian, 2000).

Preliminary Analyses

A factor analysis and a reliability analysis were performed on the class-average responses to assess the validity and reliability of the SEEQ questionnaire as used in the present study. This approach is consistent with the general agreement in ratings feedback research that the appropriate unit of analysis is the teacher and not the student (e.g. Marsh & Roche, 1993; Cashin 1995). The analyses were performed on each item and on each scale for 150 classes, representing the combination of mid- and end-of-semester ratings for the initial sample of 79 teacher participants.

Although the SEEQ is regarded as psychometrically sound, it was originally developed and tested in North America. It is therefore important to demonstrate that the instrument was appropriate for application in the Jamaican context. The results would also be of much interest to Jamaican administrators who showed interested in the instrument following favourable responses from teacher participants.

Primary Analyses

For the qualitative analysis, interview, observation, and field notes were transcribed and coded and a content analysis carried out, along three themes from the data; talking with colleagues, impact of ratings feedback, and conditions for use of peer support groups. Analysis followed an iterative and progressive process of data reduction, creating data displays, and drawing conclusions (Miles & Huberman, 1994; Patton 1990). Initially, the data was read and reread and fitted into matrices to discover

patterns and connections and make comparison of responses between groups and between the two study sites.

For the quantitative analysis, descriptive statistics were used to analyse demographic and background data across the intervention groups. The Pearson Chi-square (χ^2) test was used to identify associations between the treatment received and the demographic and background variables. Group differences in pre-feedback ratings were assessed with Student's *t*-tests. The main statistical technique was multilevel modelling but ordinary regression was also performed on the data. This makes for a multimethod approach to the data analysis.

First, standard multiple regression, single level model, was performed with the teacher as the unit of analysis as this was appropriate on theoretical grounds. Typically, teaching improvement interventions are intended to affect the practice of individual teachers. Analysis was also conducted with the department units as the unit of analysis, as this formed the peer group clusters and was the unit of randomisation. The results should therefore be of some interest. Second, multilevel regression analysis, multilevel model, was conducted because this method is recognised as the appropriate statistical procedure for analysing hierarchically structured data such as that in the present study, and as is common in the social sciences. This was the primary statistical technique.

In the context of the present investigation, a multimethod data analysis seems defensible for at least two reasons. In the first place, multiple regression analysis is a less complex and more familiar analytical technique than multilevel modelling. The results of the multiple regression might therefore seem more attractive and better interpretable by practitioners—teachers, administrators, and staff developers. In this way, the findings are more accessible to inform practice and policy.

A criticism of educational research is that research findings are not easily accessible to practitioners and policy-makers as reports tend to be technical, as if written for other researchers than to be used to improve educational practice (Hargreaves, 1999; Gorard, 2002a; Edwards, 2000). Second, performing the analysis with two different approaches serves to compare the results obtained. This would be used to determine the implications for substantive interpretation of failing to take account of clustering in the data structure on the conclusions about the relative effectiveness of the feedback intervention and probably point direction to a rigorous analysis for feedback interventions.

Multiple Regression and Multilevel Modelling

Multiple regression, with ordinary least squares (OLS) is a statistical technique that is used to examine the linear relationship between a single dependent variable and two or more predictor—also known as independent or explanatory—variables. This is useful for predictions and explanations of dependent variables from carefully chosen explanatory variables (Tabachnick & Fidell, 1996; Hopkins, Hopkins & Glass, 1996).

One of the underlying assumptions that should be satisfied for the appropriate application of OLS regression is that random error terms (residuals) are independent. Violation of this assumption of independence may arise when cluster sampling has been employed because individuals within a particular cluster can be expected to be more alike because of their common experience. This is expected to lead to measures that are somewhat dependent rather than independent, so the associate error terms will be correlated. Ignoring this statistical dependency will usually underestimate standard errors for regression coefficients, leading to low p-values that make explanatory

variables appear statistically significant—at the traditional 5% level—when they are not.

Multilevel modelling (MLM)—alternatively referred to as hierarchical linear modelling—is recognised as the most appropriate procedure for analysing clustered or hierarchical data (Snijders & Bosker, 1999; Kreft & de Leeuw, 1998; Hox, 1995; Raudenbush & Bryk, 2002; Goldstein, 2003). Unless of course, it can be shown as not appropriate for a particular situation or that there is no significant dependency, intraclass correlation, that resulted from a clustered sample design (cf. Fitz-Gibbon, 1997).

Fitting multilevel models has many advantages. First, it provides a convenient framework in which to model non-independence among variables across different levels of hierarchically structured data. As such, it yields more precise estimates of standard errors that take account of, and adequately adjusts for the dependence that may occur from the clustering of individuals in groups.

Second, MLM allows data from different levels of the hierarchy to be modelled simultaneously instead of having to choose whether to analyse data at the individual or group level, neither of which is entirely satisfactory. It therefore makes statistically efficient use of data allowing variables measured at the individual and group level to be included in the same model.

At its simplest, multilevel modelling can be thought of as multilevel regression, an advanced form of ordinary regression. In a simple two-level hierarchy, for example, regression coefficients are calculated on the first level of analysis, the resulting regression coefficients then become dependent variables for regression at the second level of analysis. An important distinction from ordinary regression is that multilevel

modelling separates between-group and within-group effects, and does not assume that individual observations are independent or that error terms are not correlated (Raudenbush & Bryk, 2002; Goldstein, 2003).

This statistical technique, developed in the mid-1980s, is used extensively in school effectiveness research (e.g. Daly & Defty, 2003; Raudenbush & Willms, 1991; Goldstein, et al. 1993; Fitz-Gibbon, 1991) but researchers into teaching effectiveness with ratings feedback are only just beginning to use the technique (e.g. Griffin, 2001; Marsh, Rowe, & Martin, 2002). No previous feedback intervention studies were found to have applied the technique.

Variables

Table 6.2 provides a brief description of the variables used in both the multiple regression and multilevel analyses. For these analyses, OLS regression and MLM, the dependent variable and all continuous explanatory variables were initially grand mean centred—that is, the overall mean was subtracted from each score to form deviation scores. Because of this centring, the regression coefficients and parameter estimates of these variables can be interpreted as the average value in the sample and the intercept now reflects the adjusted mean post feedback ratings score for the typical teacher with average values on the explanatory variables.

After centring, the scores were standardised to have a mean of 0 and standard deviation of 1. The interaction term was the cross product of the centred baseline score and the feedback variable, which was then standardised (Aiken & West, 1991). Categorical explanatory variables were dummy coded as 0 or 1, in this way the resulting

coefficients represent the group that has been assigned the value of 1, the group being contrasted.

Table 6.2
Description of Variables Used in this Study

Variable	Description
POSTMEAN	Mean post feedback ratings scores on the SEEQ, continuous variable
BASELINE	Mean pre-feedback ratings scores on the SEEQ, continuous variable
FEEDBACK	Intervention indicator, 1= feedback, 0 = control representing group assignment
GENDER	Dummy coded Male = 0 Female = 1
AGE	Continuous variable for age of participants
EXPERIENCE	Continuous variable for teaching experience
FIRST[a]	First year (vs. third year) = 1, else = 0
SECOND[a]	Second year (vs. third year) = 1, else = 0
FOURTH[a]	Fourth year (vs. third year) = 1, else = 0
APPLIED SCIENCE[b]	Applied (vs. Business) =1 otherwise = 0
HUMANITIES[b]	Humanities (vs. Business) =1 otherwise = 0
NATURAL SCIENCE[b]	Natural (vs. Business) =1 otherwise = 0
BASELINE × FEEDBACK	Interaction term of treatment received and pre-feedback ratings
TYPE	Dummy coded Small = 0, Large = 1

Note: Reference group has the lowest mean ratings: [a]Third year, [b]Business

Ethical Research Practice

Producing sound, innovative and informative research is no mean task (Schmucker, 2001). From the descriptions given in this chapter it should be obvious that the research process did not emerge in an artificially neat and organised manner that belies the complexity of conducting educational research. Although this might not be fully appreciated in some circles, no attempt was made to disguise the messiness but to tell what actually happened in the process of carrying out the research.

At a recent educational research conference, one researcher admitted to a roomful of people that because of an unexpected outcome thought was given to the idea of 'fixing' the results. It can be assured that in the case of the present investigation no attempt was made to 'fix' any aspect of the process or results because of the commitment to academic honesty and adherence to the principles of ethical research. It is hoped that this account will be of value to other researchers wishing to replicate the study.

Chapter 7

Results

This chapter reports on the quantitative analysis of the data collected with the ratings forms in two main sections, preliminary analyses and primary analyses. The preliminary analyses cover a factor analysis and a reliability analysis to assess the appropriateness of use of the SEEQ questionnaire in the research context. The primary analyses give the results from descriptive statistics, OLS regression, and multilevel modelling used to examine the effects of group-based peer consultation as investigated in the present study. The chapter closes with a comparison of results from the OLS regression and multilevel modelling.

PRELIMINARY ANALYSES

Factor Analysis

A factor analysis is a statistical procedure used for ordering and simplifying the interrelationships among many variables (Kline, 1994). The main applications are to (a) reduce a large number of variables into fewer dimensions or factors with minimum lost of information, and (b) identify the common underlying structure that explains the relationships among variables in a dataset. For the present study the factor analysis tested whether:

a. students were able to differentiate among components of effective teaching, and

b. the empirical factors confirmed the factors that the SEEQ questionnaire was designed to measure (Marsh, 1987).

The factor analysis in this study applied similar procedures used by Marsh and Hocevar (1991a). Consistent with the objective of this factor analysis, the number of factors to be extracted was determined a priori and did not apply the conventional methods of examining the Scree plot and retaining factors with eigenvalues over one. A trial analysis was run using principal axis factoring and an oblique rotation as commonly used in SEEQ research. However, the principal axis factoring extraction method followed by a Varimax rotation provided the more optimum solution.

In principal axis factoring only common or shared variance is analysed, compared with the assumption of the principal components analysis extraction method, for example, that *all* the variability in an item should be used in the analysis. A Varimax rotation attempts to achieve a 'simple structure' by minimising the number of

items that load highly on a factor, using the orthogonal assumption that the factors are uncorrelated. The aim here is to make the factor solution more interpretable.

Results of the factor analysis is summarised in Table 7.1. This presents factor loadings for items within each factor structure and the communality value, h^2, for each item in the last column. The communality is the proportion of variance that each item shares with other items in the common factors. From examination of the results it is clear that the factor structure of the SEEQ was satisfactorily maintained in the factor solution, with moderate to high loadings, .49 to .84, on the targeted factors, median loading = .67. It is important to note, as pointed out by Marsh and Hocevar (1991a) that the Overall Rating items are not intended to measure a specific component of teaching effectiveness but are strongly related to the Instructor Enthusiasm and Learning/ Academic Value factors. This was found to be true in this present study.

Taking this into account, the other 29 items loaded higher on the factors they were designed to measure than on any other factor, as indicated with the bolded figures in Table 7.1. The exception was the Organisation/Clarity scale in which all the items loaded higher on the Learning/Academic Value factor, represented in italic. It is curious that the items in the Organisation scale were grouped with the Learning items.

On the basis of how the questions were structured, it seemed that students were unable to make a clear distinction among the items relating to preparation and delivery of course materials; preparedness for classes; whether instruction was in line with course objectives; and the impact on their learning and their interest in the course. The content of this scale would probably need to be further examined if the instrument were to be adopted for used in Jamaica.

The 250 non-targeted loadings were much smaller with a few exceptions. Nine items showed high cross-factor loadings of over .40 on related factors, and an item in the Enthusiasm scale loaded highly with .53 on the Learning subscale just slightly lower than the loading on its own factor. Aside from this, the factor structure and composition of the factors were consistent with that previously reported in SEEQ research.

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy that tests whether the partial correlations among items were small, was found to be 0.959, which is well above the suggested minimum of 0.70 (Morgan & Griego, 1998). This result indicates that the degree of common variance among the items is very good and that it was appropriate to apply factor analysis. Further, the Barlett's test of sphericity that tests the overall significance of correlations within a correlation matrix was highly significant (sig. = 0.000), to suggest that correlations are real and not attributable to chance or sampling error. The correlation matrices for the scales on the student questionnaires by students and by classes are shown in Appendix P.

Separate factor analyses for the two universities and for mid-semester and end of semester ratings showed reasonably well formed factor structures similar to the results presented in Table 7.1 but with more cross-factor loadings between the Learning and Enthusiasm scales. The factor solution for University A was less clear than that for University B. Factor analysis of teacher self-ratings also showed clear loadings on the factors targeted with the exception that Breadth of Coverage loaded higher on the Organisation/Clarity factor. In summary, the results of the factor analysis support the original factor structure of the SEEQ, demonstrating its construct validity, and its suitability for use in the Jamaican setting.

Table 7.1
Factor Analysis Results of the SEEQ Items for the Total Sample after Rotation

SEEQ Scale/Item (Summarised)	Factor Loadings							
	I	II	III	IV	V	VI	VII	h^2
Learning/Academic Value								
Course challenging, stimulating	**.543**	.448	.183	.165	.265	.161	.228	.741
Learned something valuable	**.709**	.265	.332	.138	.176	.228	.268	.871
Increased subject interest	**.719**	.282	.310	.260	.199	.228	.218	.919
Learned and understood subject matter	**.683**	.142	.279	.304	.245	.291	.009	.837
Instructor Enthusiasm								
Enthusiastic about teaching	.298	**.625**	.294	.402	.196	.309	.176	.893
Dynamic and energetic	.360	**.698**	.229	.383	.249	.254	.138	.966
Enhanced presentation with humour	.331	**.612**	.177	.352	.284	.355	.124	.865
Teaching style held your interest	.530	**.562**	.245	.371	.245	.272	.104	.949
Organisation/Clarity								
Explanations were clear	*.556*	.240	.282	.360	.380	.254	.009	.943
Well explained and prepared materials	*.609*	.260	.242	.359	.338	.283	.157	.930
Lessons agreed with course objectives	*.639*	.184	.242	.365	.317	.297	.216	.912
Lectures facilitated note taking	*.515*	.253	.005	.166	.406	.257	.106	.647
Group Interaction								
Encouraged participation in discussions	.281	.171	**.759**	.239	.276	.132	.222	.885
Invited to share ideas/knowledge	.193	.141	**.839**	.260	.199	.173	.202	.939
Encouraged questions and gave answers	.404	.321	**.572**	.407	.220	.224	.206	.919
Encouraged to express own ideas	.283	.238	**.697**	.403	.207	.226	.216	.936
Individual Rapport								
Friendly towards individual students	.215	.300	.355	**.714**	.169	.220	.112	.872
Welcomed students seeking help	.268	.285	.339	**.705**	.139	.329	.158	.920
Interested in individual students	.307	.311	.288	**.723**	.165	.283	.007	.910
Accessible to students after class	.250	.167	.213	**.656**	.318	.253	.139	.750
Breadth of Coverage								
Contrasted various theories	.332	.244	.308	.211	**.674**	.246	.147	.858
Gave background of concepts	.314	.255	.308	.267	**.673**	.230	.260	.904
Gave different points of view	.319	.246	.404	.222	**.593**	.246	.263	.881
Discussed developments in the field	.321	.131	.413	.283	**.494**	.159	.244	.760
Assessment/Grading								
Feedback on assessments valuable	.283	.175	.160	.322	.216	**.752**	.131	.870
Evaluation methods fair/appropriate	.271	.231	.162	.274	.180	**.774**	.169	.899
Assessments tested course contents	.266	.258	.221	.213	.177	**.767**	.247	.917
Assignments/Readings								
Required readings were valuable	.201	.122	.325	.135	.190	.226	**.809**	.921
Readings contributed to understanding	.386	.205	.312	.172	.304	.290	**.635**	.900
Overall Rating items								
Overall course rating	*.624*	.394	.273	.324	.199	.335	.213	.926
Overall instructor rating	.489	*.492*	.255	.390	.200	.343	.156	.888

Notes: h^2 = Communality. N = 150 classes. Extraction method: Principal Axis Factoring. Rotation method: Varimax with Kaiser normalization. Factor loadings in bold are for items designed to measure each factor.

Reliability Analysis

Reliability of a questionnaire is characterised by the consistency with which it assesses a given phenomenon. As already mentioned in Chapter 2 in student ratings of teaching, internal consistency reliability—the consistency of ratings by individual students—is not particularly useful. More useful and appropriate is interrater reliability, or the degree of agreement among students within the same class rating the same teacher on the same component of effective teaching (Marsh & Dunkin, 1992; Cashin, 1995). This present reliability analysis is similar to the procedures followed by Marsh and Roche (1993) in their study of the SEEQ in an Australian setting for the first time.

The measure of interrater reliability was estimated from the class-average response using the intraclass correlation coefficient (ICC). The ICC is obtained with a one-way analysis of variance (ANOVA) that compares the within-group variability with the between-group variability. As responses from students in a class ought to be more alike than students from different classes, because of their common experience, estimates of reliability for class-average responses are therefore expected to be high when differences between classes are much larger than differences within classes. The analysis was run for the total scale and for scale scores, for the total sample of classes and then for three categories of class sizes; small, medium, and large, that closely aligns to the local context.

The interrater reliability estimate found for the total scale was very high and equals to 0.93, indicating that there was substantial agreement in the ratings in classes for the total scale, and that the items are measuring the construct teaching effectiveness. This compares favourably with that of $r = 0.89$ found in Australia

(Marsh & Roche, 1993) and $r = 0.90$ in North America (Marsh, 1987) with the SEEQ. The median interrater reliability estimates for classes on the total scale for small classes was $r = 0.82$, with $r = 0.88$, for medium size classes, and $r = 0.85$ for large classes (see Table 7.2). These estimates are indicative of a high interrater reliability for all class sizes.

Considering that scale scores are generally more reliable than the scores of individual ratings items, a comparison of estimates of reliability for the total sample and different class sizes is presented in Table 7.2. As is apparent, the scales have good reliability estimates that are above the acceptable standard of .70, with one exception. In the large class size category there is only a moderate agreement for the Assessment/Grading scale, $r = 0.56$, which is not acceptable. Generally speaking, the findings of this analysis provide strong evidence for the reliability of the SEEQ.

Table 7.2
Interrater Reliability Estimates: Total Sample and Different Class Sizes on SEEQ Scales

Scales	Class size			
	Total	Small	Medium	Large
Learning/Academic Value	.89	.87	.92	.91
Instructor Enthusiasm	.90	.86	.93	.93
Organisation/Clarity	.90	.89	.93	.84
Group Interaction	.84	.84	.85	.79
Individual Rapport	.84	.83	.86	.81
Breadth of Coverage	.86	.85	.89	.79
Assessment/Grading	.79	.82	.82	.56
Assignments/Readings	.73	.75	.74	.77
Overall Rating items	.93	.92	.94	.97
Summary statistics				
Number of classes	150	67	65	18
Minimum class size	3	3	15	25
Maximum class size	36	14	24	36
Mean class size	16	9	19	29

Note: Subscale scores formed from the mean of items in each scale.

In summary, the preliminary analyses provided good support for the psychometric properties of the SEEQ, and suggest that the instrument is valid and reliable, and can be used confidently as a source of information to assess teaching effectiveness in Jamaica.

PRIMARY ANALYSES

Participant Characteristics

Table 7.3 summarises the demographic and background characteristics of teachers in the study in terms of percentage. Of the 71 teachers about 51% were females and 49% were males with the distribution of gender within treatment groups also relatively equal instead of being dominated by any one gender. The mean age was 42 years with the typical teacher in the feedback group being only slightly older, about 43 years, compared to the typical teacher in the control group with 41years. Just over 50% of the sample had less than 7 years teaching experience from two categories combined, 0 – 6 years, whereas about 34% of teachers had over 10 years of teaching experience. Majority (72%) of participants held the rank of lecturer with only 28% holding the ranks of assistant lecturer and senior/principal lecturer. This compares well with the total population.

Just under 90% of the participants were teaching in the disciplinary areas of humanities (30%), business (31%), and natural sciences (27%), and only about 12% were located in the applied sciences. More participants in the control group were teaching in the soft disciplines (humanities and business) than in the hard disciplines (applied and natural sciences). Meanwhile, teachers in the feedback group seemed more

or less spread out across the four disciplines. Nearly 60% of the participants taught senior students, Years 3 and 4, and about 40% instructed junior students, Years 1 and 2. Just under a third (31%) of teachers in the control group compared with 26% in the feedback group taught the senior students, Years 3 and 4.

Table 7.3
Description of Participants in Percentages by Intervention Groups

Variable	Group		Total
	Control *n* = 37	Feedback *n* =34	
Gender			
Female	29.6	21.1	50.7
Male	22.5	26.8	49.3
Age			
25-30	8.5	5.6	14.1
31-40	15.5	11.3	26.8
41-50	22.5	23.9	46.5
Over 50	5.6	7.0	12.6
Mean (years)	41.0	42.9	42.0
Teaching Experience			
0–3	16.9	9.9	26.8
4–6	15.5	11.3	26.8
7–10	7.0	5.6	12.7
10+	12.7	21.1	33.8
Mean (years)	6.2	8.0	7.1
Academic Rank			
Asst. Lecturer	8.5	5.6	14.1
Lecturer	35.5	36.6	71.8
Senior Lecturers*	8.5	5.6	14.1
Discipline			
Applied Science	2.8	9.9	12.7
Humanities	18.3	11.3	29.6
Business	18.3	12.7	31.0
Natural Science	12.7	14.1	26.8
Course Level			
Year 1	5.6	9.9	15.5
Year 2	15.5	11.3	26.8
Year 3	18.3	12.7	31.0
Year 4	12.7	14.1	26.8

*Senior and Principal Lecturers

By expressing the demographic and background variables as categorical, use of Pearson Chi-square (χ^2) tests to determine the association between these variables and the treatment received found no statistically significant associations for: gender ($\chi^2 = 1.13$, $p = .29$), age ($\chi^2 = 0.89$, $p = .83$), teaching experience ($\chi^2 = 3.28$, $p = .35$), academic rank ($\chi^2 = 0.69$, $p = .71$), and the course level at which instruction was offered ($\chi^2 = 1.95$, $p = .58$).

When the disciplinary areas were grouped and categorised according to Biglan's (1973a,b) classification, *hard* (applied and natural sciences) and *soft* (humanities and business), the proportions in the disciplinary groups were found to be statistically significant only at the 10% level for the two groups ($\chi^2 = 3.05$, $p = .08$). Almost 37% of participants in the control group taught in the *soft* disciplines compared with only 21% in the feedback group. Conversely, 24% of participants in the feedback group taught in the hard disciplines, compared to about 16% in the control group. ANOVA test on the means for the disciplines by intervention groups found statistical significance at the 10% level for only Business ($F(1,20) = 2.96$, $p = .10$) in favour of the control group. Overall, teacher participants in both the control and feedback groups were similar on all measured variables.

Attitude towards the Value of Student Ratings

The effective use of student ratings requires that teachers see student views about their teaching as valuable information. Figure 7.1 illustrates the views held by teacher participants on the value of student ratings to improve their teaching as obtained from the teacher self-evaluation forms. The majority (45%) appear to be quite certain about the utility value of student ratings to improve teaching effectiveness as evidenced

by their response of 'extremely' useful. A slightly smaller percentage (32%) thought student ratings were only 'moderately' useful, and just under 25% indicated that student ratings were 'somewhat' or 'rarely' useful to enhance teaching effectiveness. A slightly larger percentage of participants in the control group (25%) than in the feedback group (18%) appear to hold more positive views about the usefulness of student ratings. This difference is likely due to chance as explained below.

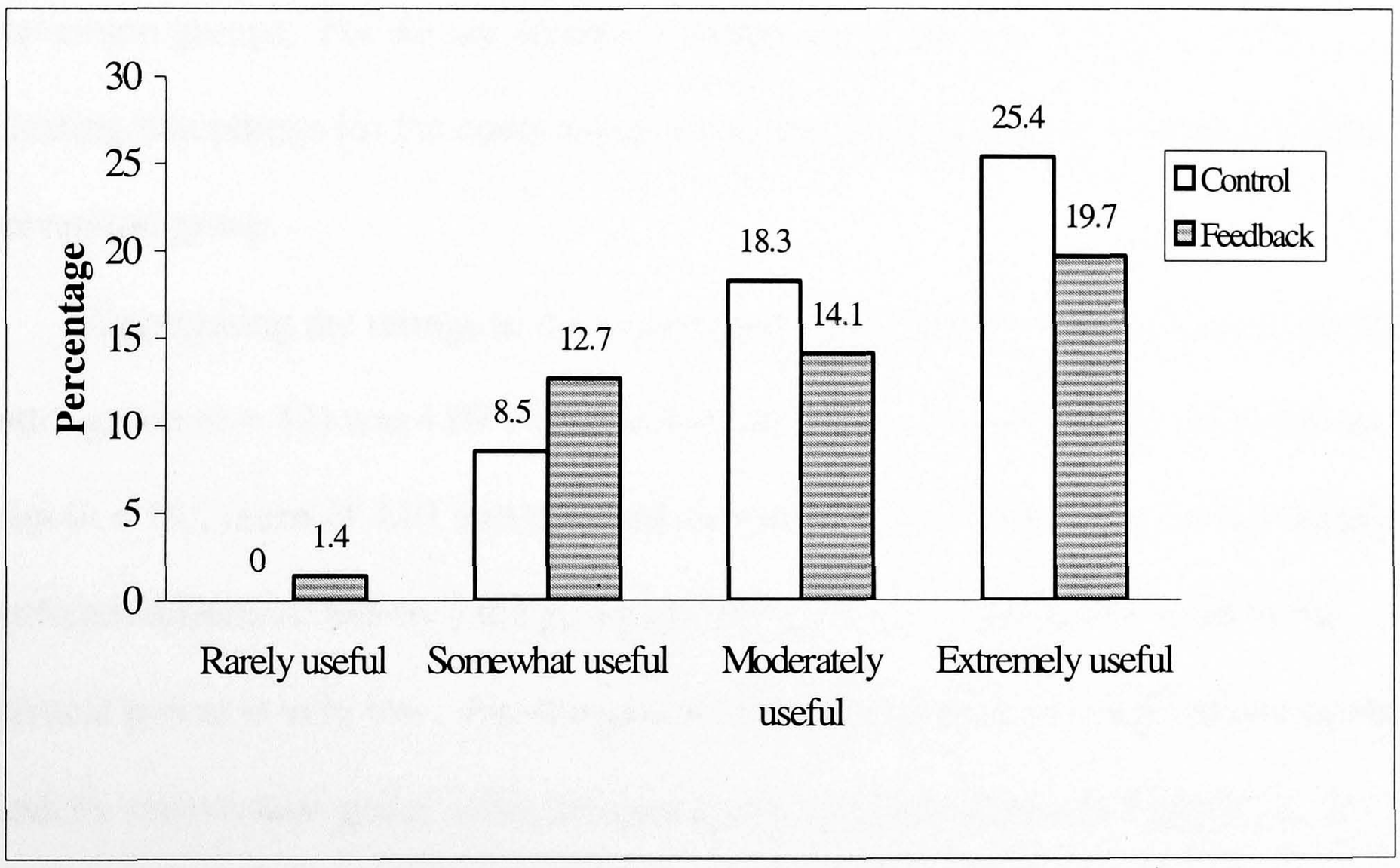


Figure 7.1 Attitude towards the Value of Student Ratings

Treating the attitude scale as ordinal, comparison of attitude ratings was performed with the non-parametric Mann-Whitney *U* test. No statistically significant differences were found between the intervention groups ($p = .18$, Mann-Whitney-U). It is recognised, though, that using a single item to represent attitude is somewhat problematic and that multiple items would have probably provided a more reliable result. Illustrations of teacher perceptions of their own teaching effectiveness as obtained with the teacher self-evaluation questionnaires are presented in Appendix Q.

Feedback Ratings Scores

Analysis of the pre-feedback ratings scores for teachers found that the control group (n = 37) had a mean of 4.08 with standard deviation of .30. The mean score for the intervention group (n = 34) was 4.00 with standard deviation of .39. The difference was not found to be statistically significant ($t(69) = 0.97$, $p = 0.34$). Analysis on scale scores also found no statistically significant difference between the control and intervention groups. For the pre-feedback ratings the effect size was $ES = -0.23$, indicating that ratings for the control group was just slightly higher than those for the intervention group.

Aggregating the ratings to the department-level (peer groups), the mean for the control group (n = 12) was 4.09 with standard deviation .20, and for the intervention group (n = 10), mean of 4.01 and standard deviation of .23. There was no statistically significant difference between the groups ($t(20) = 0.75$, $p = 0.46$), as is evident the statistical power is very low. An illustration of the distribution of pre-feedback ratings scores by intervention group using the peer group means is shown in Figure 7.2. It shows a clear trend in the data for the control group to be rated slightly higher than the intervention group before the feedback intervention was administered.

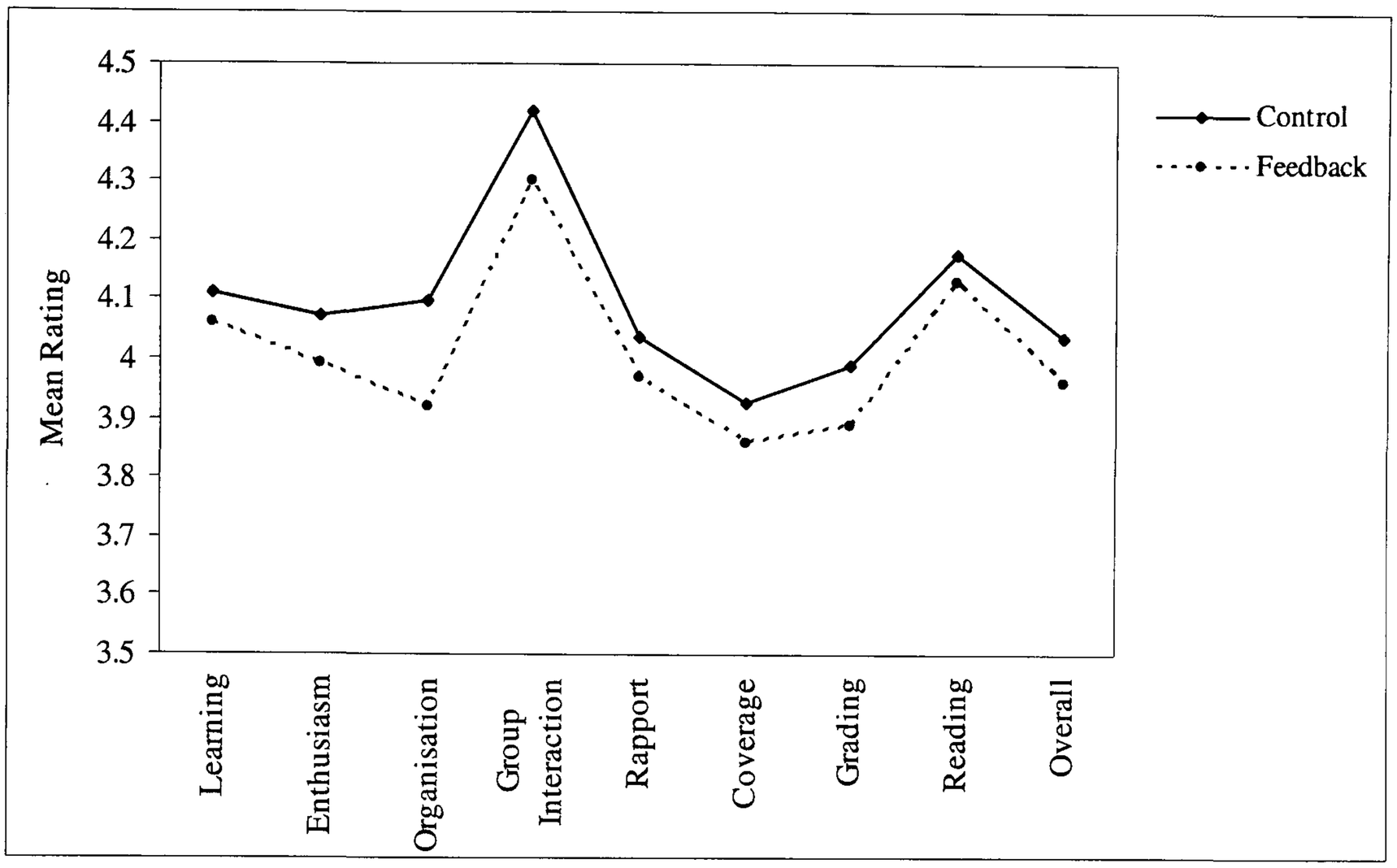


Figure 7.2 Pre-feedback Ratings on SEEQ Scales on Peer Group Means

Table 7.4 gives the means and standard deviations for the post feedback ratings on each ratings scale by intervention group at the teacher level. The Table shows that the mean ratings score for the intervention group was 4.12 with standard deviation .39, while the mean for the control group was 4.10 with a standard deviation of .45.

Table 7.4
Post feedback Ratings by Intervention Group on SEEQ Scales for Teachers

Scales	Intervention Group			
	Control		Feedback	
	n = 37		*n* = 34	
	Mean	SD	Mean	SD
Learning/Academic Value	4.11	.43	4.14	.35
Instructor Enthusiasm	4.16	.60	4.16	.49
Organisation/Clarity	4.06	.51	4.10	.38
Group Interaction	4.27	.51	4.34	.41
Individual Rapport	4.10	.51	4.13	.45
Breadth of Coverage	3.94	.45	4.00	.45
Assessment/Grading	4.08	.45	4.01	.45
Assignments/Reading	4.10	.46	4.11	.48
Overall Rating items	4.04	.55	4.05	.47
Mean Ratings Score	4.10	.45	4.12	.39

ANALYSIS I: OLS REGRESSION MODEL

Multiple regression analyses were performed separately on each scale, creating nine unique models. A final model was fitted with the mean pre- and post feedback ratings scores to provide a better representation of the effects for the feedback intervention. For each regression model the pre-feedback score served as covariate to adjust for differences in initial ratings.

The models were fitted with the post feedback score as the dependent variable and with the explanatory variables; feedback (intervention group indicator), baseline (pre-feedback score), and interaction (feedback × baseline). As recommended by Aiken and West (1991), to make the OLS regression results meaningful with interpretable values, the baseline score was centred and the interaction term formed from this centred score. All scores were standardised to have a mean of 0 and standard deviation 1. Analysis was conducted with the teacher as the unit of analysis as well as with the department units as the unit of analysis

Teacher-level Analysis

Table 7.5 contains the results of the ordinary regression analyses with the teacher as unit of analysis. The results are reported in terms of unstandardised regression coefficients (*b*). To ease interpretation, regression coefficients represent the change in the dependent variable associated with each unit change in the independent variable relative to other variables in the model. These are expressed in standard deviation units because all continuous variables were standardised. The squared multiple correlation, R^2, is the proportion of variance in the dependent variable that is

accounted for by the linear combination of explanatory variables in each model. The standard error is an index of the precision of an estimate.

Table 7.5
OLS Regression Results for Teacher-level Analysis

Scales	R^2	Baseline	Feedback	Interaction
		b (S.E.)	*b* (S.E.)	*b* (S.E.)
Learning/Academic Value	.492	0.979 (.150)	0.165 (.173)	–0.474 (.184)**
Instructor Enthusiasm	.689	0.969 (.101)	0.150 (.136)	–0.270 (.137)**
Organisation/Clarity	.495	0.881 (.146)	0.328 (.175)*	–0.287 (.183)
Group Interaction	.628	0.977 (.117)	0.411 (.150)**	–0.340 (.153)**
Individual Rapport	.662	1.001 (.108)	0.227 (.142)	–0.365 (.144)**
Breadth of Coverage	.449	0.723 (.143)	0.258 (.181)	–0.090 (.185)
Assessment/Grading	.365	0.618 (.151)	-0.031 (.194)	–0.027 (.198)
Assignments/Readings	.363	0.635 (.143)	0.086 (.194)	–0.061 (.196)
Overall Rating	.590	0.894 (.118)	0.151 (.156)	–0.237 (.158)
Mean post feedback	.652	0.991 (.115)	0.241 (.144)*	-0.320 (.148)**

Notes: Mean post feedback is a composite of nine ratings scales
All multiple R^2s and *b* coefficients for baseline scores are statistically significant ($p < .01$).
S.E. = estimated standard errors
** *t*-value statistically significant at $p \leq .05$ * *t*-value statistically significant at $p \leq .10$

For the individual scales, the results show that the largest R^2 value among the teaching dimensions fitted singly was for Enthusiasm that explained 68.9% of the variance in post feedback ratings, identifying it as highly stable component of effective teaching in this data. The components Individual Rapport (66.2%) and Group Interaction (62.8%) were also stable. In contrast, Assessment/Grading (36.5%) and Assignments/Readings (36.3%) did not remain quite so stable from the mid- to end of semester ratings period. As might be expected, pre-feedback ratings are statistically and substantially associated with post feedback ratings, with regression coefficients ranging from 0.62 to 1.00.

To interpret the estimates of effects associated with the feedback variable, it is important to keep in mind that these estimates are strictly not 'main' effects. Presence

of the feedback-by-baseline interaction term in the models means that the estimates ought to be interpreted as 'average' effects. This is because the estimates of effects for 'feedback' in the different models are conditional at the mean value of 'baseline'. Although the focus of interpretation should then be on the interaction effect, the 'average' effect is still informative (Jaccard, Turrisi & Wan, 1990; Aiken & West, 1991).

From the OLS regression results in Table 7.5 (column 4), it is clear that the effect of 'feedback' on the different components of effective teaching varied somewhat. The regression coefficients for the components Organisation/Clarity ($b = .33$, $t = 1.87$, $p = .07$) and Group Interaction ($b = .41$, $t = 2.74$, $p = .01$) are moderate and statistically significant. The lack of statistical significance on the other seven components suggests that observed differences between the groups were no different than what could be expected by mere chance, though all but one are positive. No valid conclusions can be drawn.

The 'interaction' coefficients are all negative with statistical significance on four of the nine components of effective teaching (column 5). That means, among teachers with similar ratings, those with low baseline scores in the intervention group improved in their teaching effectiveness more than similar teachers in the control group in the areas; Learning ($b = -.47$, $t = -2.57$, $p = .01$, Enthusiasm ($b = -.27$, $t = -1.97$, $p = .05$), Group Interaction ($b = -.34$, $t = -2.22$, $p = .03$), and Individual Rapport ($b = -.37$, $t = -2.54$, $p = .01$). It does appear that these components represent areas that these less effective teachers were able to make relatively immediate changes to. As pointed out by Marsh and Roche (1993), this finding clearly illustrates the importance

of using a well-designed multidimensional instrument to provide feedback on the different components of effective teaching.

The difference between the groups becomes clearer when the scale scores are pooled in the final model. The 'feedback' coefficient in the final model found borderline statistical significance, 10% level, providing some evidence of a positive effect of the intervention for all teachers at different levels of teaching effectiveness ($b = .24$, $t = 1.67$, $p = .10$). This indicates that, on average, teachers in the intervention group improved in their teaching effectiveness more than those in the control group. This final model explained 65% of the total variance.

Even clearer, is evidence of a baseline aptitude-treatment interaction in the sample, indicating that the effect of feedback and consultation is moderated by teacher's initial level of teaching effectiveness as defined by baseline scores. As can be seen from Table 7.5 the interaction effect is statistically significant and negative ($b = -.32$, $t = -2.16$, $p = .04$), indicating improvement in teaching effectiveness was stronger for the initially less effective teachers—those with low baseline scores—than for other teachers. In short, the results obtained suggest that the less effective teachers may have received the most help from the feedback intervention. Figure 7.3 displays a scatter plot with fitted regression lines for the intervention groups. This shows how the interaction appears for individual teachers.

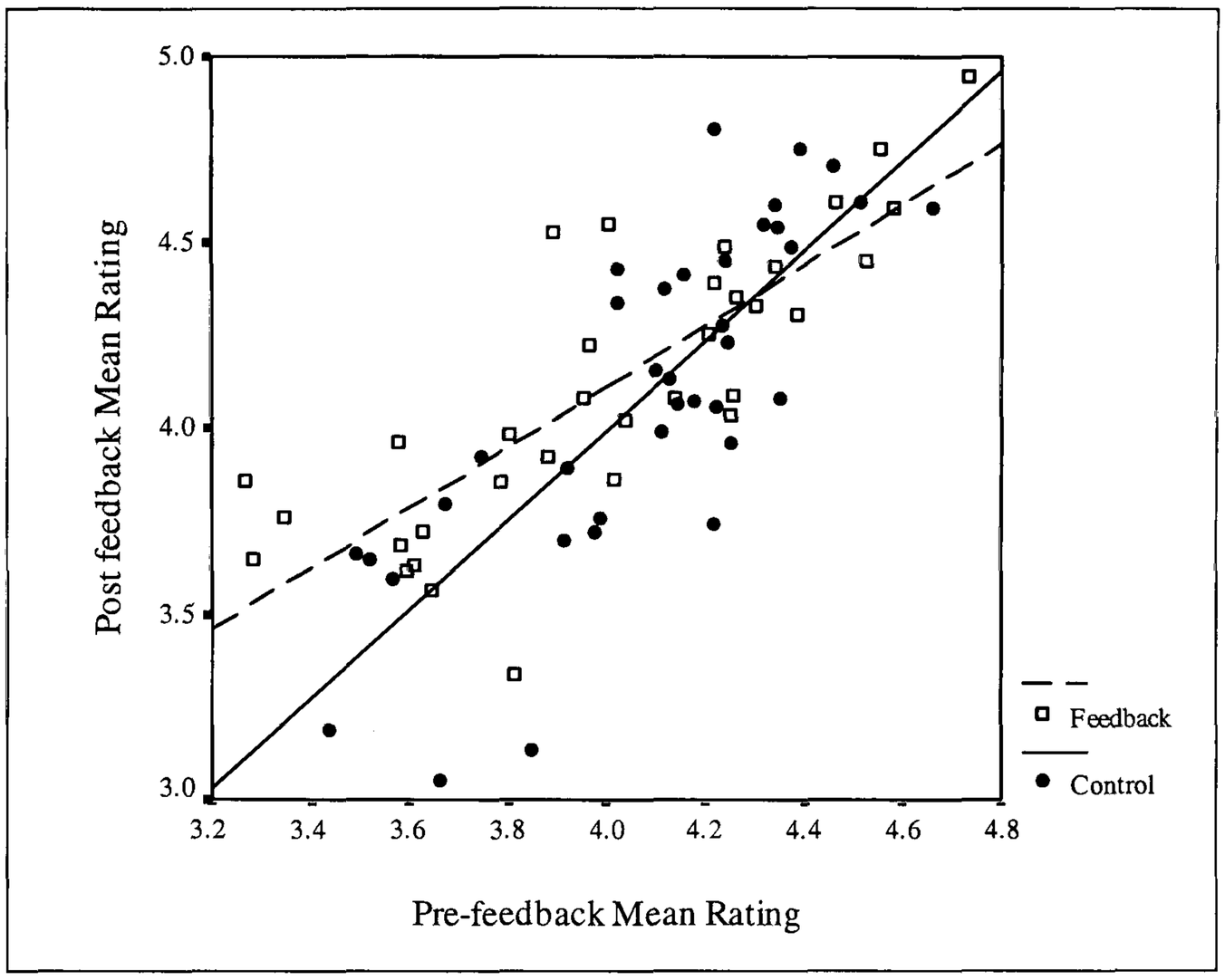


Figure 7.3 Interaction Effect for Individual Teachers by Intervention Groups

To further illustrate the feedback-by-baseline interaction the mean baseline score was divided into three levels and plotted: low (20th percentile), medium (20th to 80th percentile), and high (above the 80th percentile). Figure 7.4 displays this plot. As is clear from the graph, among staff with similar baseline scores, teachers who received low baseline ratings and who received the feedback intervention improved slightly more in teaching effectiveness than did other teachers.

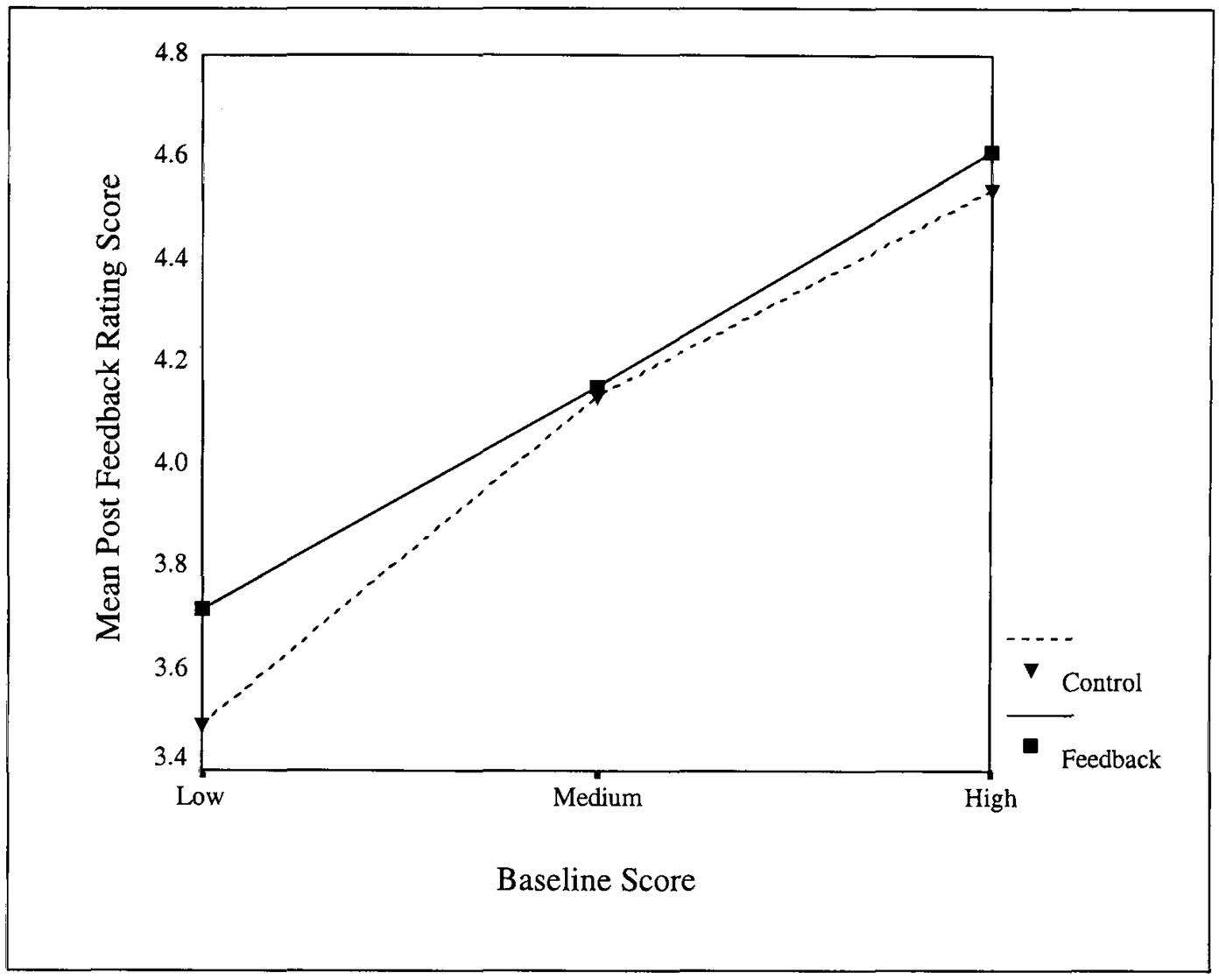


Figure 7.4 Mean Post feedback Ratings on Initial Level of Teaching Effectiveness

A Ceiling Effect?

Is the interaction effect the result of a ceiling effect? It could be said that because the relatively effective teachers had high baseline scores to begin with (close to the maximum rating or ceiling) they had limited scope for improvement. In this way, comparison with teachers who had low baseline scores and more opportunity for improvement, would give the appearance that the less effective teachers had improved more in their teaching effectiveness. If this were the case, it becomes difficult to say that the effects of the intervention were more positive for the initially less effective teachers.

It does seem unlikely, though, that the presence of a ceiling effect would provide an adequate explanation for interaction effect in this study. One good reason is that, the performance of teachers with low baseline actually remained in the 'low' band. You

could say they moved from "very low" to "low", and there is still a wide gap in mean scores even from teachers with moderate (medium) teaching effectiveness. For another reason, it is important to consider that the moderately effective teachers were not actually restricted but changed little from their initial level of teaching effectiveness. Interestingly, Pambookian (1974) found that it was the moderately effective teachers who actually improved more in their teaching effectiveness, than teachers rated as less or highly effective.

To summarise, the results of this analysis suggest that feedback and consultation in peer support groups is associated with improvement in teaching effectiveness. This effect may be generalised across teachers with different levels of initial teaching effectiveness but seemed slightly stronger for the initially less effective teachers. Before drawing any firm conclusion from these results a cautionary note must be made.

Because teachers were nested within department units, there is likely to be more similarity amongst the ratings for teachers in the same department than for those in other departments because of their common experience in the discipline. This results in a degree of dependency, making the ratings not completely independent observations. In other words, the ratings were correlated.

For ordinary OLS regression non-independence is likely to result in a downward bias in the estimates of standard errors leading to findings of statistical significance when in fact none exists. The estimates found in this analysis may therefore be misleading. Further, the teacher-level analysis overlooks the contextual effect of teaching in a particular department or discipline.

Department-level Analysis

Table 7.6 contains the OLS regression results for data generated from the teacher level data as means for the department units, which is now treated as the unit of analysis. Weighting by peer group size was adopted to take account of the variable peer group cluster sizes. Apart from this, the models were identical to those fitted with the teacher-level data.

The results show that statistically significant average feedback effects were found for the components Group Interaction, Individual Rapport, and Breadth of Coverage. These were without interaction effects. In the same way, three statistically significant interaction effects were found on the components Learning, Assignments/Readings, and Overall Rating, without a statistically significant average feedback effect. Only one component of effective teaching, Group Interaction, fully matched in statistical significance for the 'feedback' effect with the teacher level analysis, while the interaction term matched on only the Learning/Academic Value component.

Table 7.6
OLS Regression Results for Department-level Analysis

Scales	R^2	Baseline	Feedback	Interaction
		b (S.E.)	*b* (S.E.)	*b* (S.E.)
Learning/Academic Value	.619	1.281 (.272)	0.254 (.276)	–0.801 (.328)**
Instructor Enthusiasm	.651	1.086 (.229)	0.263 (.257)	–0.482 (.292)
Organisation/Clarity	.459	1.047 (.363)	0.536 (.341)	–0.486 (.422)
Group Interaction	.827	1.048 (.145)	0.635 (.189)**	–0.217 (.205)
Individual Rapport	.818	0.881 (.144)	0.396 (.185)**	0.077 (.205)
Breadth of Coverage	.688	1.060 (.200)	0.448 (.247)*	–0.446 (.273)
Assessment/Grading	.368	0.584 (.226)	–0.037 (.355)	0.012 (.399)
Assignments/Readings	.620	1.145 (.232)	0.060 (.271)	–0.743 (.295)**
Overall Rating	.692	1.139 (.216)	0.268 (.250)	–0.503 (.280)*
Mean post feedback	.750	1.140 (.187)	0.380 (.223)	–0.476 (.248)*

Notes: Mean post feedback is a composite of nine ratings scales
All multiple R^2s and *b* coefficients for baseline scores are statistically significant ($p < .05$).
S.E. = estimated standard errors
** *t*-value statistically significant at $p \leq .05$ * *t*-value statistically significant at $p \leq .10$

Turning to the coefficients for the final model that assessed overall intervention effects, it can be seen that the estimated coefficient for 'feedback' is positive and moderate ($b = .38$, $t = 1.70$, $p = .11$) but is just short of statistical significance, 10% level, because of the larger standard error. In common with the teacher-level analysis, the coefficient for the interaction effect is also negative but is statistically significant at the 10% level ($b = -.48$, $t = -1.92$, $p = .07$). Again, the results indicate that the intervention effects varied according to the initial baseline ratings scores. It must be pointed out, as explained below, estimates from this analysis are misleading so caution should be exercised in drawing conclusions from them.

At first glance, it may appear that the results from this department-level analysis reflect a causal positive relationship between the feedback intervention and teaching effectiveness. On careful consideration, however, it becomes clear that the estimates of effects from the aggregated data are measuring the effects for department units.

Therefore the results cannot be appropriately used to interpret the intervention effect for individual teachers but the effect for the peer groups and the department units they represent.

Using analysis at one level to draw conclusions at another level is referred to as 'ecological fallacy'. Analysis carried out at one level can only be reliably used to make inferences at that level, as the relationship at one level does not necessarily hold at another level (Kreft & de Leeuw, 1998; Plewis, 1998). The use of this department-level analysis is therefore limited, as intervention studies with student ratings are more interested with the application and implications for teaching effectiveness for individual teachers.

The use of aggregated data also resulted in the loss of adequate statistical power to provide reliable estimates because of the smaller sample size. It also overlooks the differences among teachers in department units. Analysis II uses multilevel modelling techniques to analyse the data that will overcome the 'unit of analysis' problem, account for the dependence of observations, and not overlook differences at the teacher level or at the department level. The results from multilevel modelling should therefore allow for more complete and accurate conclusions to be drawn.

Subgroup Analysis

Although it might have been interesting to perform a subgroup analysis—that is to estimate intervention effects on the basis of actual implementation, teachers who met in the peer groups or those who did not meet, compared to the control group—this was not considered particularly sound practice. In the first place, this analysis was not planned and based on what is known about how the intervention was implemented, it

was not plausible to expect there would be a difference. In any case, such comparisons would simply be capitalising on chance to find a difference. Additionally, because the sub-group would represent a self-selected group, and with reduced statistical power from the small group sizes, coupled with imbalance in group sizes, the results of such an analysis would be spurious leading to incorrect interpretations (Brookes, et al. 2001; Bigger, 2003).

ANALYSIS II: MULTILEVEL MODELLING

Initially, it was assumed that within a multilevel framework the data should be modelled with students as the level-1 units seeing that the ratings were actually made by students. It turned out, that the parameter estimates for the fixed effects from this analysis were quite discrepant from those found using OLS regression at the teacher- and department-level. According to Kreft (1996) the estimates of a multilevel analysis will be close to the results obtained with OLS regression but it is the standard errors that will be underestimated because of the correlated data. Comparative analyses provided by Raudenbush and Bryk (2002) seem to support this.

This was not the case in an initial analysis. As an example, the multiple regression analysis at the teacher-level yielded 'interaction' coefficient of –0.320 and standard error of 0.148. For the department-level analysis this was –0.476 with standard error of 0.248. The MLM analysis yielded an estimate of –0.062 and standard error of 0.101. As Gorard (2002b) argued—in his response to the debate on the appropriateness of OLS regression vs. multilevel modeling—when the results from the two approaches are different there is no way of deciding which one is correct.

It then became clear that the ratings data was a measure for teachers and represented an explanation of the improvement process being investigated in the present study. The students' role was simply to provide an estimate of teaching effectiveness for teachers. It is also worth noting, that it is already well established in ratings feedback research that the appropriate unit of analysis is the teacher and not the student. On this basis, four three-level MLM models were estimated with teachers at level-1, department units at level-2, and universities at level-3.

MLM Models

Models were fitted only with the mean post feedback and pre-feedback ratings scores. The models fitted are referred to as random-intercept, or more commonly variance components models, as it is only the intercept that is allowed to vary randomly across higher-level units. The slope coefficients were constrained to be fixed. The assumption is that the intercept varies as a function of the explanatory variables as well as a function of unique effects of units at the higher levels, which are assumed to constitute a random sample from among some population.

Model 1

Analysis with MLM started with fitting the unconditional or null model, Model 1. This model contained no explanatory variables and is fitted simply to apportion the variance in teaching effectiveness—the mean post feedback ratings—associated with the different levels in the hierarchy. This model is analogous to a one-way analysis of variance (ANOVA) with random effects. This model is important because it provides information about the relative influence of teachers, department units, and the institution on teaching effectiveness.

The model is defined by the equation $Y_{ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$, where Y_{ijk} is the dependent variable—mean post feedback ratings, POSTMEAN—for teacher i in department j in institution k. The term β_0 is the intercept and constant term (= 1), which reflects an adjusted mean post feedback ratings score. The residuals $e_{ijk} \sim N(0, \sigma^2_e)$, and $u_{jk} \sim N(0, \sigma^2_u)$ and $v_k \sim N(0, \sigma^2_v)$ are the random error terms for the teacher, peer group, and university levels respectively. These are assumed to be normally distributed with mean of zero and constant variance σ^2. The error terms represent the amount by which the intercept for the different levels differs from the overall mean value.

Model 2

Model 2 represents a main effects model for the feedback intervention. The equation was,

$$Y_{ijk} = \beta_{0ijk} + \beta_1 x_{ijk} + \beta_2 x_{jk} + \beta_3 x_{ijk} \qquad \text{with } \beta_{0ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk}$$

The β_1 term is the regression coefficient for pre-feedback ratings, BASELINE, which served as covariate to control for initial differences in ratings status. The β_2 term is the intervention indicator, FEEDBACK. The regression coefficient of this explanatory variable indicates the relative effect for the feedback intervention. The β_3 term is the interaction term, FEEDBACK × BASELINE which is the variable of primary interest, as it provides a better description of the relationship between the intervention and teaching effectiveness.

Model 3

Model 3 estimated the effects on teaching effectiveness due to three important teacher characteristics—gender, age, and teaching experience. In this sense, the parameters no longer represent the main effect of the intervention on teaching effectiveness but estimated effect among teachers with similar characteristics. The model was estimated as:

$$Y(\text{POSTMEAN})_{ijk} = \beta_{0ijk} + \beta_1(\text{BASELINE})_{ijk} + \beta_2(\text{FEEDBACK})_{jk} + \beta_3(\text{FEEDBACK} \times \text{BASELINE})_{ijk} + \beta_4(\text{GENDER})_{ijk} + \beta_5(\text{AGE})_{ijk} + \beta_6(\text{EXPERIENCE})_{ijk}$$

Model 4

Model 4 estimated the effects of teaching responsibility in terms of discipline and the level at which instruction is offered. The model specification was:

$$Y(\text{POSTMEAN})_{ijk} = \beta_{0ijk} + \beta_1(\text{BASELINE})_{ijk} + \beta_2(\text{FEEDBACK})_{jk} + \beta_3(\text{FEEDBACK} \times \text{BASELINE})_{ijk} + \beta_4(\text{FIRST})_{ijk} + \beta_5(\text{SECOND})_{ijk} + \beta_6(\text{FOURTH})_{ijk} + \beta_7(\text{APPLISCI})_{ijk} + \beta_8(\text{HUMANITIES})_{ijk} + \beta_9(\text{NATSCI})_{ijk}$$

Specifying a composite or full model was constrained by the small sample size. Fitting a large number of variables with such a small sample would simply be a play on chance to find statistical significance.

Presentation of Results

Parameter estimates are presented in two parts: (1) fixed effect parameter estimates that give the regression coefficients for each explanatory variable with their standard errors, (2) random effects parameters that define the variation at the different

levels of the hierarchy after taking account of explanatory variables. Regression coefficients are expressed in terms of standard deviation units because continuous variables were standardised. Expressed this way, the estimated effects may be interpreted as effect sizes (see Chapter 5).

Statistical significance of a fixed effect parameter is assessed with the *t*-statistic, defined by the ratio of a parameter estimate to its standard error, referred to the standard normal distribution. For this, the degrees of freedom depend on the number of units at the level for the coefficient being tested less the number of explanatory variables at the level in the model minus one. Nonetheless, an interpretation of statistical significance of a variable is best made in relation to the overall fit of the model more than on the testing of a single parameter. Only if the model fit improves in a significant way can one be reasonably confident in accepting the individual significant parameter (Kreft & de Leeuw, 1998; Goldstein, 2003).

Model fit is assessed on the difference between the deviances—identified by the –2*log-likelihood values—of two models being compared, which has an approximate chi-squared distribution. The degrees of freedom for the chi-square test are the difference in the number of parameters estimated in the two models.

RESULTS

Variance in Teaching Effectiveness

The results from fitting the unconditional model that partitions the variance in the mean post feedback ratings are given in Figure 7.5.

$$\mathbf{postmean}_{ijk} \sim \mathrm{N}(XB,\ \Omega)$$

$$\mathbf{postmean}_{ijk} = \beta_{0ijk}\mathbf{constant}$$

$$\beta_{0ijk} = 0.131(0.280) + v_{0k} + u_{0jk} + e_{0ijk}$$

$$\left[v_{0k}\right] \sim \mathrm{N}(0,\ \Omega_v) \ :\ \Omega_v = \left[0.124(0.156)\right]$$

$$\left[u_{0jk}\right] \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \left[0.322(0.250)\right]$$

$$\left[e_{0ijk}\right] \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \left[0.563(0.235)\right]$$

$$-2*log(like) = 194.725$$

Note: v_{0k} = institution level; u_{0jk} = department level; e_{0ijk} = teacher level

Figure 7.5 *MLwiN* Output: Results Estimating the Variance in Teaching Effectiveness

First, the intercept, β_{0ijk}, is 0.131 but this is usually of no theoretical interest. The estimated proportion of variance at level-1, e_{0ijk}, is .563, indicating that majority (56%) of variability in teaching effectiveness (post feedback ratings) is attributable to teachers themselves. At the peer group level, or the related department unit, u_{0jk}, the variance is .322, and suggests that this level accounts for 32% of the total amount of variance. Universities, v_{0k}, account for only about 12% of the variation, indicating not

variance. Universities, v_{0k}, account for only about 12% of the variation, indicating not surprisingly, that the institution has less influence on teaching effectiveness than the academic departments (discipline) in which teachers practice. But similar to Tymms' (1993) observation, the influence of departments is only on a small number of teachers compared to the more far-reaching influence of the institution. This makes the role of university administrations in improving teaching quality rather crucial.

Associated with the variance for departmental units is the intraclass correlation (ICC), measure of the amount of dependence in ratings scores, or the extent to which ratings are correlated due to the grouping of teachers by departmental units. This is computed as the ratio of the between-group variance to the total variance. It should be understood that this ICC uses a different formula from that in the reliability analysis.

The estimated ICC was $\rho = 0.32$, which is positive and moderate, showing that ratings for teachers are not wholly independent, and that there is a grouping effect due to the department unit. This also means that the assumption of independence in the OLS regression analysis would have been violated. When the ICC equals 0, there is no difference between OLS regression estimates and those obtained with MLM, because no clustering exists. The size of the ICC is an indication that the use of multilevel modelling techniques is most appropriate to analyse the data for the present study.

The finding of a moderate positive ICC is hardly surprising because of the way universities are naturally structured. The result of a grouping effect indicates that teacher grouping according to departments, confounded with disciplinary areas, has an apparent influence on teaching effectiveness.

In summary, teachers themselves seem to have the greatest influence on teaching effectiveness, academic departments and to a lesser extent the institution, also

play an important role in influencing teachers to create and maintain high levels of teaching effectiveness. Clearly, then, teaching improvement activities that aim to help faculty to develop their teaching skills within the context of their discipline may be quite beneficial for university teaching.

Effects on Teaching Effectiveness

The overall effect of the feedback intervention on teaching effectiveness was estimated by fitting Model 2 with the explanatory variables, BASELINE, FEEDBACK, and the FEEDBACK × BASELINE interaction term. The estimated coefficients and their standard errors are given in Table 7.7. The average feedback effect was statistically significant ($t(20) = 1.72$, $p = .10$) with an effect size of .24 standard deviation units (*SD*). This result says that, on average, teachers who received the feedback intervention improved in their teaching effectiveness by roughly one-quarter of a standard deviation more than teachers in the control group.

Table 7.7
Multilevel Results Estimating Effect on Teaching Effectiveness

	Model 2	
Fixed Effects	Parameter	S.E.
Intercept	–0.120	0.106
Baseline	0.977	0.113
Feedback	0.241	0.140
Feedback × Baseline	–0.311	0.144
Random Effects		
Level 3–University	0.003	0.013
Level 2–Peer group	0	0
Level 1–Teacher	0.341	0.058
-2*Log-Likelihood	125.642	

Note: S.E. = standard error;

Of greater interest is the interaction term to determine the extent to which the intervention effect varied as a function of baseline ratings. A statistically significant negative interaction effect between feedback and baseline was found (–.31 *SD*, $t(68) = -2.16$, $p = .03$). This estimate of effect implies that among teachers with similar baseline, those with low baseline ratings scores in the intervention group improved in their teaching effectiveness more than teachers in the control group with similar ratings. In precise terms, the intervention effects were slightly larger for the initially less effective teachers.

As anticipated, the effect of 'baseline', was substantial and highly statistically significant (0.98 *SD*, $t(68) = 8.65$, $p < .0001$). Adding the three explanatory variables accounted for 67% of the total variance. Only about 39% (.563 –.341)/.563) of the variance at the teacher level was explained but virtually all the variability attributable to universities, approximately 98%, was explained. At the department level, the variance completely disappeared and was estimated to be exactly 0. This is natural and happens— Snijders and Bosker (1999) explained—when the difference between group means is less than would be expected on the basis of within-group variability.

A comparison of deviances reveals that Model 2 resulted in a statistically significantly reduction in the deviance by 69.08 from that obtained with the null model. It is apparent that Model 2 is appropriate and fits the data better:

$$\begin{aligned} \text{Model fit} &= \text{Deviance Model 1} - \text{Deviance Model 2} \\ &= 194.725 - 125.642 \\ \chi^2\,(3) &= 69.083,\ p < 0.0001 \end{aligned}$$

To determine whether there was a difference in teaching effectiveness between the institutions, another model was estimated to include a 'type' dummy variable

representing university type and size. No statistically significant difference in teaching effectiveness was found (–.30 *SD*, $t(1) = -1.81$, $p = .32$). Adding the 'type' variable hardly changed the estimates of effects for the other explanatory variables or their significance as found in Model 2. This could also be seen from the proportion of variance explained, 68%, compared with 67%, found in model 2. There is no evidence to suggest that teaching effectiveness is different in the institutions.

To summarise, on average, the effects of feedback and consultation in peer support groups is positively related to improvement in teaching effectiveness that can be generalised across all teachers but the benefits seem to be stronger for the initially less effective teachers.

What Size of Effect?

Is the size of effect for the feedback intervention large enough to be important? Effect size is used to test the relative efficacy of an intervention or its practical significance. Coe (2001) emphasised that it is the practical significance rather than statistical significance that is important to teachers and administrators in education. In fact, Carver (1978) described the emphasis on statistical significance as a "corrupt form of scientific method", pointing out that a statistical significant difference is not necessarily important. The concern should therefore be how much difference an intervention makes and whether the size of effect is large enough to be of some importance or small enough to be ignored.

The effect sizes found for the intervention in the present study could be judged as "small" by Cohen's (1977, 1988, 1992) standards (see Chapter 5), and in relation to that usually found for individual consultation. The 'main' effect estimate found with

this group-based intervention mirrors that (mean effect = .20) found by Cohen (1980) when teachers receive ratings feedback alone without consultation. Yet, in the context of a new research topic, effect sizes of .24 for a 'main' effect and .32 for interaction effect may be regarded as quite good and certainly large enough to merit consideration of the feedback intervention as a promising teaching improvement strategy.

Any low-cost intervention that may increase the acceptance and use of student ratings feedback by a large number of teachers, and results in even a modest positive effect on teaching effectiveness is certainly useful to universities. Such an intervention may represent a worthwhile investment for further research, refinement and development.

Effect of Teacher Characteristics

Model 3 estimated the effects of teacher gender, age, and experience on teaching effectiveness. Table 7.8 reports the parameter estimates and their standard errors. Not surprisingly, the estimated coefficients for teacher characteristics are very small, in fact with near-zero effects, with one exception. This indicates that teacher characteristics are weakly related to teaching effectiveness. These coefficients were not expected to be large or even moderate to support the literature of a weak relationship between teacher background characteristics and student ratings.

Table 7.8
Multilevel Results Estimating Effect of Teacher Characteristics

	Model 3	
Fixed Effects	Parameter	S.E.
Intercept	–0.169	0.130
Baseline	0.979	0.112
Feedback	0.272	0.139
Feedback × Baseline	–0.316	0.148
Gender (Female)	0.066	0.141
Age	–0.173	0.078
Experience	0.035	0.083
Random Effects		
Level 3–University	0.003	0.011
Level 2–Peer Groups	0	0
Level 1–Teachers	0.316	0.054
-2*Log-Likelihood	120.167	

Note: S.E. = standard error;

Of the three effects, only the 'age' effect is shown to have slightly more influence on teaching effectiveness, which is different from what may be expected by mere chance when gender and teaching experience were taken into consideration. The coefficient was statistically significant and negative (–0.17 *SD*, $t(65) = -2.22$, $p = .03$), suggesting that student ratings of teaching effectiveness decreased with age. Younger teachers were slightly more likely to be rated higher in teaching effectiveness than older teachers.

Feldman (1983) and Piccinin and Moore (2002) also found that teaching effectiveness was negatively associated with teacher age. In reflecting on this finding Feldman (1983) pointed to these teachers' deep-rooted beliefs and values, and their conception of teaching. Piccinin and Moore (2002) suggested that older teachers may not readily admit to difficulties in teaching and may be more resistant to embracing change in methods and techniques.

Adding age, gender, and teaching experience to the model did not change in any marked way the estimated average feedback effect but it came close to statistical significance at the 5% level (.27 *SD*, $t(20) = 1.96$, $p = .06$). The interaction coefficient remained virtually unchanged from that in Model 2 (.32 *SD*, $t(65) = -2.14$, $p = .04$). Controlling for the teacher characteristics does not appear to have an important influence on the effects of the feedback intervention.

Fitting model 2 to the data explained 69% of the total variance. Comparing the random effects of model 2 and model 3, it is readily seen that controlling for possible differences due to the teacher characteristics explains only about 7% (.341-.316/.341) of the variation between teachers, indicating that the model did not explain much additional variance in teaching effectiveness.

With teacher age, gender, and teaching experience accounting for so little of the variation, it can only be assumed that there are other important characteristics that were not captured in this data set—perhaps, enthusiasm, reputation, or even conception of teaching—that would best explain the difference in teaching effectiveness between teachers. That teacher characteristics do not add significance to the model is confirmed by the lack of improvement in model fit:

$$\begin{aligned} \text{Model fit} &= \text{Deviance Model 2} - \text{Deviance Model 3} \\ &= 125.642 - 120.167 \end{aligned}$$

$$\chi^2(3) = 5.475, p < 0.14$$

Effect of Course Characteristics

Model 4 models the effects due to course-level and discipline on teaching effectiveness. From the results in Table 7.9 it is obvious that the effects are mostly small and not statistically significant. This finding is not surprising because the ratings feedback research literature does indicate that these variables are not very influential on teaching effectiveness (e.g. Watchel, 1998). With four categories for course levels, 3 dummy variables were created with year 3 serving as reference because it had the lowest mean rating. The results in Table 7.9 show there is no real difference in teaching effectiveness for teachers who offer instruction at the year levels 1, 2, and 4 compared to instructing at year level 3. There is no evidence that teaching effectiveness is related to the level at which teaching occurs.

For the disciplines, business (lowest mean rating) served as the reference category. The results indicated that the difference between business and the applied sciences was statistically significant with teachers in business appearing to have improved somewhat more in their teaching effectiveness than those in the applied sciences (–.33 *SD*, $t = -2.19$, $p = .03$). The student ratings literature does show that of all the disciplines, the applied sciences tend to receive the lowest student ratings. Differences with teachers in humanities and natural sciences seemed very likely to have occurred by chance. Differences were not statistically significant. Overall, in this data set there is no strong evidence to suggest that teaching effectiveness is different for the disciplinary areas in which instruction is offered.

Table 7.9
Multilevel Results Estimating Effect of Course Characteristics

	Model 4	
Fixed Effects	Parameter	S.E.
Intercept	–0.317	0.203
Baseline	0.931	0.112
Feedback	0.297	0.140
Feedback × Baseline	–0.299	0.140
First [a]	–0.058	0.144
Second [a]	0.146	0.146
Fourth [a]	0.162	0.140
Applied Science [b]	–0.330	0.151
Humanities [b]	0.093	0.149
Natural Science [b]	0.166	0.148
Random Effects		
Level 3–University	0	0
Level 2–Peer groups	0	0
Level 1– Teachers	0.313	0.053
-2*Log-Likelihood	118.958	

Note: Reference category: [a] Third Year Students [b] Business
S.E. = standard error;

Once the effects of course level and discipline were controlled for, the effect size for 'feedback' increased marginally from that found in Model 2 and even gained statistical significant at the 5% level (.30 *SD*, $t(20) = 2.12$, $p = .05$). The coefficient and statistical significance of the feedback-by-baseline interaction remained essentially unchanged (–.30 *SD*, $t(62) = -2.14$, $p = .04$).

This model, adjusting for possible differences due to course characteristics accounted for about 70% of the total variance in post feedback ratings, but this variation is not appreciably different from 67% observed in Model 2. At the teacher-level the total unexplained variability goes down from .341 (model 2) to .313 (Model 4), indicating that course characteristics explained only about 8% of the variation between teachers. Clearly, effects of the course characteristics fitted are no much stronger than teacher characteristics in explaining differences in teaching effectiveness.

It may be concluded on the basis of this data that teaching effectiveness is not highly related to course and teacher characteristics. With much of the difference between teachers remaining unexplained it is clear that teaching effectiveness is more likely due to more important factors, which may include teaching skills and subject-matter knowledge. Adding course characteristics does not improve the fit of the model to the data:

$$\begin{aligned} \text{Model fit} = & \text{ Deviance Model 2} - \text{Deviance Model 4} \\ & = 125.642 - 118.958 \end{aligned}$$

$$\chi^2\,(6) = 6.684,\ p < 0.35$$

COMPARISON OF RESULTS FROM OLS REGRESSION AND MULTILEVEL MODELLING

In this comparative analysis, only the estimates of effects and their standard errors, and the variance explained as obtained for the final models estimated with ordinary regression and Model 2 in the multilevel regression are considered. The relevant statistics from each analysis are presented in Table 7.10. Although identical models were estimated none of the estimated coefficients fully matched but the estimates of effects from the multilevel analysis are quite similar to those from the teacher-level analysis and markedly different to those found with the aggregated data in the department-level analysis.

To illustrate, looking at Table 7.10 it can be seen that the estimated coefficients for 'feedback' in the teacher-level analysis (0.241) and the multilevel analysis (0.241) are identical but that for the cluster-level analysis (0.380) is quite different. The

standard errors for the multilevel and teacher level are also virtually identical while those for the department-level are almost twice as large. The multilevel model analysis accounted for just about the same amount of variance as the teacher-level OLS regression analysis. As Raudenbush and Bryk (2002) explained, the estimates from multilevel modelling will be close to the teacher-level analyses.

Table 7.10
Comparison of Multilevel with OLS Regression Results for Effect on Teaching Effectiveness

	Multilevel Analysis		***Teacher-level Analysis***		***Department-level Analysis***	
	Coefficient	*S.E.*	*Coefficient*	*S.E.*	*Coefficient*	*S.E.*
Baseline	0.977	0.113	0.991	0.115	1.140	0.187
Feedback	0.241	0.140	0.241	0.144	0.380	0.223
Feedback × Baseline	–0.311	0.144	–0.320	0.148	–0.476	0.248
Variance explained	0.666		0.652		0.750	

Note: S.E. = standard error;

In spite of the difference in the estimated coefficients, all three analyses do seem to support the conclusion that the intervention effects depend on the initial ratings status, and that the less effective teachers may have received slightly more benefits from the feedback intervention, considering the strengths and weaknesses of each type of analysis. This convergence of evidence is important as it enhances the validity of findings. The average feedback effect came close to statistical significance at the 10% level in all three analyses.

Variations in the results stem from the differences in the sample sizes used in the analyses. The teacher-level analysis and MLM both use the data for individual teachers but the department-level analysis pooled this data to the group level. This results in a reduced sample size and reduced statistical power. Statistical power is further reduced

by the presence of a positive intraclass correlation that inflates the variance of response—in spite of the use of simple randomisation—and by the variable group sizes, indicating that the means for the groups were not equally reliable.

This unreliability in the group means can result in extreme values from small groups strongly influencing estimates, leading to low precision and regression coefficients with large standard errors. The inverse variance weighting employed in MLM protects against any instability associated with unequal cluster sample sizes. MLM will also produce more efficient estimates because it is relatively unaffected by the degree of grouping in the data (Kreft & de Leeuw; 1998; Aitkin & Longford, 1986; Raudenbush & Bryk, 2002).

Looking at the statistics presented in Table 7.10 it does seem that there is good agreement between the regression coefficients estimated and their standard errors by the teacher-level OLS regression and MLM, so in this case the OLS estimates do not appear to be particularly biased and there is no difference in substantive interpretation of intervention effects. Nonetheless, the teacher-level OLS regression results could not be used to draw firm conclusions about the intervention effects. The appropriate evaluation of intervention effects would have to be on the basis of department units (peer-groups) because the intervention was administered at that level.

Looking once again at the results in Table 7.10 and comparing estimates by the department-level analysis and MLM, it does seem that analysis of the aggregated data may have resulted in misleading estimates. This would have resulted in a quite different interpretation of the intervention effects. Furthermore, as previously discussed, basing conclusions on aggregated analysis regarding teacher effectiveness would be committing an "ecological fallacy". Therefore, on the knowledge that MLM takes

account of intercorrelations among variables and so tend to produce more precise estimates, the estimated coefficients obtained by the multilevel models are expected to be the more accurate assessment of intervention effects in the present study.

On the other hand, satisfactory analysis with MLM requires a large data set. In educational research the common problem is that of small sample sizes. It is almost impossible to meet the requirement of 25 groups with at least 25 individuals in each, as recommended, for example, by Paterson and

Goldstein (1991). While this may raise questions about the reliability of estimates found for the present study, looking at the bigger picture this requirement will certainly limit the application of multilevel modelling to ratings feedback intervention studies, with implications for the accurate representation of feedback intervention effects.

Summary of Quantitative Results

The evidence presented in this chapter clearly indicates that augmenting feedback from student ratings with consultation in peer support groups is significantly associated with improved teaching for teachers at all levels of teaching effectiveness. The effects, however, seemed stronger for the less effective teachers and this did not result from a ceiling effect. The convergence of evidence from the ordinary regression and multilevel modelling validates the conclusion about the effects of the feedback intervention.

The results also show that there is little or no effect of teacher age, gender, teaching experience, their discipline, and the level at which instruction is offered on teaching effectiveness, supporting the literature. Although the effect size found for the feedback intervention is not considerably large, in the context of research with an

innovative strategy, this effect size is good and encouraging to facilitate further research.

The application of multilevel modelling was also demonstrated. The results showed that aggregating the data to the level of department units, the unit of randomisation, would have exaggerated the estimates of intervention effects with implications for interpretation. The multilevel modelling approach produced less biased estimates than the OLS regression, properly accounted for the nested nature of the data, and provided good insights into the variation in teaching effectiveness.

The MLM results showed the majority of variance in teaching effectiveness is attributable to teachers themselves but that departments and the institution also have some influence on teaching effectiveness. Because the sample size is small, however, estimates of effects might not be reliable so the results must be interpreted with care.

Chapter 8

Teacher Perceptions of Consultation in Peer Support Groups

This chapter reports on how teachers experienced the intervention and their perceptions on the efficacy of consulting over ratings feedback in peer support groups. It presents the qualitative analysis of the interview and is useful to gain a better understanding of the results from the quantitative data. The chapter is organised according to the three themes that emerged from the data: (a) talking with colleagues (b) impact of ratings feedback, and (c) conditions for use of peer support groups.

Introduction

As already mentioned in Chapter 6 individual semi-structured interviews were conducted among 35 of the 40 teachers in the intervention group. The interviews involved all 16 teachers who did not meet in the peer support groups and 19 of the 24

who met with their peer partners at least once whether or not end-of-semester ratings were collected for these teachers. Questions concentrated on how they had benefited from talking over ratings feedback with colleagues, whether they were able to make any changes to their teaching, and strengths and weaknesses of the intervention.

Written rather than audio taped notes were taken during the interviews as already mentioned in Chapter 6. These notes were supplemented by observation data and field notes as the researcher acted as facilitator of the group meetings and was able to make notes regarding the functioning of the groups but this was not systematic. Observation data and field notes provided important additional information that would not have been ascertained from the interview notes alone. Once again, it is important to note that these responses should be interpreted with care as many teachers volunteered to participate in the study simply to "help the researcher".

Talking with Colleagues

Teachers were more or less willing to talk about their teaching with their colleagues but for the most part dialogue in the groups was more in terms of students, than on any deficiency on the part of teachers themselves. The novelty of the approach may have played a role here. Talking openly about one's teaching with colleagues is not normal practice for these teachers. It is reasonable to expect that with sustained dialogue over time the talk would shift to teaching and its influence on learning. Although implied, at no time was reference explicitly made to improving the quality of student learning.

It was evident that lack of social cohesion and collegiality in one department unit may have influenced participation in the group process. This was the only peer

group in which all its members could not seem to find the time to meet. As another indicator, the typical response to the question on willingness to talk with colleagues was "it depends on who is in the group". As would be expected success of this model of consultation greatly depends on the extent to which teachers feel there is respect, confidentiality, trust, and feel safe enough to open up about their teaching. When asked whether the process had enhanced relations with colleagues, the general response was that it was too soon to tell.

Two main issues were explored in the interviews under this theme: (a) effect, if any, of talking over ratings in a structured group, and (b) factors that inhibited meeting peers in a group setting.

Talking over Student Ratings

The 19 teachers who met with their peer partners contributed responses for this section. The tendency was to focus on the written comments made by students rather than on the ratings form items. Baxter (1991) and Tiberius, et al. (1989) also found that teachers were more likely to respond to students' open-ended comments than the ratings on items. The main benefits from use of peer support groups in this context came in the form of opportunity to reflect, discovery, and motivation to make changes to practice.

Reflection

Motivation to "look more closely" was the predominant comment on benefits received from talking with colleagues in a group setting. Several teachers commented that they became 'aware' which led them to pay more attention to the ratings:

> *Yes, the meeting made me aware of more things to take into consideration and motivated me to act on it [the ratings].*

In a sense it prompted me to look closely at my teaching methods.

It [meeting with peers] propelled me to try new methods ... and realise the significance of ratings.

Discovery

For many teachers the collaborative learning arrangement was a process of discovery or becoming alert to certain aspects of practice. The comment was that they had come to realise that certain views from students were not unique to them. They were somewhat surprised to learn that students were saying similar things about their colleagues:

It was rather interesting to see how the responses are common across lecturers. I have come to realise that there could be some objectivity in ratings .... I am now more appreciative of the results.

When one teacher who had the reputation as a good teacher shared students' written comments, one peer partner could be heard saying, "I never knew they say that about you too".

A focal point in the discussion among teachers in the different groups was their tendency to rate courses as being more difficult or as having a heavier workload than did students. In this case, the feedback differed markedly from students' frequent protests in classes. On this basis many teachers questioned whether students really took the exercise seriously or understood what they were doing for that matter. On the other hand, speaking in the context of the discipline, teachers were able to express some understanding for the ratings students made in certain respects. The group process was also highly valued by the less experienced staff members. One junior teacher commented:

As this is my first teaching assignment, interacting with the more experienced teachers and those with large classes proved rather interesting and enlightening. At least I'll have some ideas when my time comes.

Another benefit for these teachers, drawing on the observation data, is the opportunity to actually examine the ratings in some detail. Although ratings were returned one week before the first group meeting was scheduled, it was observed that many teachers used the meeting time to examine the ratings, making comments as they perused them. Also from observation, was the opportunity for debate on certain aspects of process but this was limited in the groups. For example, when one teacher commented that the ratings form should not ask students to comment on the pace of the class, the other teachers were able to challenge this view.

Changes to Teaching

The opportunity for dialogue and discussion with colleagues might have had some impact on practice as many teachers indicated that the process led them to modify certain aspects of their teaching. The comment was that they took action on the feedback ratings that they probably would not have taken if they were working alone.

To illustrate, when one teacher shared students' complaints about being called "dunce", the peer partners were able to exchange experiences and ideas on how they deal with similar issues. In the interview the teacher delightfully indicated that the earlier approach was revised and reported improved relationship with students. At the same time, a few teachers were critical that the intervention was not tailored to their specific needs. Some felt that the better approach would be to provide individual consultation first and then move to the group level.

The value of the peer group interaction could also be assessed from responses on the elements of the intervention that were most valuable. Basically all the teachers who participated in the group process identified "meeting with colleagues" as one of the most helpful aspects and a strength of the feedback intervention. They seemed to have appreciated the opportunity to interact with their peers because, in the words of one teacher, "… the opportunity to reflect is lacking otherwise".

Factors Affecting Participation

When asked about the factors that affected their ability to meet peers in a group setting, without exception all respondents, those who had not met with their peer partners at all, and those who met only once, singled out the lack of time. This might be understood against the background that teachers were not given special time off or other incentives to participate in the study. In this sense, the study has some ecological validity.

The view that time was a constraining factor was expressed by references to the lack of common time for partners to meet and the lack of time due to heavy teaching loads along with administrative duties.

Lack of common time

Even though teachers were in the same building or shared the same office space it proved rather difficult for group members to meet because they had such different work schedules:

> *With different class schedules it is not easy to find a time for group members to meet during the day.*
>
> *There's just no common time to meet.*

There were no clear indications that lack of engagement in the process resulted from lack of motivation but the events of the semester and the busyness associated with first semesters probably magnified the sense of not having enough time:

> *I really wanted to do this but as you can see this semester has not been a good one.*

The reference was to the interruption to classes due to the rain associated with the tropical storms and disturbances associated with the general election as mentioned in Chapter 6.

Lack of unpressured time

Academics in Jamaica might not be as actively engaged in the research function as their counterparts in say North America and Europe but these academics are equally pressured by the demands associated with their teaching and service functions. Some academics teach as many as four different courses per semester in addition to institutional and community service obligations. It is not uncommon for a teacher's workday to extend from 8 am to 9 pm, Monday to Friday, and 9 am to 5 pm on Saturdays. This is necessary to facilitate part-time and evening programmes, and because of limited classroom space. For senior lecturers lighter teaching loads mean more administrative duties.

Many teachers expressed feeling overwhelmed with academic and administrative duties that left them with little time to share with colleagues:

> *This place is nothing but work. …You can't even find time for yourself … it's just too much … _____ [name of teacher], had so many meetings we were not able to meet again as planned.*

> *I just couldn't find the time as hoped with having to prepare documents for____ [a quality assurance exercise]. It's just crazy around here"*

Worryingly, the pressure of teaching and other academic obligations is such that academics do not seem to have time to attend to their own professional development.

In summary, data for this theme seem to suggest that teachers are willing to talk with their colleagues about their student ratings but may be constrained by the conditions of their work. Those who actually participated in the group process found the experience interesting and indicated that they made changes to their teaching because of the group interaction.

Impact of Ratings Feedback on Teaching

Thirty-five of 40 teachers in the intervention group provided interview data relating to this theme. Overall, teachers believed that receiving mid-term ratings feedback actually helped to improve their teaching as they were able to make changes to their teaching but only "minor" ones. The general comment was that the time was just too short to make changes. The ratings information would be used in the following semester.

Impact on Teachers

For many teachers receiving early feedback on their teaching and having to examine it in some detail not only made them aware of student views but also provided the opportunity to reflect on their teaching and their role as teachers:

> *I became aware of students' views relating to my teaching. This prompted me to look more closely at my delivery and myself, contemplate thorough preparation … and take teaching more seriously.*

On reading the report I became more conscious of my role as an instructor and the importance of student views. Yes, I became more aware.

The feedback was important because I became conscious of things that students pointed out.

The results somehow caused me to reflect on my performance. It may just be that teaching was taken for granted before.

The concern of many teachers was the gap that existed between students' ratings and their own ratings of teaching. The use of bar graphs to compare the ratings from students and self-ratings seemed particularly powerful. Awareness of this discrepancy may have actually encouraged many teachers to attend to the ratings, reflect on their teaching, and take action to modify certain areas of classroom practices:

> *Just knowing how students think, … gives you an opportunity to assess yourself, especially when you figure all along you are doing the right thing, and finding that students think otherwise.*

Further evidence of the role of self-ratings is provided from the fact that 46% of teachers interviewed identified self-ratings and the comparison between student and teacher self-ratings as one of a most valuable aspect of the intervention. Teachers were generally appreciative of the form in which the ratings results were reported.

Only two references were made to the psychological effect of ratings. One teacher commented:

> *In a way I am angry, for I try so much to get students involved and to have them complain … it's just so frustrating.*

Another expressed rather dispiritedly that,

> *Students just want you to spoon feed them and when you don't they say all sorts of things.*

Impact on Teaching Practice

Exactly 84% of the teachers interviewed indicated that they had modified their teaching style, and pace of delivery on the basis of the ratings feedback received. This indicates that both teachers who met with and did not meet with colleagues in the groups apparently used the ratings feedback to effect changes in teaching. A small number felt that complaints from students were out of their control. Teachers seemed to have liked the idea of receiving mid-semester ratings but commented that the time in which to make changes was just too short.

The most frequently cited impact on classroom practices was in the form of (a) changes in teaching strategies and techniques, (b) and attempts to improve interpersonal relationship with students. For example, one teacher reported:

> *As a result of students' views about workload, a planned in-class test was changed to a take home exercise. Students were appreciative of this.*

For another:

> *I make myself more available to students. I now have an open door policy where students can come and see me if they see the door open.*

Yet another teacher recalled how she adopted ideas from the TIPS sheets in response to students' complaints that the class was boring. Students responded positively to the changes. For a few teachers adjustments were made to the pace of instruction, and attempts made to increase student participation in classes.

In the context of the limited intervention time, the set of teaching ideas was only marginally helpful. It was generally felt that it was too much material to read. Only a

few teachers indicated that they actually made immediate use of the teaching tips. For sure, teachers were appreciative for the set of tips but noted that they would "be of benefit for next semester".

There is reason to believe that that interest in the ratings feedback increased because the ratings questionnaire had at least face validity among teachers. The questionnaire was seen as comprehensive and being able to provide good information on teaching strengths and weaknesses. The presentation of the ratings also came in for special mention.

Training for Students

One interesting finding from the interview data was the general concern that students should be trained in how to rate their teachers. It cannot be said that these teachers were being critical of their ratings or the questionnaire, which was identified as a strong point of the intervention. They simply felt that students might not have taken the process seriously, or exercised care in making ratings because of lack of training:

> *I'm not sure whether students know what they are doing when they respond to ratings forms. There is need to educate students on how to rate their teachers.*
>
> *I'm just not sure that students understood the questions. Training should be provided on how to answer ratings questionnaires.*

It is possible that this conclusion resulted from discussing the ratings results with students. Teachers were encouraged to close the 'feedback loop' by sharing the results with students to gain a better understanding of the views. A few teachers indicated that changes were negotiated with students. There was also the suggestion teachers themselves should be trained in how to rate their teaching.

In sum, this category of responses suggest that teachers will use ratings feedback to inform their teaching practice if ratings are timely, reports are informative, and they have adequate time in which to implement and evaluate changes to teaching.

Conditions for Use of Peer Support Groups

A key interest of this study was to explore teachers' views on the conditions under which a group-based peer model of consultation could be effectively implemented as a teacher professional development activity. To the extent that teachers were positive in their views about learning from ratings feedback in peer support groups indicates that they do value collaborative arrangements to improve their teaching if given the chance. Nonetheless, they view that successful use of this model depended on: (a) availability of time for colleagues to meet and talk, (b) support from administrators, (c) provision of good feedback information.

Sufficient Learning Time

From the responses of teachers, it is clear that a key condition that should be satisfied to use group-based peer consultation successfully, is to make available an adequate amount of time to facilitate dialogue in the groups and to have regular meetings. Teachers also saw it as necessary to have sufficient time to learn from the ratings to be to able to make, implement, and evaluate changes to teaching. Against the background that the ratings were collected later than expected, coupled with the time used for processing, the overwhelming observation was that ratings should be provided early in the semester.

Support from Administrators

All interview respondents firmly expressed the model could only work if it was supported by administrators, including course and programme leaders, in terms of scheduling, co-ordinating, providing good feedback information, and providing the time and resources required to facilitate the process:

> *For this to work it would have to be a scheduled activity. When left up to colleagues it might not happen.*
>
> *If lecturers are left on their own it is unlikely that discussions will take place.*

Consistent with the literature, teachers drew attention to the role of department leaders in facilitating teacher collaboration for teaching improvement. Teachers' belief that they required support from administrators may stem from the variable work schedules and time demands that make it not too easy for teachers to meet unless space and time has been created for the activity.

Feedback Information

The other important condition that was seen as necessary for the use of this group-based model was for the provision of adequate and timely feedback information. Making reference to the detailed nature and form of presentation of the ratings report, the questionnaire, and the other materials provided as part of the present study, teachers felt that they needed accurate and good quality feedback information to work with.

In sum, the implementation and effective use of peer support groups as a consultation process, as far as teachers are concerned, depend on the extent to which administrators are willing to facilitate and support the process of allowing them to learn in a collaborative way with disciplinary peers.

Summary of Qualitative Results

The results of this qualitative analysis data show that teachers in this present study generally perceive that consultation over student ratings in peer support groups is practical and appropriate to improve teaching effectiveness. Generally speaking, teachers seemed comfortable to open the doors on their teaching and talk with their peers over the student ratings feedback, albeit they were more willing to disclose the written comments than ratings on questionnaire items. Overall, teachers welcomed the opportunity to talk about teaching, reflect, exchange experiences and ideas, and tell their stories about teaching from the perspective of their disciplinary areas. All teachers declared that the intervention was worth their time and effort.

Teachers spoke about how the group process provided the opportunity to reflect on their teaching. Teachers also expressed that they modified certain aspects of their teaching on the basis of the ratings feedback received and the interaction in the peer groups. This is more than what probably would have happened if they were working alone, in isolation. Teachers recognised the limitation of the short duration of the intervention, indicating that it only allowed for minor changes to be made to teaching.

There are, however, factors related to the working environment that seems to work together to make consultation in peer support groups that more difficult and challenging. The notion of "no time", variable work schedules, the pressure of heavy teaching loads, coupled with administrative duties, and the very structure of departments themselves, presented clear threats to teachers' enthusiasm and willingness to collaborate on teaching improvement.

Participation in the intervention, like many staff development programmes, was voluntary. As such, it represented an additional activity to already full schedules that

were not compatible with their colleagues to begin with. Against this background of lack of time and workload issues teachers had little 'free' time to engage in a voluntary activity. Although teachers may have become aware of their own responsibility for teaching effectiveness, they evidently felt that without administrators supporting and facilitating the process, this model of consultation was unlikely to be sustainable or be effective for that matter.

Chapter 9

Discussion

In this final chapter some of the more important results of the present study are considered. It begins with a brief overview of the problem of the study. The chapter then presents a summary of key findings, which is followed by a more detailed discussion of the main finding of the research and implications for practice. The chapter concludes with some directions for future research. As the present study probably represents the first of its kind, it should be recognised that this is a tentative discussion on the potential of group-based peer consultation as an alternative model to individual consultation.

Overview of Study

An established finding in ratings feedback research is that merely providing teachers with feedback from student ratings has only a modest effect on teaching effectiveness. The effect is substantial with support from consultation. The traditional, individual consultation model is effective in improving teaching effectiveness but is

now arguably much too costly for universities to afford. Consequently, teachers are left on their own, without support and guidance on how to use ratings feedback effectively. And yet, the use of student ratings as an indicator of teaching quality is becoming more crucial in universities.

The present study is a response to the lack of support to university teachers to use ratings feedback to improve their teaching with a relatively cost-effective strategy. The purpose of the study was to investigate the effect on teaching effectiveness of augmenting ratings feedback with consultation in peer support groups, a group-based consultation model. A randomised controlled trial was conducted among a sample of 71 teachers in two universities. The data was analysed with both OLS regression and multilevel modelling. Teachers' experiences and perceptions on the efficacy of the intervention were also explored through semi-structured interviews.

Summary of Findings

Several issues of one kind or another have been discussed in Chapters 7 and 8 of this thesis. A summary of some of the key findings of the study are as follows:

- The SEEQ is a valid and reliable instrument to assess teaching effectiveness in the Jamaican context.
- Consulting over ratings feedback in peer support groups is associated with improvement in teaching effectiveness (*ES* = .24). The effects of the feedback intervention are slightly stronger for the initially less effect effective teachers (*ES* = .32).
- Results of OLS regression and multilevel modelling results converge upon the effectiveness of group-based peer consultation to improve teaching effectiveness.

- Teacher and course characteristics namely, gender, experience, course level, and discipline, have relatively little effect on student ratings of teaching effectiveness.
- There is a small inverse relationship between teaching effectiveness and teacher age.
- Teachers are willing to consult over their ratings feedback in groups with colleagues if given the opportunity of learning space and time for dialogue and reflection.
- Teachers perceive group-based peer consultation is a practical strategy to improve teaching effectiveness but feel that effective use will depend on:

  (a) availability of adequate time and space for talk with colleagues, (b) level of support from administrators in terms of department structures and processes, and (c) provision of quality feedback information.
- Collaborative learning from ratings feedback is constrained by teaching workload, and other conditions of work.

Contribution of the Meta-Analysis

The findings of the meta-analytic review (see Chapter 5) represent an important contribution to the literature on the practices and strategies that may be expected to work for the most successful consultative feedback. Prior to this, the consultant might have been seen as the critical element in the consultation process. For example, Cohen (1991) suggested that effectiveness of consultative feedback resulted from the informational and emotional support the consultant provided. Instead, the results of the meta-analysis indicate that the consultant is only one of several elements important for effective consultation over student ratings. It now seems to be important to focus on the content, design, implementation, and quality of interaction in the consultation process.

It is these elements that underpin the association with improvement in teaching effectiveness.

Acknowledging the limitations of the meta-analysis, the conclusions must only be considered as tentative. Yet, the findings represent the best available knowledge at the present time that can be used to guide policy and practice. The strategies and practices that may be useful to maximise the effects of consultation over student ratings include:

- Opportunity to talk about one's teaching
- Active involvement of the teacher in the consultation process rather than passively receiving information
- Quality feedback information gathered with a well-developed instrument and from multiple sources
- Opportunity to interact with peers
- Provision of adequate learning time in the consultation process
- Comparison of teacher self-evaluation with student ratings

It also happens that similar features were identified from the experimental portion of the present research with group-based peer consultation. This clearly suggests that the strategies used in individual consultation may not be very different for group-based consultation.

The implication of the results of the meta-analysis is that, in providing opportunities to support teachers in learning from student ratings, as many of these features as possible should be included to optimise the effect on teaching effectiveness. From this meta-analysis, then, it seems reasonable to suggest that a less costly, yet

effective consultation process, such as group-based peer consultation can be organised to support teachers to learn to use ratings feedback to improve their teaching.

Discussion

Ratings Feedback on Teaching Effectiveness

The result of the present study is that feedback from student ratings augmented by consultation in peer support groups is significantly associated with improvement in teaching effectiveness. The benefits of the feedback intervention, however, are slightly stronger for the initially less effective teachers. This finding for group-based peer consultation is consistent with research by McKeachie, et al. (1980), Piccinin, Cristi and McCoy (1999), and Marsh and Roche (1993) for individual, one-to-one consultation.

The importance of the finding that the less effective teachers may have benefited more from the feedback intervention should not be taken too lightly. As Coe (1998) has pointed out, contrary to popular belief, feedback may not always increase performance but may actually be more detrimental. For student ratings feedback, Pambookian (1974) and Litzleman, et al. (1998) have shown that negative feedback ratings to teachers, who are already performing poorly, without appropriate support, can lead to frustration and further erode teaching performance.

Roche and Marsh (2000) expound that teachers who receive poor ratings may either recognise the need for change and seek help, dismiss the ratings as invalid, or reallocate their effort towards other interests such as research or administration. Even worse, they may resort to tactics such as lenient grading and reducing workload, in order to receive favourable student ratings. To the extent that this might be happening has been identified as one reason for the dumbing down of the curriculum and driving

down of academic standards (Trout, 2000; Sproule, 2000). It can be understood then, lack of support with student ratings feedback holds strong negative consequences for the quality of university teaching.

On the face of it, it may seem appropriate to suggest that staff developers target poorly performing teachers for consultation, helping them to become more effective, except that this option is problematic. For one thing, it is a mistake to think that only the less effective teachers have problems with teaching (Fullan & Hargreaves, 1992). The other thing is that, these teachers are already embarrassed to know that their teaching is not going well (Wragg, et al. 2000; Stanley, Porter & Szabó, 1997). Any intervention that focuses directly on these teachers is likely to have a devastating effect on teaching self-confidence and self-esteem and probably reduce teaching performance even further.

This issue can be understood in relation to self-efficacy theory, where teachers who may have received low student ratings doubt their capability to facilitate student learning (Bandura, 1982, 1993). With particular reference to ratings feedback Roche and Marsh (2000) referred to this phenomenon as 'teacher self-concept'. That is to say, the perception teachers hold about their teaching effectiveness. As discussed in Chapter 5, Roche and Marsh (2000) found that teachers adjust their self-perception upwards or downwards when they receive ratings feedback. Teacher self-concept has a strong influence on motivation and effort to bring about change and improve teaching practice.

But colleagues can support a positive teacher self-concept. Convery's (2001) experience may be used as an example. He recalled, "Being valued by my critical friends as a teacher-researcher rather than as a failing teacher validated my changing

self-image and provided both encouragement and direction" (p. 140). For Convery support from colleagues in examining teaching meant that those areas needing improvement could be identified rather than avoided. This lends some support to the value of group-based peer consultation as a strategy for improving teaching for all teachers, provided it is implemented properly. The value of this model of consultation is underpinned by the opportunity for teachers to engage in interactions with a mix of colleagues—new and old, effective and less effective, senior and junior—in their departments. At the same time, there is the opportunity for mutual support, dialogue, and generation of improvement strategies in a collective manner.

Indeed, as the present study has shown, because teachers had access to the knowledge and experiences of colleagues in their departments, they soon realised that their problems were similar. Consequently, they were reassured to know that they were not the only ones experiencing problems. As suggested by Elliott (1991), this makes it easier to tolerate any loss in self-esteem and teachers soon become more open to student feedback. Feeling more confident as teachers, they may then be more willing to change their teaching for improvement to come about.

Departmental and Institutional Influence

An interesting feature of the present study is the use of multilevel modelling which partitioned the variation in teaching effectiveness. Although hardly surprising to learn, the results indicate that apart from individual teachers themselves, there is an apparent influence of the department, confounded by discipline, on teaching effectiveness. This suggests that, merely targeting individual teachers may not be sufficient to improve the quality of university teaching, although this is relevant. Teaching within academic departments should also be targeted for improvement, which

does not come about by simply accounting for the teaching effectiveness of individual teachers (Elton, 1998).

This idea of targeting teaching within departments fits closely with the views that already exist in the related literature that there are clear differences in teaching and learning in the disciplines—in terms of goals, values, culture, philosophies, orientations, and styles—and that teachers identify with and have allegiances to their disciplines and to their areas of specialism. Learning how to improve teaching cannot therefore be separated from the context in which teaching takes place, that is, the department units (Becher, 1994; Smeby, 1996; Becher & Trowler, 2001; Knight, 2002a; Shulman, 2000). At the same time, improving teaching cannot be seen as *the* responsibility of the individual teacher. As noted by Knight (2002b), it is now generally expected that departments and institutions make a substantial contribution to teaching quality.

This perspective resonates with the notion of situated learning and communities of practice (Lave & Wenger, 1991; Wenger, 1998), in which learning about and improving practice would be undertaken as a collective activity by teachers within their departments, through professional communities that are nurtured by institutions. That means more than activities simply taking place within departments. The advantage of this approach is that in the short-term teaching improves for individual teachers and over time the quality of teaching within departments is likely to improve from the collaborative culture that has been built up.

Teacher responses seem to corroborate the literature by drawing attention to the significant role department leaders—and university administrations in general—play in leading teaching improvement (Ramsden, 1998; Lafferty & Fleming, 2000; Knight, 2002b). In discussing the influence of department leaders on teaching, Knight

and Trowler (2000) contended that the exhortation to teach better is an exercise in futility unless department cultures are conducive to better teaching. Likewise, Evans (2001) commented that despite government policy and rhetoric, it is administrators who can foster positive work attitudes by creating and sustaining work contexts that are conducive to effective teaching and learning. It is clear, department leaders have some responsibility to create the conditions, and opportunities that will allow teachers to experience collaboration and mutual support with student ratings feedback.

At the level of the institution, university administrations have a facilitative role rather than direct responsibility for teaching quality with far-reaching consequences. From this level, administrators are seen as 'drivers' and 'enablers' of a quality culture within institutions (Gordon, 2002). Quality improvement therefore has to be facilitated through developing, maintaining, and monitoring institutional policies and processes that are needed to provide the conditions and opportunities necessary for teachers to improve teaching.

Improving quality in teaching, according to Biggs (2001), "cannot be left to the sense of responsibility or to the priorities of individual teachers. The institution must provide the incentives and support structures for teachers to enhance their teaching [effectiveness]" (p. 229). In the context of student ratings this calls for university administrations to recognise the need to locate teaching evaluation within a comprehensive system that provides teachers with prompt and detailed ratings feedback with adequate support on how to use the information to improve their teaching (e.g. Arreola, 2000; Ory, 2000).

There is also the need to commit resources to validate ratings forms to provide teachers with the best available feedback information. Although there is the reality of scarcity of funds, implementing teaching evaluation systems is such an expensive venture that it makes good sense, to go an extra mile to ensure that ratings feedback is being used constructively to contribute to quality improvement in the institution. In any case, teachers generally believe that administrators can do what they *really* want to do. So, it may be more a matter of willingness and commitment to quality than the lack of funds.

An important point to highlight from the results of this study is that teachers need adequate time to become acquainted with ratings feedback, experiment with ideas and monitor changes before improvement can be expected. Many administrators it seems trivialise the complexity of teaching and learning, by wanting to "see" immediate improvement when teachers receive ratings feedback, even with some support. Teachers should be allowed enough time to work with student ratings before improvement can reasonably be expected. For this reason, Bernstein, Jonson and Smith's (2000) suggestion from research with a peer review group that institutions adopt alternating periods of improvement and evaluation, may also apply to the use of ratings feedback. This is particularly important if indeed the intent is improving teaching rather than just accountability.

As illustrated in this research, the fact that teachers were given the chance to work with ratings feedback solely for the purpose of teaching improvement may have given teachers the opportunity to concentrate on using the feedback information to improve their teaching. In practice, this could take the form of, say, a two-year period when ratings feedback would be protected from scrutiny for accountability purposes and

used only for teaching improvement, which could be brought forward during the accountability period.

The analogy is with external quality monitoring for teaching and research that is assessed once every few years. This offers institutions the opportunity to use the information from a review to improve its systems before they are assessed again. Used in this way, more teachers may come to view student ratings as an improvement tool, which is different from the present dominant view of student ratings as a managerial device.

Collaboration for Teaching Improvement

Analysis of the qualitative data provides some evidence that teachers are willing to collaborate on teaching improvement if given the opportunity of learning time and learning space. Despite the public nature of the feedback intervention, teachers felt comfortable enough to talk about their student ratings with colleagues; share their experiences; and exchange ideas on practices that they had successfully used. However, the opportunity for collaborative learning was limited by the nature of their work, and the very structure of their departments. This fits well with what Hargreaves (1993, in Clement & Vandenberghe, 2000) referred to as 'constrained individualism', which basically results because of constraints in the workplace that prevent collaboration.

Another thing that was made quite clear in this research, is noted in the literature but somehow ignored by university administrations is that, attending to student ratings feedback is not automatic. Ratings feedback is only more likely to be utilised within a support structure than when individual teachers are working alone in isolation, even with access to self-help resource materials.

As an academic working in the capacity of staff developer for many years, Piccinin (personal communication, April 10, 2003) is still amazed at how little use teachers actually make of ratings feedback. However, as demonstrated in the present study, this lack of use may not be directly related to lack of motivation, lack of commitment to effective teaching, or even lack of appreciation of the value of student ratings. Rather, the problem may be attributed to the conditions under which teachers work and the problem of time.

The literature on the work of university teachers (McInnis, 2000; Taylor, 1999; Lafferty & Fleming, 2000; Light & Cox, 1999) does point to the intensification of academic work that is changing the work roles and workloads of academics in dramatic ways. Not least of which is the idea of 'accounting' for everything and the 'gentle' reminders to 'publish or perish'. Consequently, teachers are left with little time to attend to ratings feedback, let alone find time to collaborate on improving teaching. As Currie, et al. (2000) has pointed out, as universities become more 'greedy' teachers are forced to make choices on aspects of their work:

> Academics have to put in extra hours, wrestle with how to use their time, and decide what will be sacrificed for what. Increasingly, that part of their work, which is meaningful to them is displaced (p. 271).

and from a quotation in Easthope and Easthope (2000):

> I just look at it now and I mean, I don't do as good a job now as I did three years ago, or four years ago, because there is just not the time. If you have a 25% increase in workload something has got to give and basically it's the preparation and the marking…There will be less of that time for everything (p. 55).

For teachers in this study learning how to improve teaching was facilitated through dialogue and discussion, and social interaction with disciplinary peers in the structured group settings. This is as Wenger (1998, 2000) said, participating in a group, a community of practice, is vital to meaningful professional learning about practice and how to make it better. Little (2003) adds that strong teacher communities are important contributors to improving teaching and institutional reform.

It is also evident from the results that voluntary assembly of teachers may be harder than it appears. Yet, any attempt by administrators, to mandate and regulate collaborative learning from student ratings amounts to 'contrived collegiality' (A. Hargreaves, 1997). This runs the risk of collaboration becoming a burden with no teaching improvement following, especially if opportunities of time and space are not made available to facilitate such collaboration.

So, imposing collegiality will not suffice. Organising department structures and commitment to supporting collaborative arrangements such as group-based peer consultation may stand a better chance of helping teachers to collaborate to improve teaching. If department systems and structures are not adjusted to support teacher collaboration, teaching improvement activities can only remain voluntary, leaving teachers to juggle participation with academic and personal obligations. It is almost certain that efforts to improve teaching will continue to have only limited effects and only a few teachers continue to commit themselves to involvement in such activities.

Despite the short duration of the feedback intervention, the results of the study provide some insights into the likely benefits for teachers of discussing student ratings with peers in a group process: opportunity for reflection on practice, drawing on experience to solve problems, benefits of mutual support, new ways of thinking about

practice, experimentation with new methods and strategies, sharing experiences and knowledge, becoming aware of one's practice as seen through the eyes of students, engagement in discussion about the nature of practice, and the opportunity for teachers to learn with and from one another.

Scholarship of Teaching

The results of the present study seem to indicate that even at a time when there is so much emphasis on the need for effective teaching; teachers are so bogged down with academic work and administrative duties that they overlook their own need to continue learning by engaging in professional development activities. This is not made any easier, of course, by the competing demands for their time or the fact that such activities are voluntary. And yet, as seen from the MLM results, most of the variance in teaching effectiveness may be attributed to teachers themselves.

It may therefore be necessary for teachers to become more concerned about their own professional development. ·Besides, the vision for teaching and learning that underpins reform efforts in higher education requires that teachers develop a scholarly approach to teaching. According to Richlin (2001), scholarly teaching involves on-going learning about teaching and learning to be able to promote quality student learning. This leads into one form of scholarship of teaching, which is making public and opening to the scrutiny of peers, how one's teaching is actually helping students to experience quality learning.

The idea of extending 'scholarship' to teaching was first suggested by Ernest Boyer. In Boyer's (1990) terms scholarship means "engaging in original research…stepping back from one's investigation, looking for connections, building

bridges between theory and practice, and communicating one's knowledge effectively" (p. 16). Writing in the capacity as vice-president of the Carnegie Foundation for the Advancement of Teaching, Hutchings (2004) articulated the scholarship of teaching and learning as that which:

> begins with questions about how and under what circumstances students learn, and with a commitment to inquiry and evidence about those questions. It invites faculty to bring their habits, skills, and values as scholars to their work as teachers (n.p.).

Being scholarly about one's teaching therefore entails investigating issues relating to student learning, monitoring its impact, and being committed to improving one's classroom and advance practice beyond it. Importantly, this also involves deliberately seeking feedback and on-going reflection on one's teaching. However, these processes are best undertaken through dialogue and in collaboration with colleagues than as a solitary process (Light & Cox, 2001; Hutchings & Shulman, 2000). Hutchings (2004) adds, "Teaching, like any craft or art, advances when people find like-minded colleagues to work with, review their efforts, and push them to the next stages of thinking" (n.p).

Once again, group-based peer consultation is useful as it provides the opportunity for teachers to examine the evidence on their teaching, share what they are doing to enhance student learning in their classrooms, and engage in dialogue about the nature of teaching and learning in their discipline. Group-based peer consultation also allows for teachers to experiment with new teaching ideas, and share with colleagues the outcome of such experimentation.

Evaluation of the Intervention

As this is a preliminary attempt to explore the effectiveness of group-based peer consultation, it is important to examine factors that may have moderated the effectiveness of the intervention and which future research may need to give careful consideration. These are discussed under the headings: (1) design, (2) implementation, and (3) evaluation.

Intervention Design

The intervention was designed to provide support to teachers to learn from ratings feedback and develop improvement strategies for those areas of teaching for which low student ratings were received. Because the intervention was offered in peer support groups, the process may have provided less support with ratings feedback than the conventional one-to-one approach, even though self-help materials were provided. It was assumed that teachers would have engaged in private reflection with the help of the resource materials. From interview responses it was clear that this did not happen in the majority of cases. Certainly, this may have moderated the intervention effects.

The effectiveness of the intervention might have also been attenuated by the *John Henry* effect, which L'Hommedieu, et al. (1990) contended is evident in the need for ratings to be collected for the control group. It is argued that, the anticipation of student ratings is likely to influence classroom behaviour during the semester under examination. In turn, this may artificially raise the ratings for the control group, making it that more difficult to measure the effects of ratings feedback. Alongside this is the fact that teacher self-ratings were administered to the control group, which may have prompted these teachers to scrutinise their teaching more than would be the case in a true no-treatment control trial.

The time allocated to the intervention was clearly insufficient for teachers to engage in any meaningful dialogue about teaching improvement. The talk that resulted is similar to the kind of interaction described by Day (1997) as 'anecdotal exchanges and the trading of techniques'. It was also of the type Argyris and Schön (1974) termed 'single-loop' learning as the talk centred on the technical aspects of teaching. The opportunity for teachers to think deeply about teaching and learning and engage in 'double-loop' learning never arose. And yet, it is this type of learning opportunity that teacher professional development is expected to offer.

It is reflective dialogue and examination of one's conception of teaching that will allow teachers to consider how their teaching relates to student learning, and which will help them do more than just tinker with teaching. But as pointed out by Golby and Appleby (1995), teachers do not readily confront their problems from a reflective stance. This might even be more so in the absence of training and sufficient time. Because the approach was new to teachers many would not have had the competencies needed to engage in reflection at a deep level. From their two-year study with experienced teachers Wildman and Niles (1989) concluded that reflection requires substantial training. Reflection has to be learned.

Yet, as it emerged, the circumstances in this study do underscore the challenges associated with designing and implementing staff development programmes in most cases. Teachers are not able to spend quality time in such programmes and most are reduced to transmitting information about teaching through one-shot workshops/ seminars, attendance at conferences, or similar events.

Intervention Implementation

It is possible that the intervention effect could be larger but the intervention was not implemented as intended. Or that the intervention may have been implemented more effectively in some peer groups than in others, resulting in differential effects of the intervention and biasing the overall estimate of effect. And further, many of the participants in the intervention group were not fully exposed to the intervention. Only about 15% of teachers actually received the full intervention as intended, feedback ratings and meeting with peers at least twice; 45% received approximately half, feedback ratings and meeting with peers once; whereas 40% received only feedback ratings.

Although participants expressed that the intervention caused them to reflect on their teaching, they were only able to make minor changes because of the limited timeframe of one semester and the short intervention period. The common response was that the information would be used to make changes in the following semester. Once again, this point was raised by L'Hommedieu, et al. (1990) as a limitation of ratings feedback research.

Because changes to teaching within the same semester may be limited, students may in fact give teachers low end-of-semester ratings for what appear to be failure to incorporate suggestions from the mid-semester ratings. Although this issue was not examined in the present study, taking the utterances of students at the time of administering the end-of-semester ratings forms into consideration, it does seem that this might have been the case. This limitation, however, does draw attention to the need to close the 'feedback loop' by sharing ratings results with students and informing them

of changes made or might not be possible at the present time, as a result of their feedback.

Added to this is the fact that, meeting with peer partners for only once over a single four-week period, for on average forty minutes or so, during the lunch hour or when groups members could "spare a few minutes", certainly does not represent quality time in which to talk about teaching in a substantive way.

Intervention Evaluation

It might be that the intervention will influence teaching behaviours to a greater extent, but this would only be detected in a follow-up as the effect may actually emerge over time. Stevens and Aleamoni (1985) suggested that the effects of consultation may persist up to ten years. Also, given the events of the semester, it is not clear if the intervention would have been more effective had it been implemented at a different time.

At the same time, it highlights how student ratings might be used for summative evaluation without reference to the context in which teaching occurs. A limitation of the use of student ratings is that the influence of context, including workload and administrative duties or other situational circumstances, is often not considered when judging teaching quality (Koon & Murray, 1995).

Implications for Practice

The result of the study is preliminary but it does appear to point to a possible solution to the need for teachers to be supported in using student ratings feedback in the absence of individual consultation. These results, quantitative and qualitative, have a

number of interrelated implications in contributing to an understanding of the potential of group-based peer consultation, as an alternative model to individual consultation:

1. ***Structure departments to support teacher collaboration***

The qualitative data seem to suggest that group-based peer consultation may best be facilitated when it is integrated within the structure of departments. This may require that departments facilitate and co-ordinate the process, at least initially, rather than leaving it up to teachers. This means that departments may have to re-organise their structures to accommodate the process by making some relatively low cost adjustments to timetabling arrangements, allocation of duties, reward structures, and importantly to the norms of privacy by facilitating more talk among teachers about learning and teaching. Eventually, such actions could help to nurture communities of practice through which teachers may become even more committed to working together, sharing, and exchanging ideas to enhance the quality of teaching and learning in their discipline for the benefit of departments.

2. ***Provide effective departmental leadership***

Related to the point made above, is the need for department leaders to balance departmental management with leadership. Effective departmental leadership involves a commitment to contribute to quality improvement by helping teachers to teach better. This requires department leaders to offer not just lip service but explicit support through initiating, developing, and maintaining the structures, processes, and systems that may provide opportunities for teachers to collaborate and learn from student ratings.

3. *Establish comprehensive teaching evaluation systems*

Teacher professional development is essential to transform quality, and quality in teaching is crucial to the credibility of a university. This may be a good enough reason for university administrations to show that they value and support teachers by putting in place a comprehensive teaching evaluation system. Such a system may provide guidance and support for teachers to use student ratings feedback to improve their teaching and align with institutional goals for quality improvement. This system could include:

- Appropriate structures and processes to encourage teacher collaboration within departments.
- Education and communication of the research findings on student ratings of teaching to increase appreciation for ratings feedback.
- Provision of quality feedback information on teaching performance collected with well developed instruments and from multiple sources and presented in a form that adds value to the information.
- Staff development programmes that support teachers to address their specific needs, learn with their disciplinary peers, and at the same time provide the opportunity for teachers to examine their beliefs and values concerning teaching and learning.
- Development of policies and systems to ensure that all teachers participate in some form of consultation over student ratings on a regular basis to improve, develop, and maintain effective teaching practice.
- Place more emphasis on the use of student ratings for teaching improvement rather than for accountability purposes.

4. ***Allow adequate learning time with student ratings***

Before student ratings can be expected to improve teaching, teachers need plenty of time and resources to learn from the ratings feedback, to experiment with ideas, and monitor changes to their teaching. That means improvement in teaching on the basis of ratings feedback cannot reasonably be expected to be immediate but must take place over time. It may be important that teachers be given the opportunity to use student ratings solely for the purpose of teaching improvement for a specified period of time, free from scrutiny for accountability purposes. This may help teachers to focus on learning how to improve their teaching rather than waste time to resort to tactics to get favourable ratings from students.

5. ***Commit to scholarly teaching***

Teachers may need to become more committed to their own professional development and working with their colleagues within universities. Teachers should expect to take time to reflect on and examine their practice, and participate in the scholarship of teaching. This would involve giving serious attention to the views of students, systematically trying out new ideas and sharing the findings with colleagues, and engage in learning with and from colleagues, within and outside their departments.

Directions for Future Research

In addition to the issues raised in evaluating the intervention that could be considered in designing future studies, there are other limitations of the study with implications for future research. The first concerns the generalisability of the sample. The findings of the study are limited to teachers in Jamaica, although the characteristics, learning situation, and context of work for these teachers may not be much different

from university teachers elsewhere. It is important for future research to replicate the study in a variety of different settings to strengthen generalisability of findings and clearly establish the effectiveness of the feedback intervention.

Also related to generalisability is the fact that participants were volunteers who were probably already highly motivated and committed to improving their teaching practice to begin with. More likely, however, is that the relationship between the researcher and participants coupled with their knowledge of the purpose of the research may have influenced teachers, especially their interview responses.

Second, duration of the intervention was too short to adequately address teachers' needs and help them to become more aware of their beliefs about teaching and learning, in order to encourage meaningful changes rather than just tinkering with teaching. It is now generally expected that staff development programmes incorporate opportunities for teachers to examine their conceptions to teaching. Future research should replicate this study with a longer intervention period, helping teachers to articulate their values and beliefs about learning and teaching.

Third, for the most part ratings feedback studies are conducted within one semester with little or no follow-up. In this study many teachers indicated that they were unable to implement meaningful changes to their practice within the same semester. It would therefore be interesting for a follow-up study to determine whether teachers actually follow through to make changes to practice operating on their own, as well as determine the impact of the group process on the relationship between teachers.

Fourth, consistent with the literature, teachers and course variables included in the present study accounted for only a small amount of the variance in teaching effectiveness. There is, however, no reason to assume that including other teacher or

course variables would have explained much more variance. It may be important for future research to explore the effect of departmental and institutional variables to better understand the impact that departments and the institution have on teaching quality.

Fifth, although the group process was short in duration, it may have been an important element because the act of meeting and talking as indicated in the interview responses encouraged teachers to reflect on and consider changes to their teaching. It is plausible, however, that it was the ratings alone, or the ratings and self-help materials, rather than the group process that contributed more to the outcome. It may be interesting for future research to compare the effects of giving ratings feedback only, ratings feedback and a group process, and ratings feedback with self-help materials. This would provide useful information and possibly more options to staff developers in designing programmes to support teachers with ratings feedback. It may also be interesting to assess the costs of implementing and maintaining these different types.

Sixth, the present focus in quality improvement is for accountability not only at the level of the individual teacher but also at the programme, department, and institution level. Improving teaching quality therefore rest on a collaborative culture within universities. It may prove interesting for future research to examine how best a 'group-based peer consultation programme may be effectively integrated into department structures to nurture collective responsibility for improving teaching quality. Related to this is the need to determine just how cost-effective the strategy really is.

Seventh, the interaction between teachers in the groups were not documented and evaluated to determine the factors associated with the most effective groups. As it seems that group-based peer consultation is a promising strategy, it is important to

understand how the group process facilitates improvement to be able to effectively organise groups to maximise the effects of the consultation model.

A final implication is that, because the study was limited by low statistical power due to the small sample size, this has implications for the reliability of estimates as already mentioned in Chapter 7. Future research may wish to replicate the study with a larger sample to be able to use multilevel modelling techniques efficiently.

Conclusion

This study was an initial attempt to explore the possibility that group-based peer consultation may be an alternative to individual consultation. The results of the randomised controlled trial indicate that group-based peer consultation is an effective means to improve teaching effectiveness. Importantly, the empirical evidence is complemented by teacher satisfaction with the model as a practical approach to learning from ratings feedback.

It does seem, however, that the potential of a group-based peer consultation model is likely to be maximised if, it is embedded in the structure of departments, there is a supportive learning environment that provides plenty learning time and space, and teachers are provided with good quality ratings feedback information. If indeed university administrations are serious about improving the quality of teaching and learning, they need to value and support teachers by providing the structures and processes that will provide the opportunities for teachers to collaborate to learn from student ratings how to become more effective.

This preliminary study has contributed evidence-based knowledge on group-based peer consultation as an effective support mechanism with student ratings of

teaching. It is, however, just one piece of research and complex problems such as improving teaching and learning are not usually solved with one experiment. It is now necessary to refine this model of consultation and improve upon its design and implementation. For this reason, more research is needed to demonstrate its potential to improve the quality of university teaching.

Appendices

Appendix A	Letter to Administrators	266
Appendix B	Letter to Teachers	268
Appendix C	Consent Form	269
Appendix D	Procedures for Administering Student Questionnaires	270
Appendix E	SEEQ Rating Forms	271
Appendix F	Teacher Self-Evaluation Questionnaire	275
Appendix G	Frequency of Talk with Colleagues about Teaching	277
Appendix H	Letter with Ratings Report	278
Appendix I	Guidelines for Peer Support Groups	279
Appendix J	Sample of Ratings Report	280
Appendix K	Ratings Interpretation Guide	283
Appendix L	Teaching Ideas and Practical Strategies (TIPS) Sheets	285
Appendix M	Interview Guide	301
Appendix N	Letter for Control Group	302
Appendix O	Letter of Appreciation	303
Appendix P	Correlations of SEEQ Factors	304
Appendix Q	Teacher Self-Ratings by Intervention Group	306

APPENDIX A: Letter to Administrators

September, 2002

Dear

My name is Angela Penny and I am a graduate student in the School of Education at the University of Durham, England. I am requesting permission to conduct research at _____ among lecturers to evaluate the use of student ratings feedback for teaching and professional development as part of the requirements for a doctorate degree. My research supervisor is Professor Carol Fitz-Gibbon who may be contacted at telephone number 0191 374 3480 or e-mail address c.t.fitz-gibbon@durham.ac.uk.

The project is entitled: **Using Student Ratings Feedback for Professional Development.** The objective of the study is to evaluate the impact of peer consultation and student ratings feedback on teaching effectiveness. The study is necessary to provide reliable information to university teachers on how student ratings can be used to improve their teaching and for their own professional development.

Rationale for the study: When teachers are provided with student ratings feedback, they are expected to use the ratings as a tool to improve their practice. However, it takes more than providing teachers with ratings data to improve teaching performance. Research supports the process of consultation with at least one other person to provide support in interpreting, reflecting, and exploring change strategies.

A potentially effective strategy is the use of a peer support group, a structured group process to support learning from the ratings feedback. Peer support groups offer reflection space to talk about teaching and identify change strategies with the support of peers, as well as space to address issues, pick up ideas, and carry them out, as it relates of enhancing practice.

Overview: Lecturers who volunteer to participate in the project will be randomly assigned to a no feedback group or a feedback/consultation group. The control group will receive no treatment but the feedback/consultation group will meet in peer support groups for no more than 3 hours over a four-week period to talk about teaching.

Benefits:

For teachers:

- Increased awareness of the need to see teaching in a community of practice rather than a privatised activity.

- An opportunity for deeper levels of reflection, experimentation, and learning about teaching and how to make student ratings feedback work to achieve excellence in teaching.

- An opportunity to share the experience of examining student ratings in a supportive manner along with the experience of doing action research in one's classroom.

For the institution:

- An opportunity to pilot a study on research into teaching, essential to building a culture of learning, an objective many universities are anxious to achieve.

- Involvement would demonstrate the institution's seriousness in assuring quality, as well as provide information on the way forward for staff development plans and policies.

- A change in desire and motivation in using student feedback should affect the quality of teaching and learning in a positive manner.

- Improved atmosphere in the classroom, among faculty members, and with administrators as it relates to the collection of student ratings.

Risks: There are no foreseen risks to participating in this study. Data collected will be used only for research purposes, and accessed only by the researcher and the research supervisor. All data will be maintained with the strictest of confidence in keeping with ethical research procedures. The University will not be identified in the research report.

I do believe that this research project should be able to make a valuable contribution to the quality assurance efforts of your University. Please contact me via e-mail: a.r.penny@durham.ac.uk for answers to any query you might have, as well as any ideas you might have for this proposed study. In addition, I am able available to meet with you on a date and time that is convenient to you to discuss the research in more details.

Sincerely yours

Angela Penny

APPENDIX B: Letter to Teachers

Dear Colleague

I am requesting your participation in a research project that will evaluate the impact of student ratings feedback with consultation on teaching effectiveness and professional development. The study, which forms part of my doctoral programme, is necessary to provide reliable information to university teachers on how student ratings can be used to promote quality in higher education for both teachers and students.

Rationale for the study: When teachers are provided with student ratings feedback, they are expected to use the ratings as a tool to improve their practice. However, it takes more than providing teachers with ratings data to enhance teaching. Teachers need to know how to improve and they need help and support in effecting changes. Research evidence supports the process of consulting with at least one other person to provide support in interpreting, reflecting, and exploring change strategies.

Overview: Volunteers will be randomly assigned to a "no-feedback" group or a "feedback/ consultation" group. All volunteers will be asked to administer a short questionnaire in their classes at week 5 and week 13 in the semester. Persons selected for the "feedback/consultation" group will meet with peers for no more than 3 hours over a four-week period to reflect on the results, talk about teaching, address issues, and pick up ideas as it relates to enhancing practice. Persons assigned for the "no-feedback" group will receive their training at the end of the semester.

Benefits: By participating in this study you will have an opportunity to share with others the experience of deeper levels of reflection and interpretation of ratings results in an atmosphere that is supportive and non-evaluative. Participation also offers much scope for personal and professional development. There is also the opportunity to share the experience of doing action research in one's classroom.

Risks: There are no foreseen risks to participating in this study. Ratings results will be sent only to you, and accessed only by the researcher. Further, the project is not in any way affiliated with the student ratings system used in this institution. You are also assured that all data collected will be used for research purposes only and will be maintained with the strictest of confidence. Neither you nor your institution will be identified in the research report.

Should you decide to participate, or if you have any questions or concerns please contact me by telephone at extension 3135, or e-mail: arpenny@hotmail.com, by _____.
Your assistance, in providing evidence and information on how university teachers might effectively use student ratings to enhance teaching excellence while promoting change and development in teaching in higher education, as well as contribute to knowledge in this area, will be greatly appreciated.

Sincerely yours

Angela Penny

APPENDIX C

CONSENT FORM

USING STUDENT RATINGS FEEDBACK FOR TEACHER PROFESSIONAL DEVELOPMENT

Researcher: Angela R. Penny Date: September 2002

Thank you for agreeing to participate in this study that will take place during the semester September to December, 2002.

This form outlines the purpose of the study and provides a description of your involvement and rights as a participant.

Purpose: The purpose of the study is to evaluate the use of feedback from student ratings and peer consultation for improving teaching effectiveness and for teacher professional development.

Requirements: Your involvement will require the following:

- 10-15 minutes of class time to administer a student questionnaire
- Completion of a self-rating form
- Attendance at peer support group sessions if you are assigned to receive the intervention

Confidentiality: All data will be treated confidentially. Information obtained about you will not be shared with the University; neither will your identity be disclosed in the research report.

Participation: Your participation in this research is voluntary, and you may withdraw from the study at any time for any reason and without prejudice.

Consent

I have read this consent form and understood the information. I agree to participate in this study.

_____ _____

Participant's signature Researcher's signature

_____ _____

Date Date

APPENDIX D

PROCEDURES FOR ADMINISTERING STUDENT QUESTIONNAIRES

Instructions

1. The forms shall be administered preferably at the beginning of the class period and students should be given at least 10 minutes to complete the forms.

2. Ask for a student volunteer to distribute and collect the questionnaires. The lecturer should leave the room while the questionnaire is being administered.

For the student volunteer

Please read the following instructions:

I am going to hand out a questionnaire that will ask you to evaluate several aspects of this course and the teaching. This exercise forms part of a project to determine how student ratings can be used to improve teaching in the university and is not related to the evaluation that _____ will administer at the end of the semester.

This questionnaire is the first of two that your teacher has agreed to administer in this class to receive feedback about this course and on his/her teaching.

The completion of the questionnaire is entirely voluntary, but your lecturer would like to obtain your views. Your response is anonymous; please DO NOT WRITE YOUR NAME ON THE FORM.

Your lecturer will not see the questionnaires but will receive a summary report to allow him or her to make necessary changes to his or her teaching. Your feedback would be very much appreciated.

APPENDIX E: Mid-term Rating Form

STUDENTS' EVALUATION OF EDUCATIONAL QUALITY (SEEQ)

The purpose of this survey is to give your lecturer your views about his/her teaching. Please base your response on his/her teaching in this class. Your name is NOT required and all information is confidential.

Lecturer: …… Course: …… Date: ……

Please indicate by circling the most appropriate number to indicate the EXTENT of your agreement with the following statements as description of this subject by using the following scale: ⟶	**1= Strongly Disagree 2=Disagree 3=Neutral 4=Agree 5=Strongly Agree**				
LEARNING AND ACADEMIC VALUE					
1. You are finding the course intellectually challenging and stimulating	1	2	3	4	5
2. You are learning something which you consider valuable	1	2	3	4	5
3. Your interest in the subject is increasing as a consequence of this course	1	2	3	4	5
4. You are learning and understanding the subject materials of this course	1	2	3	4	5
LECTURER ENTHUSIASM					
5. Lecturer is enthusiastic about teaching the course	1	2	3	4	5
6. Lecturer is dynamic and energetic in conducting the course	1	2	3	4	5
7. Lecturer enhances presentation with the use of humour	1	2	3	4	5
8. Lecturer's style of presentation holds your interest during class	1	2	3	4	5
ORGANIZATION/CLARITY					
9. Lecturer's explanations are clear	1	2	3	4	5
10. Course materials are well prepared and carefully explained	1	2	3	4	5
11. Proposed objectives agreed with those actually taught so you know where the course is going	1	2	3	4	5
12. Lecturer gives lectures/tutorials that facilitated taking notes	1	2	3	4	5
GROUP INTERACTION					
13. Students are encouraged to participate in class discussions	1	2	3	4	5
14. Students are invited to share their ideas and knowledge	1	2	3	4	5
15. Students are encouraged to ask questions and were given meaningful answers	1	2	3	4	5
16. Students are encouraged to express their own ideas and/or question the lecturer	1	2	3	4	5
INDIVIDUAL RAPPORT					
17. Lecturer is friendly toward individual students	1	2	3	4	5
18. Lecturer makes students feel welcome in seeking help/advice in or outside of class	1	2	3	4	5
19. Lecturer has a genuine interest in individual students	1	2	3	4	5
20. Lecturer is adequately accessible to students during office hours or after class	1	2	3	4	5
BREADTH OF COVERAGE					
21. Lecturer contrasts the implications of various theories	1	2	3	4	5
22. Lecturer presents the background or origin of ideas/concepts developed in class	1	2	3	4	5
23. Lecturer presents points of view other than his/her own when appropriate	1	2	3	4	5
24. Lecturer adequately discusses current developments in the field.	1	2	3	4	5
ASSESSMENT/GRADING					
25. Feedback on assessments/graded material was valuable	1	2	3	4	5
26. Methods of evaluating student work are fair and appropriate	1	2	3	4	5
27. Assessments/graded materials test class content as emphasized by the lecturer	1	2	3	4	5
ASSIGNMENTS/READINGS					
28. Required readings/texts are valuable	1	2	3	4	5
29. Readings, homework, etc. contributes to appreciation and understanding of the subject	1	2	3	4	5
OVERALL RATING **(1=Very poor…2= Poor…3=Average…4=Good…5=Very Good)**					
30. Overall, how does this class compare with other classes at this institution?	1	2	3	4	5
31. Overall, how does this lecturer compare with other lecturers at this institution?	1	2	3	4	5

BACKGROUND SUBJECT/CLASS CHARACTERISTICS

Subject difficulty, relative to other subjects, is:
1 = Very Easy
2 = Easy
3 = Medium
4 = Hard
5 = Very Hard

Subject workload, relative to other subjects, is:
1 = Very Light
2 = Light
3 = Medium
4 = Heavy
5 = Very Heavy

Subject pace is:
1 = Too Slow
2 = Slow
3 = About Right
4 = Fast
5 = Too Fast

1. Your Gender:
☐ Male ☐ Female

2. Your expected subject mark:
☐ U – D+ ☐ C – C+ ☐ B – B+ ☐ A

2. In comparison to other units, how easy is it to get good marks in this subject?
☐ Very Easy, ☐ Easy, ☐ Average, ☐ Difficult, ☐ Very Difficult

3. Level of interest in this subject before the start of the unit:
☐ Very Low, ☐ Low, ☐ Medium, ☐ High, ☐ Very High

4. Year in course:
☐ First, ☐ Second, ☐ Third, ☐ Fourth

OPEN-ENDED COMMENTS

Please indicate the important characteristics of this lecturer/class that you feel are most important for him/her to improve (particularly aspects not covered by the rating items).

1. .....
.....
.....
.....
2. .....
.....
.....
.....
3. .....
.....
.....
.....
.....

Please use the additional space to clarify any of your responses or to make other comments.

.....
.....
.....
.....
.....
.....

Thank you for your feedback

APPENDIX E: End of Term Rating Form

STUDENTS' EVALUATION OF EDUCATIONAL QUALITY (SEEQ)

The purpose of this survey is to give your lecturer your views about his/her teaching. Please base your response on his/her teaching in this subject. Your name is NOT required and all information is confidential.

Lecturer: …… Course: …….. Date: ……..

Please indicate by circling the most appropriate number to indicate the EXTENT of your agreement with the following statements as description of this subject by using the following scale: ⟶	**1= Strongly Disagree 2=Disagree 3=Neutral 4=Agree 5=Strongly Agree**				
LEARNING AND ACADEMIC VALUE					
1. You found the course intellectually challenging and stimulating	1	2	3	4	5
2. You have learned something which you consider valuable	1	2	3	4	5
3. Your interest in the subject has increased as a consequence of this course	1	2	3	4	5
4. You have learned and understood the subject materials of this course	1	2	3	4	5
LECTURER ENTHUSIASM					
5. Lecturer was enthusiastic about teaching the course	1	2	3	4	5
6. Lecturer was dynamic and energetic in conducting the course	1	2	3	4	5
7. Lecturer enhanced presentation with the use of humour	1	2	3	4	5
8. Lecturer's style of presentation held your interest during class	1	2	3	4	5
ORGANIZATION/CLARITY					
9. Lecturer's explanations were clear	1	2	3	4	5
10. Course materials were well prepared and carefully explained	1	2	3	4	5
11. Proposed objectives agreed with those actually taught so you knew where the course was going	1	2	3	4	5
12. Lecturer gave lectures/tutorials that facilitated taking notes	1	2	3	4	5
GROUP INTERACTION					
13. Students were encouraged to participate in class discussions	1	2	3	4	5
14. Students were invited to share their ideas and knowledge	1	2	3	4	5
15. Students were encouraged to ask questions and were given meaningful questions	1	2	3	4	5
16. Students were encouraged to express their own ideas and/or question the lecturer	1	2	3	4	5
INDIVIDUAL RAPPORT					
17. Lecturer was friendly toward individual students	1	2	3	4	5
18. Lecturer made students feel welcome in seeking help/advice in or outside of class	1	2	3	4	5
19. Lecturer had a genuine interest in individual students	1	2	3	4	5
20. Lecturer was adequately accessible to students during office hours or after class	1	2	3	4	5
BREADTH OF COVERAGE					
21. Lecturer contrasted the implications of various theories	1	2	3	4	5
22. Lecturer presented the background or origin of ideas/concepts developed in class	1	2	3	4	5
23. Lecturer presented points of view other than his/her own when appropriate	1	2	3	4	5
24. Lecturer adequately discussed current developments in the field	1	2	3	4	5
EXAMINATIONS/GRADING					
25. Feedback on examinations/graded material was valuable	1	2	3	4	5
26. Methods of evaluating student work were fair and appropriate	1	2	3	4	5
27. Examinations/graded materials tested class content as emphasized by the lecturer	1	2	3	4	5
ASSIGNMENTS/READINGS					
28. Required readings/texts were valuable	1	2	3	4	5
29. Readings, homework, etc. contributed to appreciation and understanding of the subject	1	2	3	4	5
OVERALL RATING **(1=Very poor…2= Poor…3=Average…4=Good…5=Very Good)**					
30. Overall, how does this class compare with other classes at this institution?	1	2	3	4	5
31. Overall, how does this lecturer compare with other lecturers at this institution?	1	2	3	4	5

BACKGROUND SUBJECT/CLASS CHARACTERISTICS

Subject difficulty, relative to other subjects, was:
1 = Very Easy
2 = Easy
3 = Medium
4 = Hard
5 = Very Hard

Subject workload, relative to other subjects, was:
1 = Very Light
2 = Light
3 = Medium
4 = Heavy
5 = Very Heavy

Subject pace was:
1 = Too Slow
2 = Slow
3 = About Right
4 = Fast
5 = Too Fast

1. Your Gender:
☐ Male ☐ Female

2. Your expected subject mark:
☐ U – D+ ☐ C – C+ ☐ B – B+ ☐ A

2. In comparison to other units, how easy is it to get good marks in this subject?
☐ Very Easy, ☐ Easy, ☐ Average, ☐ Difficult, ☐ Very Difficult

3. Level of interest in this subject before the start of the unit:
☐ Very Low, ☐ Low, ☐ Medium, ☐ High, ☐ Very High

4. Year in course:
☐ First, ☐ Second, ☐ Third, ☐ Fourth

OPEN-ENDED COMMENTS

Please indicate the important characteristics of this lecturer/class that you feel are most important for him/her to improve (particularly aspects not covered by the rating items).

1. .....
.....
.....
.....
2. .....
.....
.....
.....
3. .....
.....
.....
.....

Please use the additional space to clarify any of your responses or to make other comments.

.....
.....
.....
.....
.....

Thank you for your feedback

APPENDIX F

TEACHER SELF-EVALUATION QUESTIONNAIRE

Please indicate the EXTENT of your agreement with the following statements as descriptions of *your* teaching in the class you have chosen to conduct SEEQ by circling the appropriate number on the following scale: ⟶

1=Strongly Disagree
2=Disagree
3=Neutral
4=Agree
5=Strongly Agree

		SD	D	N	A	SA
LEARNING AND ACADEMIC VALUE						
1	Students are finding this class intellectually challenging and stimulating	1	2	3	4	5
2	Students are learning something they consider valuable	1	2	3	4	5
3	Students' interest in the subject is increasing as a consequence of this class	1	2	3	4	5
4	Students are learning and understanding the subject materials of this class	1	2	3	4	5
INSTRUCTOR'S ENTHUSIASM						
5	I am enthusiastic about teaching the class	1	2	3	4	5
6	I am dynamic and energetic in conducting the class	1	2	3	4	5
7	I enhance presentation with the use of humour	1	2	3	4	5
8	My style of presentation hold students' interest during class	1	2	3	4	5
ORGANIZATION/CLARITY						
9	My explanations are clear	1	2	3	4	5
10	Course materials are well prepared and carefully explained	1	2	3	4	5
11	Proposed objectives agree with those actually taught so students know where the course is going	1	2	3	4	5
12	I give lectures (tutorials) that facilitated taking notes	1	2	3	4	5
GROUP INTERACTION						
13	Students are encouraged to participate in class discussions	1	2	3	4	5
14	Students are invited to share their ideas and knowledge	1	2	3	4	5
15	Students are encouraged to ask questions and are given meaningful answers	1	2	3	4	5
16	Students are encouraged to express their own ideas and/or question me	1	2	3	4	5
INDIVIDUAL RAPPORT						
17	I am friendly toward individual students	1	2	3	4	5
18	I make students feel welcome in seeking help/advice in or outside of class	1	2	3	4	5
19	I had a genuine interest in individual students	1	2	3	4	5
20	I am adequately accessible to students during office hours or after class	1	2	3	4	5
BREADTH OF COVERAGE						
21	I contrast the implications of various theories	1	2	3	4	5
22	I present the background or origin of ideas/concepts developed in class	1	2	3	4	5
23	I present points of view other than my own when appropriate	1	2	3	4	5
24	I adequately discussed current developments in the field.	1	2	3	4	5
ASSESSMENT/GRADING						
25	Feedback on assessments/graded material was valuable	1	2	3	4	5
26	Methods of evaluating student work are fair and appropriate	1	2	3	4	5
27	Assessments/graded materials tested class content as emphasized by me	1	2	3	4	5
ASSIGNMENTS/READINGS						
28	Required readings/texts are valuable to students	1	2	3	4	5
29	Readings, homework, etc. contributes to appreciation and understanding of the subject	1	2	3	4	5
OVERALL RATING						
	(1=Very poor…2= Poor…3=Average…4=Good…5=Very Good)					
30	Overall, how does this class compare with other classes at this institution?	1	2	3	4	5
31	Overall, how does this course compare with other courses at this institution?	1	2	3	4	5

BACKGROUND SUBJECT/CLASS CHARACTERISTICS

You may wish to anticipate what the majority of students are likely to say to the following questions:

Subject difficulty, relative to other subjects, is: **1 = Very Easy 2 = Easy 3 = Medium 4 = Hard 5 = Very Hard**

Subject workload, relative to other subjects, is: **1 = Very Light 2 = Light 3 = Medium 4 = Heavy 5 = Very Heavy**

Subject pace is: **1 = Too Slow 2 = Slow 3 = About Right 4 = Fast 5 = Too Fast**

IMPORTANT COMPONENTS OF TEACHING EFFECTIVENESS

Please indicate how important you consider the following factors to be in teaching this subject effectively. **1 = Very important; 2 = Important; 3 = Somewhat important; 4 = Not important**

Learning/Academic Value	1	2	3	4
Instructor Enthusiasm	1	2	3	4
Organisation/Clarity	1	2	3	4
Group Interaction	1	2	3	4
Individual Rapport	1	2	3	4
Breadth of Coverage	1	2	3	4
Assessment/Grading	1	2	3	4
Assignments/Readings	1	2	3	4

BACKGROUND CHARACTERISTICS

1. Name: .....

2. Your gender: ☐ Female ☐ Male

3. Age group: ☐ 25 – 30 ☐ 31 – 35 ☐ 36 – 40
   ☐ 41 – 45 ☐ 46 – 50 ☐ 51 – 60

4. Your rank: ☐ Assistant Lecturer ☐ Lecturer
   ☐ Senior Lecturer ☐ Principal Lecturer
   ☐ Other (Please specify) .....

5. Your Discipline:.....

6. Your teaching experience at this level:

   ☐ Less than 1 year ☐ 1 to 3 years ☐ 4 to 6 years
   ☐ 7 to 10 years ☐ Over 10 years

7. How often do you talk to colleagues about your teaching?

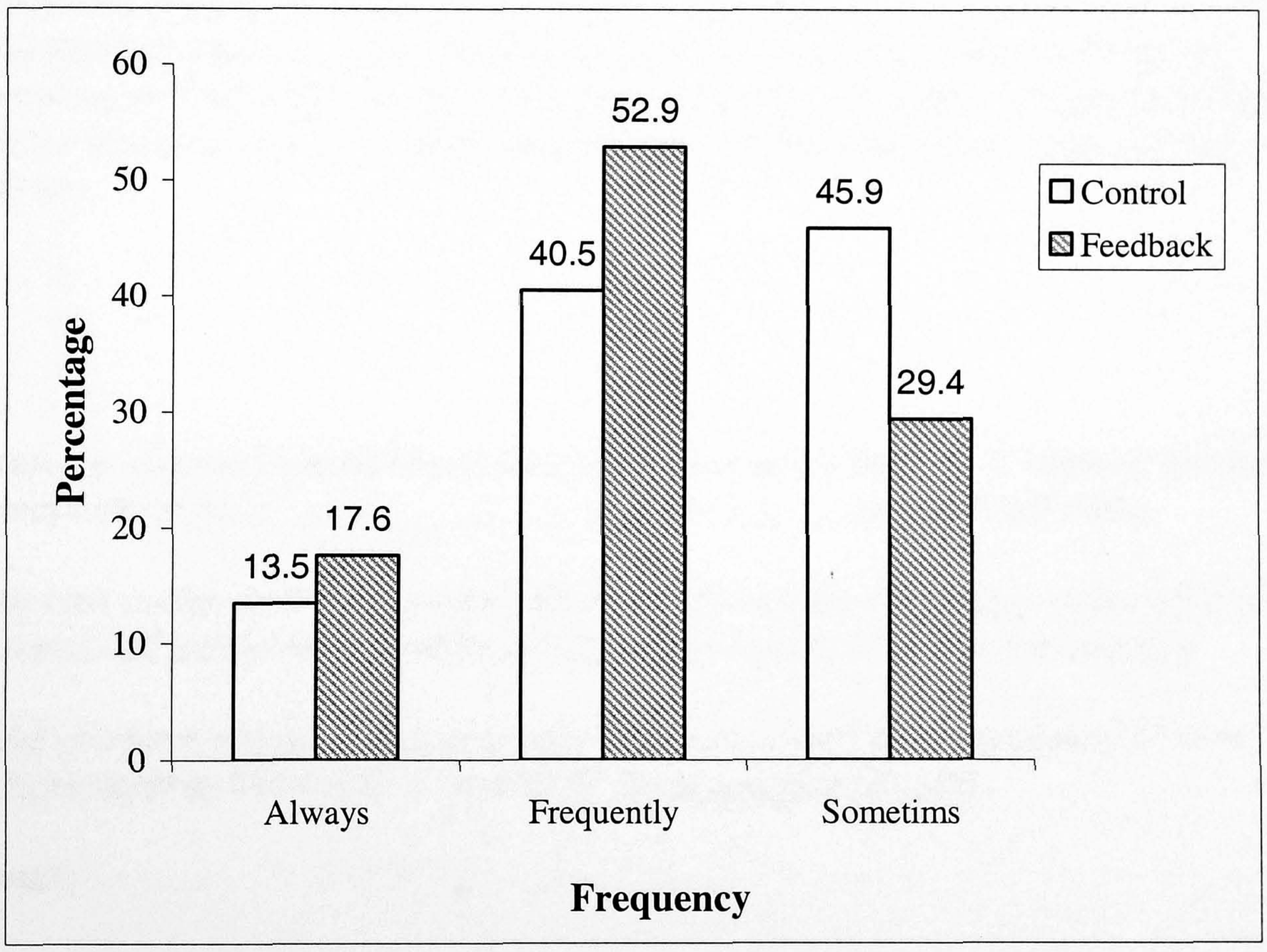
   ☐ Always ☐ Frequently ☐ Sometimes ☐ Rarely ☐ Never

8. How useful do you find student ratings for improving teaching?

   ☐ Extremely useful ☐ Moderately useful ☐ Somewhat useful ☐ Rarely useful

Appendix G

Frequency of Talk about Teaching with Colleagues



APPENDIX H: Letter with Ratings Report

Dear Colleague

Thank you for agreeing to participate in the research project that will explore the use of student ratings feedback for teaching improvement in higher education.

As indicated already the study uses an experimental design. I am pleased to inform that your department/section was randomly selected for the intervention which entails interpreting and reflecting on the ratings feedback with your peers in a group setting. From the teachers who volunteered to participate in the study the members of your group are:

1.
2.
3.

To start the process I would like to meet with your group for a short briefing session on the intervention on _____ at _____ in your staff room.

Please find enclosed in your results packet the report from the ratings collected from your class and additional information that will be examined at our first meeting.

Should you have any questions or concerns please contact me at extension 3135 or telephone number 941-0286 or via e-mail, arpenny@hotmail.com.

Sincerely

Angela Penny

APPENDIX I

GUIDELINES FOR PEER SUPPORT GROUP

Objective: To learn a method of consultation that supports individuals in learning from student ratings feedback.

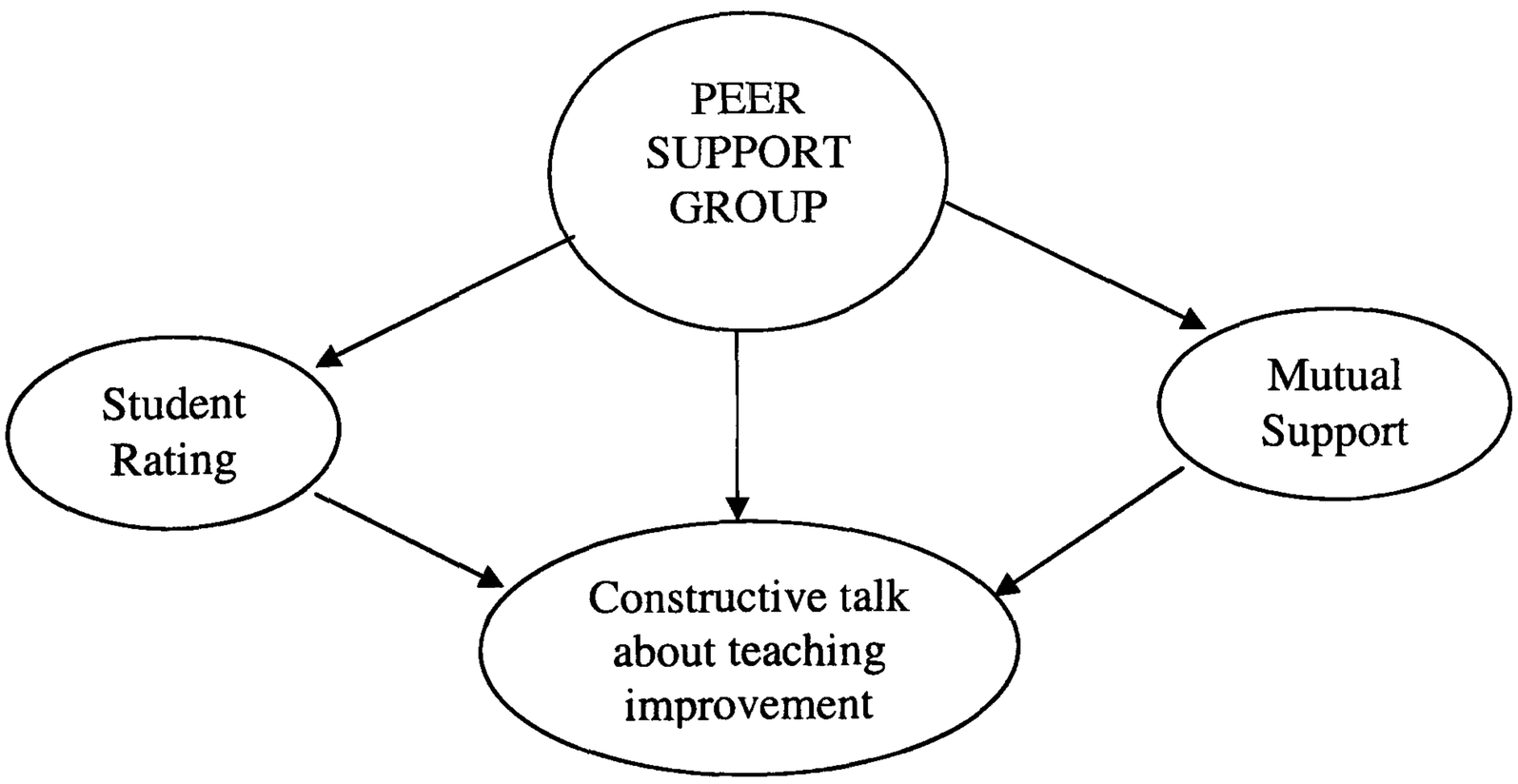
Rationale: This peer support group is organised as a structured group process to consult over student ratings feedback and involves informal meetings with peers to learn from student ratings to improve teaching effectiveness. It offers an opportunity to gain and give motivational, emotional, and informational support to colleagues to improve teaching. It provides a reflection space, where peers can talk about teaching and learning, pick up ideas, and carry them out. Improvement is needed not because there is a perceived problem but to build commitment to excellence.

Benefits: An opportunity to learn with others the experience of deeper levels of reflection and interpretation of ratings results in an atmosphere that is supportive and non-evaluative. NOT group work but learning together. NOT to develop a shared understanding but to share what one has come to understand.

Guidelines

- Maintain strict confidentiality on the discussion in the groups
- Use the following process
  - present to group area targeted for improvement, based on low rating
  - discuss the problem by sharing experiences, insights, knowledge
  - generate possible improvement strategies, using TIPS for reference
  - plan and implement change
  - share with the group results of implemented strategy
- Maintain contact with telephone and e-mail between meetings

Elements of the peer support group process

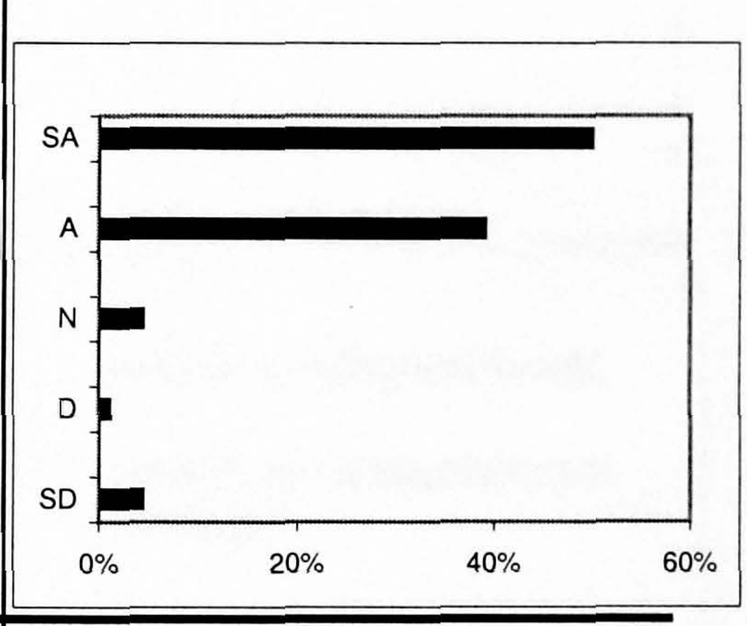


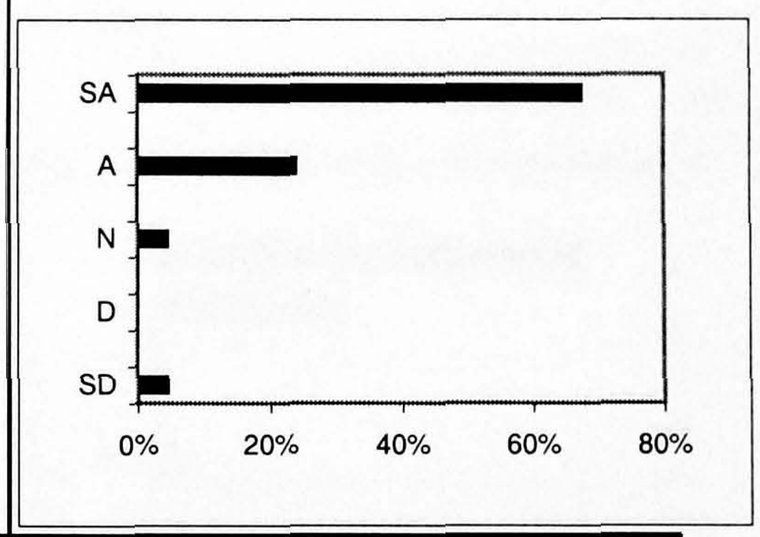
APPENDIX J

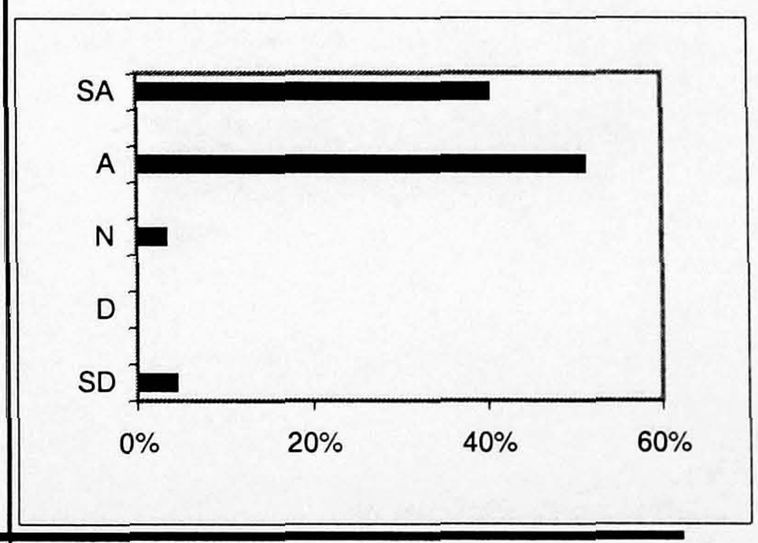
SEEQ Ratings Report

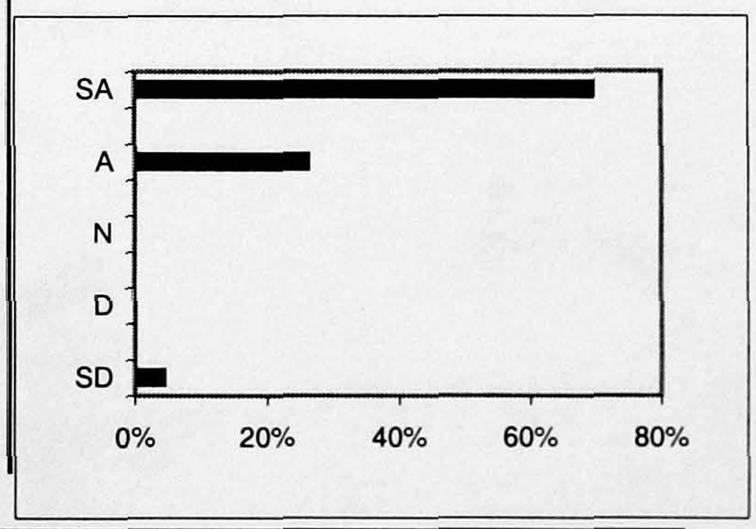
Institution: University A
Name: Miss R. Ratings
Course: Communication
Period: Mid-semester
No. responding: 23

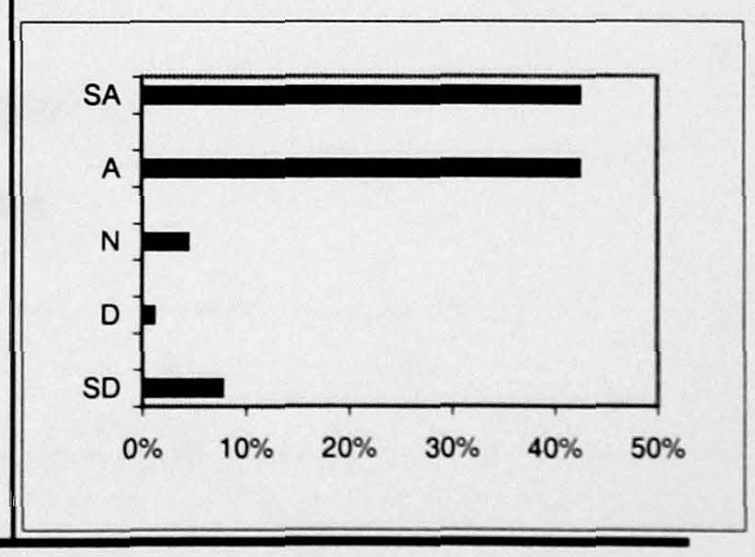
SUMMARY OF ITEMS	SD	D	N	A	SA	Mean
Learning and Academic Value						
1. Finds the course intellectually challenging	1	1	1	10	10	4.2
	4%	4%	4%	43%	43%	
2. Learning something considered valuable	1	0	1	1	20	4.7
	4%	0%	4%	4%	87%	
3. Interest in the subject is increasing	1	0	2	11	9	4.2
	4%	0%	9%	48%	39%	
4. Learning and understanding subject materials	1	0	0	14	7	4.0
	4%	0%	0%	61%	30%	
Factor Mean =	**4.3**			**Stan. Dev.**		**1.0**
Enthusiasm						
5. Enthusiastic about teaching	1	0	0	5	17	4.6
	4%	0%	0%	22%	74%	
6. Dynamic and energetic	1	0	0	7	15	4.5
	0%	0%	0%	1%	3%	
7. Enhances presentation with humour	1	0	2	3	17	4.5
	4%	0%	9%	13%	74%	
8. Style of presentation holds interest	1	0	2	7	13	4.3
	4%	0%	9%	30%	57%	
Factor Mean =	**4.5**			**Stan. Dev.**		**0.9**
Organization/Clarity						
9. Explanations are clear	1	0	0	12	10	4.3
	4%	0%	0%	52%	43%	
10. Carefully explained course materials	1	0	1	14	7	4.1
	4%	0%	4%	61%	30%	
11. Proposed objectives agree with lessons	1	0	1	12	9	4.2
	4%	0%	4%	52%	39%	
12. Class facilitated note taking	1	0	1	9	11	4.1
	4%	0%	4%	39%	48%	
Factor Mean =	**4.2**			**Stan. Dev.**		**1.0**
Group Interaction						
13. Encouraged to participate in discussions	1	0	0	5	17	4.6
	4%	0%	0%	22%	74%	
14. Invited to share ideas and knowledge	1	0	0	6	16	4.6
	4%	0%	0%	26%	70%	
15. Encouraged to ask questions	1	0	0	9	13	4.4
	4%	0%	0%	39%	57%	
16. Encouraged to express own ideas	1	0	0	4	18	4.7
	4%	0%	0%	17%	78%	
Factor Mean =	**4.6**			**Stan. Dev.**		**0.9**
Individual Rapport						
17. Friendly toward individual students	2	1	0	11	8	3.8
	9%	4%	0%	48%	35%	
18. Students feel welcome in seeking help	2	0	0	11	10	4.2
	9%	0%	0%	48%	43%	
19. Genuine interest in individual students	2	0	1	10	9	3.9
	9%	0%	4%	43%	39%	
20. Accessible after class	1		3	7	12	4.3
	4%	0%	13%	30%	52%	
Factor Mean =	**4.6**			**Stan. Dev.**		**0.9**



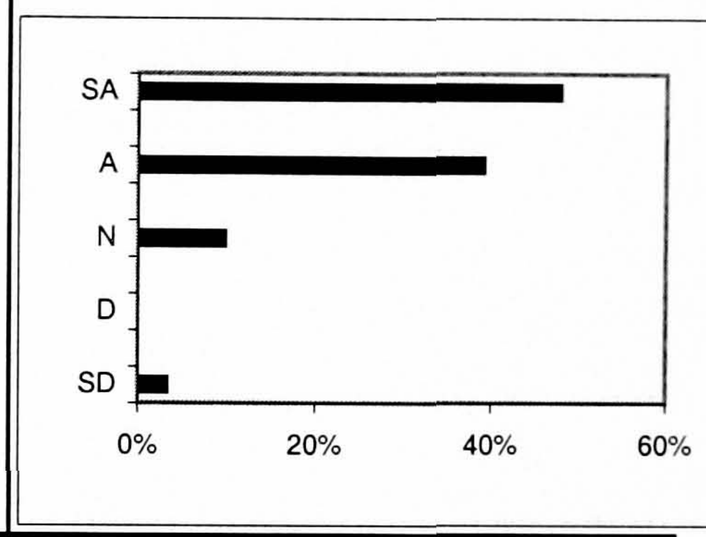




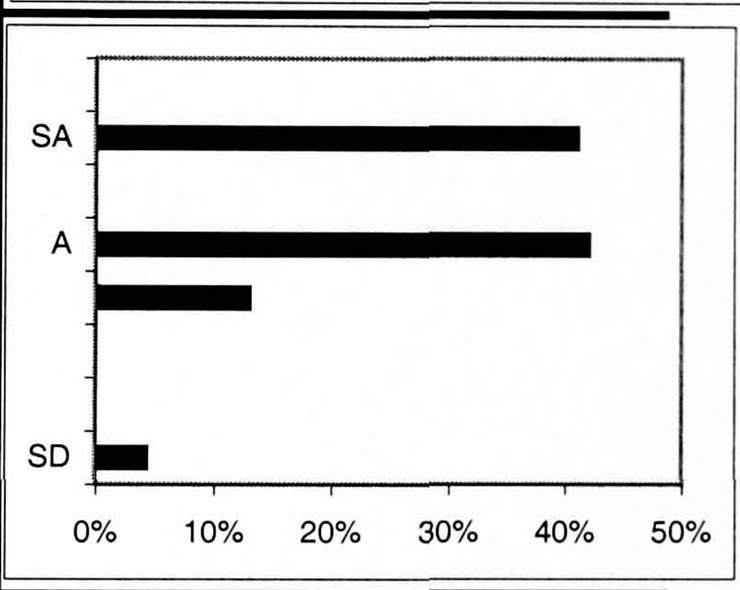




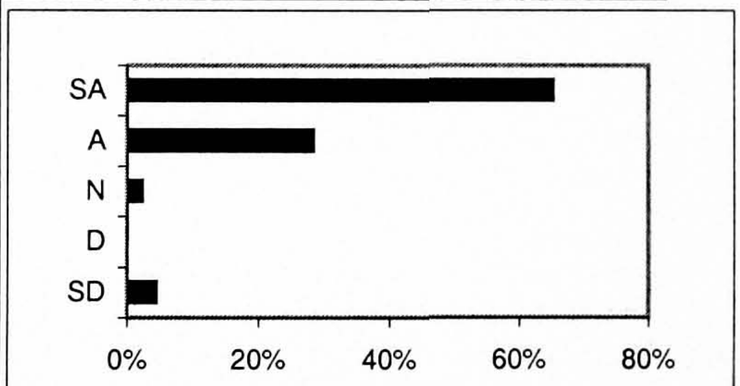
Breadth of Coverage						
21. Contrasts various theories	1	0	4	11	7	4.0
	4%	0%	17%	48%	30%	
22. Presents background of concepts	0	0	4	11	8	4.2
	0%	0%	17%	48%	35%	
23. Presents other points of views	1	0	1	8	13	4.4
	4%	0%	4%	35%	57%	
24. Discusses developments in field	1	0	0	6	16	4.6
	4%	0%	0%	26%	70%	
Factor Mean =	**4.0**			**Stan. Dev.**		**1.2**



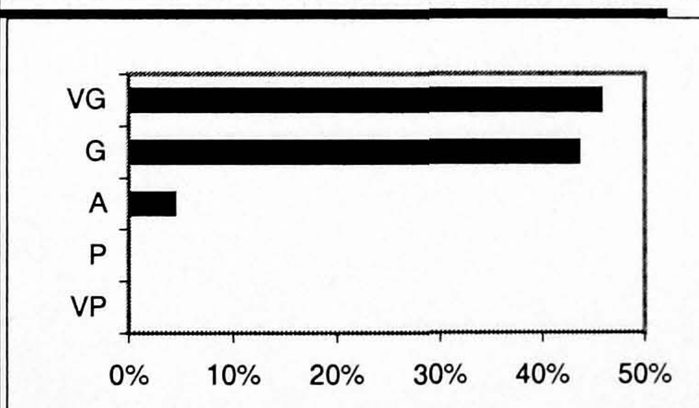
Assessment/Grading						
25. Valuable feedback on assessments	1	0	3	11	8	4.1
	4%	0%	13%	48%	35%	
26. Fair evaluation of work	1	0	3	10	9	4.1
	4%	0%	13%	43%	39%	
27. Assessments test class content	1	0	3	8	11	4.2
	4%	0%	13%	35%	48%	
Factor Mean =	**4.3**			**Stan. Dev.**		**0.9**



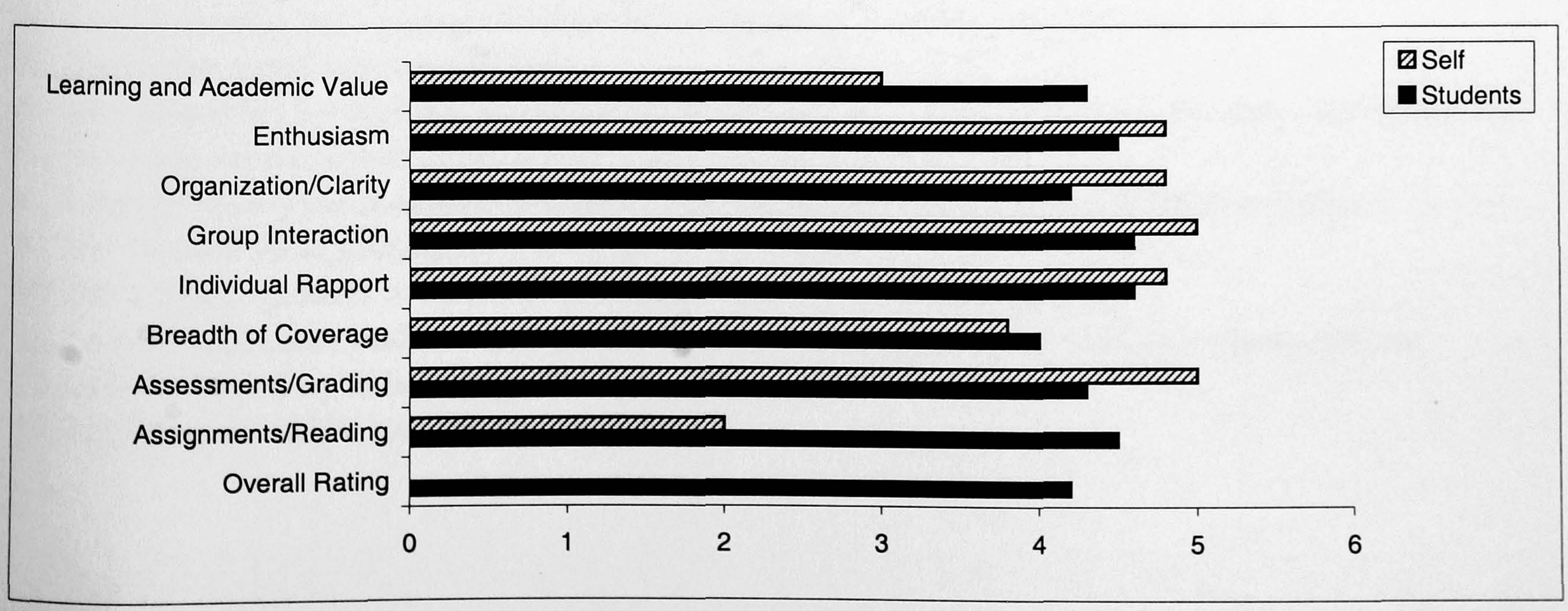
Assignments/Readings						
28. Required readings are valuable	1	0	1	7	14	4.4
	4%	0%	4%	30%	61%	
29. Readings contribute to understanding	1	0	0	6	16	4.6
	4%	0%	0%	26%	70%	
Factor Mean =	**4.5**			**Stan. Dev.**		**0.9**



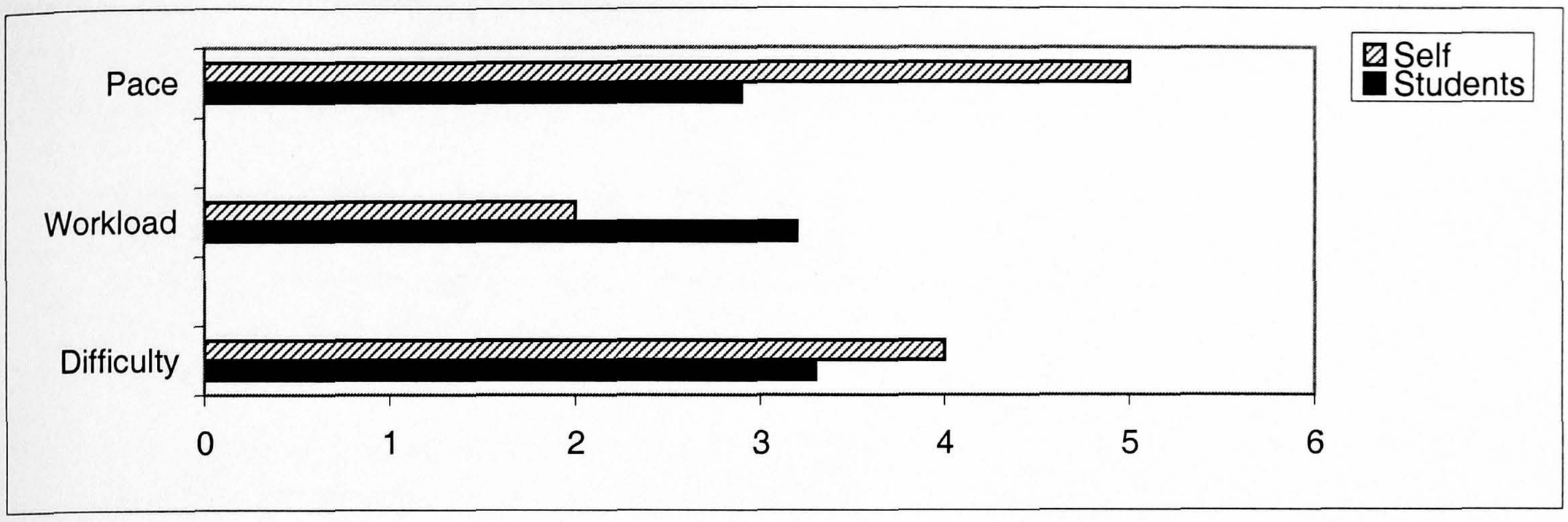
Overall Rating	**VP**	**P**	**A**	**G**	**VG**	
30. Overall comparison with other classes	0	0	1	12	9	4.2
	0%	0%	4%	52%	39%	
31. Overall comparison with other teachers	0	0	1	8	12	4.1
	0%	0%	4%	35%	52%	
Factor Mean =	**4.2**			**Stan. Dev.**		**1.2**



COMPARISON OF SELF-RATINGS WITH STUDENT RATINGS



SUBJECT DIFFICULTY/WORKLOAD



Please see questionnaire for key to scale

IMPORTANCE RATINGS AND WRITTEN COMMENTS

Importance ratings:

Learning and Academic Value	Important
Enthusiasm	Very Important
Organization/Clarity	Very Important
Group Interaction	Very Important
Individual Rapport	Very Important
Breadth of Coverage	Important
Assessment/Grading	Important
Assignments/Readings	Very Important

Written Comments:

1. This class is great, very interactive, very practical, dynamic
2. Lecturer facilitates discussions on principles.
3. Lecturer is open to new ideas, has a genuine interest in students
4. Course material isn't limited to Jamaica/Caribbean but is international which is good
5. The course needs user-friendly text
6. She intimidates her students and in the future she may need to do a course in dynamics and leadership
7. She gives valid examples from a very wide cross section of learning
8. Lecturer has a good command of subject material and teaches in a clear and concise manner
9. Not too sure about grading style and exactly what is expected in assignments
10. Use of guest speakers is very helpful in making the course more practical
11. Lecturer stimulates students to think and do their own research rather than spoon feed students
12. Lecturer should try to be less serious
13. More time should be spent on each topic

APPENDIX K: Interpretation Guide

Interpreting Your Students' Evaluation of Educational Quality (SEEQ) Results

Although student ratings of teaching is important and provides useful information for analysing teaching effectiveness, they are only one factor and should <u>not</u> be considered in isolation from other sources of information.

The report of student ratings of your instruction describes the general reactions of students to the subject and your teaching. The report is designed to give you feedback to enable you to enhance your teaching, and not necessarily to measure your teaching effectiveness.

The report summarises students' responses to items, which comprised the factors of the SEEQ questionnaire in terms of the number and percent of students selecting each response category, means for items and factors, standard deviations for factors, and a bar graph. The report also highlights discrepancies between your own ratings and ratings actually received from students.

In interpreting the results, first be sure the ratings reflect a representative sample of the class. If less than half of the class responded to the questionnaire the results may not be representative of the class as a whole. Carefully consider information such as distribution of responses by items, means, medians, and standard deviations.

The mean gives the typical student response on an item and the factor. However, it may not be particularly informative when the responses vary widely. A good practice is to examine the pattern of responses to determine the usefulness of the mean. If there is a wide spread in responses the mean is not a good measure of student responses.

The standard deviation gives an index of agreement or disagreement among students. A standard deviation of less than 1.0 (on a 5-point scale) indicates a relatively good agreement among students. Deviations above 1.2 indicate a divided class on that item. This may represent differences in the nature of the student (e.g. age, background, interest, etc.).

It is important to note your highest and lowest rated items. See whether your strengths or weaknesses cluster on any of the factors of the SEEQ questionnaire.

As a rule of thumb, there is usually cause for concern when a third of the students give low ratings to some aspect of a course. In looking at your highest and lowest rated items, try to identify specific teaching behaviours that might have caused students to give those ratings. If you do this exercise with a colleague who has administered the same form in his or her class, you can exchange examples of behaviours that lead to high ratings.

Remember that a low rating on a particular item or factor signals the need for further investigation NOT hasty judgement and action. In deciding what to ignore and what to consider, take into account the goals and nature of the course, nature of the students, whether the course is required or is an elective, level of the course, and your teaching style.

Analysing the Scores

Learning/Academic Value

A high score for this factor suggests that students are challenged and stimulated, consider their learning in the class to have been worthwhile, believe their interest in the subject has increased and are conscious of having understood the subject matter. Overall student who rate the course highly are expressing feelings of accomplishment on challenging learning tasks.

Instructor Enthusiasm

A high score in this factor suggests that students' interest and attention has been aroused, that their enthusiasm for the subject has been increased and that they are more motivated to learn. Students whose interest in and enthusiasm for a subject are aroused are likely to have better learning outcomes.

Organization/Clarity

A high score in this factor suggests that students believe the course was well structured, class materials were well prepared, and lesson objectives were reached. Students believe the instructor has a good knowledge of the subject and explanations were clear.

Group Interaction

A high score in this factor suggests that students found social interaction beneficial to the learning process. They were encouraged to ask questions and to give answers, and to share ideas and knowledge with one another.

Individual Rapport

A high score in this factor suggests that students have found the instructor approachable and accessible. It also suggests that students feel welcomed, reinforced, and encouraged.

Breadth of Coverage

A high score in this factor suggests that different theories were contrasted, background ideas and concepts were provided and different points of view and current developments were discussed. Students' knowledge and understanding of the subject was, therefore, increased.

Assessments/Grading

A high score in this factor suggests that students perceive the assessment to be fair and relevant and that feedback received was valuable.

Assignments/Reading

A high score in this factor suggests that students found readings and assignments valuable and that the learning experiences were meaningful and contributed to their understanding of the subject.

Overall Rating

A high score in these items suggests that the class and lecturer compare favourably with other classes and other lecturers in the institution.

Written Comments

Read written comments carefully to pinpoint specific complaints or suggestions for improvement. Then determine whether the complaints are justified. If the worry is legitimate, identify specific steps you can take to address the weakness. Keep in mind that students give few detailed suggestions on how to improve a course; they are better at spotting problems.

REMEMBER! Student ratings should be interpreted within the context of the class.

USING YOUR RESULTS TO IMPROVE TEACHING AND LEARNING

1. COMPARE self- and student ratings. Identify items that differ substantially and think about why this might be the case. Give particular attention to determine discrepancies between your importance ratings of the factors and student ratings.

2. EXAMINE the pattern of students' responses on specific items, not just the global ones. Take a closer look at items on which views vary widely. Why might this be so? Where written comments are provided read the comments to understand more. Now with this information, identify areas of strengths and weaknesses.

3. ANALYSE ratings further. Determine whether items receiving high or low ratings are related. Of course, do take into account the characteristics of the subject and students to explain response patterns for each group.

4. PLAN for improvement. In light of your reflection on the results select two or three areas, in the first instance, to target your improvement efforts. You might choose the factor about which you had the most difference between your self-ratings and student views. Decide on actions and strategies to enhance your teaching in conjunction with discussions with colleagues and information contained in the ***Teaching Ideas and Practical Strategies*** (TIPS) leaflets.

5. IMPLEMENT your plan and monitor the effects of changes by getting further information and feedback on your teaching by talking with colleagues and collecting additional student ratings.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they selected no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 1: Learning/Academic Value

In essence this factor denotes subjective feelings of success obtained through participation in a course and/or at the hands of a particular teacher. Students who are challenged and stimulated, who consider their learning through the course have been worthwhile, whose interest in the subject was increased, who are conscious of having understood the subject matter and who generally rated the course highly are clearly expressing feelings of accomplishment on challenging learning tasks.

Teaching for Learning and Academic Value

1. Take into account what your students want to know.

Being aware of the students' expectations for a course is a critical prerequisite for obtaining their attention, interest and participation, according to a number of outstanding lecturers in various disciplines. "It is important to be seen to be aiming to meet their needs, rather than simply following a checklist of things to be taught" say a lecturer in Physics.

2. Assign "minute papers" at the end of some lectures.

Several outstanding lecturers have found that asking just two or three open-ended questions after some lectures is a very effective means of establishing what students are understanding, and what their preferences are regarding content and method. Some also found it useful in allowing students to see what they have achieved in a session

The Physics teacher who invented this process says, "Several times during the term – about once a week – I come to class a little early. I write two questions on the board:

a. What's the most significant thing you learned today?
b. What question is uppermost in your mind at the end of today's class session?

"Then, I make my presentation for 49 minutes. One minute before the end of the period, I say to the class, 'Take out a piece of paper. You have one minute to answer these two questions.' My students sign the papers and pass them to the centre aisle. I pick them up on my way out of class. I give them to my reader to check off the names of those who turned them in. Now, I read the papers. I find I can tell whether I am getting my points across. I can tell what problems students are having. I clarify difficult points next time we meet. I can identify students in trouble early in the term. If a student gives me off-the-wall responses, I invite him/her to come see me. One of the common problems is that a student has the course prerequisites on paper, but not in his/her head. Furthermore the 'minute paper' process causes students to listen more actively. All the way along during the class session they are saying to themselves, 'is this the most significant thing, I'm going to learn today!' Toward the end of the hour they have to wonder, 'Well, what question is uppermost in my mind at the end of this session?' Students' writing improves. Responses I get in the last week of the term are more articulate, and longer than those at the beginning."

3. Touch base repeatedly with the fundamentals or basics.

Students like to be challenged, but they need to feel confident and well-prepared to meet the new challenges, and that often means consciously retreating a little to reinforce the foundations on which new material is based," according to an outstanding lecturer in Education.

One Engineering lecturer believes that too much of science and engineering is presented to students in a rote, plug-in-the numbers way.

"There are thousands of formulae," he points out, "but all of these are variations of a limited number of basic ideas or theories." "These basic ideas are 'ideal theories' from which are derived all the 'approximate' or 'technical theories' which engineers use."

"I try to teach my students how to judge when they can use an approximate theory with confidence and when they are obliged to go to a more rigorous level. In this way, I keep touching base with the fundamentals to reinforce students' understanding of them."

4. Stress the most enduring values or truths in your discipline.

"I stress the permanent values in literature, the emotional responses that a particular novel or collection of novels elicits from us all," says one English lecturer. "I try to get my students to understand why they respond to a given novel in the way they do."

After a class has discussed how they feel about a novel, the common emotions it arouses – he tries to lead them to analyse, understand, and explain why nearly everyone feels the way they do. He poses questions such as: What must literature be like in order to get us to respond the way we do? Why does a particular novel affect everyone in the same way? "Behind all my questions is the search for a way of analysing and discussing literature that will explain the most with the fewest assumptions."

5. Confound yourself, and let your students "rescue" you occasionally.

Asking open-ended questions, which the students can sense, are mystifying to you personally is a great way to encourage relevant discussion and to model your enthusiasm for discovering the secrets of the subject.

A distinguished lecturer in Education reports that "When I ask myself a question that initially seems puzzling to all of us, it lets my students know that I'm not omniscient, and that it's all right to ask questions or get confused occasionally. The class really responds to that."

APPENDIX L

6. **Get to know your students: Where they are "at"; and what they relate to.**

Knowing your students is important for a number of reasons. Several outstanding lecturers stressed that new learning must begin from what student are already familiar with. "Otherwise they quickly become confused, disinterested or anxious," a lecturer in Education explains.

A Physics lecturer noted that students will work harder to solve a problem that appears to be relevant to them. Rather than restricting problems and issues to assignments questions, she delivers many of her lecturers around puzzles and quandaries.

Getting to know what gets the class "fired up", or what they relate to is a strategy that several lecturers have recommended. This requires establishing good rapport and making time to chat with students in non-teaching situations.

7. **Encourage your students to form small study groups and send representatives to see you about difficulties their groups are having.**

One Humanities teacher who does this say, "Although I encourage my students to come see me about problems they are having with my course, first year students are often loathe to do that. By encouraging them to form study groups, I am trying to help them get to know at least some of their fellow students and to take advantage of what they can learn from one another. Also, it seems to be easier for some students to come to me for assistance if they 'represent' a group, because their problems are then seen as common to many students not just the group's representative. Staff members can be very intimidating for some first years, even those of us who try very hard not to be. Also many of these students were at the top of their high school classes and it is difficult for them to adjust to the competition at university. While it is difficult for them to admit that they don't understand something, there is a certain comfort in knowing that some of their fellow students are in the same boat and that by joining forces they can help one another."

8. **Schedule an individual appointment with each student.**

An outstanding lecturer in Education stressed the importance of knowing and treating students as people, rather than simply as students. "This is central to making the material relevant, opening up discussion, and generally meeting their learning needs," she explained.

A lecturer in Statistics felt that he was not being successful in generating class discussion. At the end of the third week, still unable to encourage class participation, he decided to pass around a sheet of paper with a list of 10-minute blocks of time when he would be available for individual appointments.

Each of his students was required to sign up for one of the 10-minute appointments. They were told that the chief purpose was for him to get to know his students better and to listen to any complaints or suggestions they might have.

"I found that this was a real ice-breaker," he explains. "Even though most of our discussions were mainly chit-chat, some of my students used the opportunity to indicate problems they were having in the course or to make suggestions about course improvements. Perhaps the chief benefit was that it gave me an opportunity to get to know my students. As a result, they seemed to feel more comfortable asking and answering questions in class.

9. **Assign "thought problems" which are typical of the problems faced by professionals.**

A Forestry teacher assigns weekly "thought problems" which are the same type of questions professional foresters are asked, such as, 'What is killing that tree?' not 'Name six factors which can kill trees."

Using real-life problems to encourage thoughtful reflection and/or discussion in this way, rather than requiring solutions in the form of assessable assignments, can be a particularly useful way to avoid overwhelming students with the complexities inherent in such tasks.

10. **Have your students keep a journal of their learning experiences during the course.**

A journal can be a very effective way to facilitate student' reflection on their own learning during a course, leading to greater understanding and appreciation of the subject. It is important, however, to ensure that students have been familiarised with the process of journal writing and the benefits they can expect from it.

This may be required as an assessable project, or simply recommended as effective preparation for class discussions, presentations.

11. **Begin each lecture by letting your students know what you are going to talk about and why.**

An Engineering lecturer refers to this as his "battle plan". At the beginning of the hour, I give my students a battle plan so they know where the discussion is going and can follow it more easily," he says. "For example, I tell my students that I'm going to discuss such-and-such a topic for the first twenty minutes, show them how to use it in the next twenty minutes, and them take questions in the last ten minutes. By laying out exactly what I am going to do, I eliminated a lot of student confusion. I don't want students spending the hour wondering, 'Why is he talking about that?' or 'What does that have to do with anything?' instead of concentrating on what I have to say."

12. **Assign provocative or controversial topics for papers.**

"I find that the quality of the papers I get often depends on the quality of the assignment I give" says a teacher of Business Administration. He tries to give provocative topics as paper assignments. For example, in a recent assignment he asked his students to respond to the question, "If you were working in a company that illegally pollutes the environment what would you do and why?" Giving provocative, assignments not only challenges his students and makes for more interesting reading but also diminishes the chance that the papers will be plagiarised.

One lecturer who successfully engages students this way warns that it is important, even when deliberately trying to be provocative or "realistic", to choose topics that the students are "ready" to deal with in the context of the material being covered.

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university teachers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they selected no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 2: Instructor Enthusiasm

A minimal condition for learning is that attention be aroused. Stimulus salience, that is, the extent to which a stimulus stands out against a background, is known to be crucial in evoking interest and attention. It is to be expected, therefore, that teachers who impress students with their enthusiasm, dynamism and energy and who make judicious use of humour will have students who are interested and attentive. Moreover, teacher enthusiasm can vicariously induce enthusiasm for the subject in students. Students who rate their teacher highly are more likely to model their behaviour towards a subject upon that teacher. Thus, the latter's enthusiasm can be acquired by students. Students whose interest in and enthusiasm for a subject are aroused are likely to have enhanced achievement in learning the subject. This factor is especially relevant to the principle that learners must be motivated to learn.

Teaching for Enthusiasm

1. "Open with gusto" and "Finish strong"

One lecturer points to the advantages of giving special thought to beginning and ending each lecture.

"The opening should secure students' attention and give them the desired mental set. Get off to a good start. Do something to command attention from the outset. Put punch into your opening."

Have some form of attention-getter...a gadget, or piece of hardware whose operation depends upon the principles of the day's lessons usually excites attention. Carefully planned questions or statements can also develop the curiosity necessary to get students' attention."

"The ending is as important as the beginning. Don't' let a class session fade into non-existence. Make an impressive ending. For example, end with a question for the class to cogitate and answer before the next meeting; a quotation conveying the essential theme; a summary; a miniature review (keep it brief); or what to do before the next class."

Giving students a strong sense of having achieved something worthwhile or useful by the end of a session has been recommended by several outstanding lecturers as an effective way to end a lesson and motivate students.

2. Focus on five or six different students each day and give your lecture as if you were talking to them individually.

Many speech teachers encourage people to think of a lecture as an enlarged or public conversation. Several excellent lecturers told us that they lecture to a large class (50-200 students) in the same way they talk to a few students.

"By focusing on a few student, I am more relaxed and informal. This helps me to concentrate more on the ideas I want to convey than the impression I may be making," one teacher said, "I think that, as a result, I speak with more expression and conviction."

3. Exaggerate everything about your presentation in a large auditorium class.

A teacher of Economics believes that physical exaggeration and a bit of hyperbole are keys to successes in lecturing in a very large auditorium. "You have to remember that 800 students constitutes an audience, not a class in the normal sense," he points out.

"In front of a very large audience, everything you would do in lecturing to a class of 30 or even 100 looks small, stiff, and formal. You have to exaggerate everything, make it all 'larger than life,' if you want to capture an audience.

4. Begin class with an incident, example, or anecdote to get your student' attention.

An attention getter does not have to be "gung ho" or "whiz-bang" – carefully planned questions or statements which are provocative, controversial or paradoxical can be quite effective.

A lecturer in History says that he often begins by reading aloud a short passage from a primary source or a story to illustrated his major theme or point in the lecture. " For example, I start out by stating that the Wizard of Oz is a parable for progressivism and read passages from it to illustrate this major thesis. I then get students to help identify the different characters and what they represent. I usually end with a quotation that pulls together what I have been trying to say," he says. "Also whenever possible, I try to link the past with current events, to show how the topic is important for the present."

5. Make diagnostic and practice audio tapes.

Although the audio quality of most home tape recordings is not good enough to diagnose fine points of pitch, inflection, articulation, and pronunciation, it can be used effectively to note whether you speak too slowly or too rapidly, whether you vary your tone and inflection sufficiently to hold your students' attention and communicate meaning, and whether you articulate clearly and/or forcefully enough to be heard and understood.

Audio tapes can also be used to check the organization of your lecture presentation and the clarity of your explanation. Repeated audio taping will also allow you to monitor your improvement on any of these variables.

6. Colour code your lecture notes with cues to "slow down", "pause and get attention", "demonstrate with gestures", or other stage directions.

One of several lecturers who use this says, "Because I have a tendency to speak too rapidly, I find these colour codes helpful as cues to slow down when introducing new ideas, explaining a concept, or summarising major ideas and the relationships between them. This also frees me to speak at

my own norm fast clip when making transitions or giving examples," he says.

7. **Use dramatic pauses and repetition to draw students' attention to the main ideas.**

Several teachers stress the need for repetition (using different examples) to communicate the most important points in their lectures.

Dramatic pauses are another way to highlight important ideas. A History teacher says that she used to tell her students, "The main point is..." but in the matter-of-fact manner, almost as an aside. "I discovered that many of my students did not get the message," she explains. "Now I indicate a main point by pausing to get my students' full attention and then saying emphatically, 'this is the really important consideration!" then I pause again to be sure they are prepared to write it down. If not, I restate the importance of what is to follow."

8. **Videotape a segment of your class.**

Several teachers have had their classes videotaped. One Zoology teacher has had is lectures videotaped many times. "The first time was a shattering experience," he says, "but it is the most effective kind of feedback you can get. I have found videotape invaluable for getting rid of annoying mannerisms, for learning to vary the speed of my delivery and to put more expression and greater clarity into my explanations."

9. **Develop effective ways to encourage students to see you about their difficulties.**

"Enthusiastic teaching is reflected not only in how you relate to and represent the subject matter, but in how you related and respond to the student," says a lecturer in Health Sciences.

Lecturers rated highly by their students as most welcoming and accessible, used means such as scribble boards, or notepads on doors as a means of "attracting reticent students to see them, and of proving a more relaxed, informal atmosphere. Such strategies can also help to promote a higher level of enthusiasm for the subject.

10. **Take care to communicate your genuine concern for students after class.**

"Students can be very sensitive to non-verbal messages implying that you are not genuinely interested, and this can quickly turn them off seeking help or pursuing an interest in the subject," warns an outstanding Physics lecturer. "Some lecturers seem to fear that any further encouragement of their students to drop in would leave them inundated. It is actually quite easy to learn how to avoid negative rejection of student's request, without devoting your entire day to them."

She explained that she made a point of never making students feel unwelcome. If a student dropped in at an inappropriate time, she would maintain a positive attitude, saying, for example, "I'd love to see you – how about 4:30?" rather than "I can't see you now, I'm busy – try again later."

11. **Begin or punctuate, your lecture with a "joke of the week," especially in large early Monday morning classes.**

One lecturer admits that his jokes are pretty bad, but finds that his students appreciate his efforts anyway. "I hate Mondays and I hate early mornings even more," he explains. "A joke related to the course content, to education, or to life in general tends to help get everyone awake," he says.

The source of his jokes? "One source is the students themselves," he says. "I encourage students to bring me jokes I can use. In that way my 'bad' jokes are their 'bad' jokes as well."

12. **Invite guest speakers to your course.**

An English teacher sometimes invites professional actors to talk about their interpretations of a scene or a role from a play his students are studying. "It's very important to make clear to a guest what you expect of him or her in order to ensure that it is an educational experience for your students,"

An Architecture teacher prepares his guests well in advance so that they know exactly what is expected of them. "Practising architects are asked to submit working drawings, models, photos, and publications on one of their buildings so that my students will be well acquainted with their work beforehand," he says.

"Students are asked to submit a set of questions to a guest speaker beforehand about his/her work. Designated students are given responsibility to see that the questions are addressed to the speaker."

13. **Vary the pace and type of instructional activities in course.**

One excellent teacher says that he conducts each class meeting differently "to keep my students off balance. Students always know what topic will be covered in a given session, "he says, "but they don't always know how it will be handled."

An English teacher also believes that his wide variety of teaching strategies accounts for his high ratings on interesting style of presentation. "I read whatever I can find on teaching in my discipline," he says, " and I borrow shamelessly from other instructors when it comes to pedagogical strategies."

Some of the variations used by excellent teachers include: student panel discussions, guest speakers, slides, films, overhead transparencies, blackboard work with coloured chalk, role-playing and simulations, and a wide variety of group discussion techniques.

14. **Focus your lectures around a common object, event, or phenomenon which exemplifies the major concepts of the course.**

A lecturer in the Biological Sciences calls this his "potato lecture." "Biology is an empirical discipline; it depends on observation and investigation. I pass out potatoes to all 700 students in the class and begin a Socratic dialogue about the kinds of things they can observe about their potato. I have to overcome their previous experiences," he explains. "Although potatoes are familiar objects to them, they don't have the foggiest idea what is a potato is. I stress what you can get out of everyday experiences by asking the right questions. I poll them on their observations, help them ask questions and describe ways they could investigate answers." An interactive exercise around a common phenomenon tends to "break the ice" between faculty and students even in a large lecture course.

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and Davis, B. G., Wood, L., & Wilson, R. (1993). *Berkeley Compendium of Suggestions for Teaching with Excellence.* University of California, Berkeley.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they select no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 3: Organisation/Clarity

The essential ingredients of this factor are structure and clarity. By cueing learners about the organisation of subject matter, by providing advance organisers, by scheduling student exercises and assignments appropriately and by inducing appropriate cognitive schemata teachers assist students' memory retrieval and formation of linkages between new material and material previously learned. These principles of teaching and learning are time-honoured and widely accepted elements of information processing theories of learning. While clarity is clearly an expected outcome of careful preparation and good organisation, it can be important as a correlate of teacher knowledge of the subject, with teacher uncertainty producing vagueness which inhibits student understanding. Students who perceive instruction to be well organised and clear are, thus, likely to enjoy enhanced knowledge and understanding of course material. The Organisation factor is pertinent to several accepted principles of teaching and learning.

Teaching for Organisation and Clarity

1. **Empathise with the students' difficulties in learning the material for the first time.**

   "It is important to distinguish between appreciating the difficulty students have in understanding new material, and the rather simpler but less effective option of allowing the subject difficulty to act as an excuse for the lecturer's quality of teaching or the students' quality of learning," according to an outstanding lecturer in Education.

   A faculty member in the sciences says that he noticed that he had taught the course better the first time than he did the second time. "When I asked myself why, I realised that in preparing the course for the first time, I really had to work hard to master certain parts of the material in order to explain it to my students. The next time, however, these concepts no longer seemed difficult to me. Unfortunately, I forgot that they would still be difficult for the students. Now I colour code all of my lecture notes, keying the parts that students are likely to find difficult and making a special effort to make points very clear."

   A Physics teacher also tries to put himself in the students' shoes. "After I have finished writing up a set of lecture notes," he say, "I review them carefully, asking myself: 'What might my students find hard to follow in that line of reasoning?' 'What examples might make that more clear?' this has now become the most important part of my lecture preparation.

   Several faculty members report making notes to themselves of explanations that worked well and those that didn't. They also keep records of the kinds of errors students most commonly make in assignments and exams as a reminder of what students find most difficult to understand.

2. **Keeping a journal.**

   One History teacher has found it very effective to keep a brief journal or diary for each course. "After each lecture, I jot down a few notes about how the class went: explanations and examples that worked well and those that didn't, students' difficulties with the text, techniques for generating discussions, and so forth. If something went very badly, I correct it at the next meeting. For the most part, however, I keep the journal to help me improve the course next time."

   Although journal of this type could be beneficial to any teacher, its value is greatest for new instructors or for faculty members teaching a new course or a course they teach only every few years.

3. **Tell them what you are going to tell them; then tell them what you told them.**

   Although it may appear to be an over-simplication, many excellent teachers cite the old adage, "tell them what you going to tell them; then tell them what you told them." In the case of lecturers on complex subjects, the general principle is a good one which can be adapted to major topics within a lecture as well as to the overall lecture itself.

4. **Use lots of concrete or memorable examples.**

   Most excellent teachers agree that the choice of examples is very important, favouring those that are anecdotal, personal, or humorous because they find that students tend to remember these best. "I use concrete examples wherever possible," says an Anatomy teacher. "For instance, I describe a particular body organ by comparing its size or texture to an object familiar to students, like a walnut."

   An Economics teacher also places importance on using concrete examples of interest to students. "I use specific examples whenever I can. In talking about inflation and price controls I'll use the Nike sneakers or Sony Walkmans rather than apples or a general product."

5. **Begin and end your lecture or discussions with a summary statement.**

   "Students crave both continuity and sense of closure," one lecturer explains. "They do not like unfinished presentations. At the same time, because none of us likes repetition, I try hard to use different words and examples in each summary. The best way I have found to avoid redundancy is to note on an index card the exact words I have used at the end of a lecture, so that I am reminded to vary them in the brief recapitulation I give at the beginning of the next class meeting."

   A teacher of Business Administration also uses this technique. "Because each concept in this course builds upon what has gone before, it is important for students to see how each new topic relates to what they have already learned as well as to what they will be learning in the coming weeks. I find the most effective way of doing this to begin with a brief

APPENDIX L

summary of what came before, followed by a brief preview of what will come next."

6. **Explicitly call attention to the most important ideas in each lecture.**

"I began to emphasise the main points about ten years ago," says one Political Science teacher, "when I discovered that you can't rely on undergraduates to intuitively know what the most important points are. You have to tell them."

Faculty members in several disciplines stress the need to call students' attention to the most important ideas being presented. Some teachers announce the importance of an idea before presenting it, saying such things as "This is really important, so you have to be alert." Other teachers emphasise the most important things to remember is..." or "This is so important that everyone of you should have it engraved on a gold plaque and hung over your bed!" as one teacher of Computer Science puts it. "There is no point in my students having to guess what is important if I can tell them," he says.

7. **Rework completely your lecture notes each time you teach a course, particularly if you are in a rapidly changing field.**

"It's important to completely redo my notes each time I teach the course," says an Economics teacher. "In this way, over a period of six to eight years, they change quite radically. This is partly because the field is changing, but it is also because my own ideas continue to develop."

Although the myth of the teacher who teaches with yellowed and musty notes is almost unheard of in a major university today, the importance of re-creating lecture notes each time a course is taught – even if back-to-back within the same year – was stressed by nearly all excellent teachers as way of keeping themselves fresh and interested as well as interesting to their students.

8. **Prepare clear, interesting and uncluttered overheads or other visual displays to enhance and clarify your presentation.**

A number of award-winning lecturers stressed the importance of clear, uncluttered overheads. A lecturer in Music notes that overheads should never contain more than three or four succinct points or ideas. "Colour and space are the key ingredients in making the information on overheads accessible to students," she says.

A lecturer in Chemistry prefers to prepare overhead transparencies by hand to give him more freedom to be creative, so that students maintain attention and interest. "When students can easily see the point of a transparency, they can follow the logic of a lecture. The whole presentation appears to be clearer and more organised, " he explains.

9. **Rephrase explanations of major points several times.**

"Repetition leads to learning," one Science teacher says. "I repeat major points several times from a different direction or in different words."

"No single explanation will be clear to all students," points out a teacher of Business Administration. "By using different language or different examples, I maximise the chances that every student will eventually understand."

An Engineering teacher reports that he develops the same point in two or three different modes, for example, mathematically, verbally, and graphically.

10. **Acknowledge the difficulty of concepts students are likely to find hard to understand.**

"Acknowledging difficulty avoids the risk of belittling the students' efforts in mastering the concept, or the students themselves if they do not master the material easily," according to an exceptional Chemistry lecturer. "It is important to admit to the difficulty of understanding material for the first time, but not to make that difficulty an excuse. A good way of achieving this aim is to offer a special 'strategy' for mastering the material, such as '...so listen carefully...', or '...so remember this simple example'".

One Engineering teacher says, "I consciously cue students to the most difficult ideas by saying such things as, 'Almost everyone has difficulty with this one, so listen closely.' Because the level of students' attention varies throughout the hour, it is important to get everyone listening carefully before introducing a new concept or explaining a difficult point".

11. **Take your students' perceptions into account when using and assessing your organisational structure.**

A number of lecturers were initially surprised by feedback from students that suggested deficits in organisation. "I realised, in consultation with my students, that while my organisational format made sense to me, it was confusing to them because other related subjects were organised according to different framework," one lecturer explained. "We – the students and I – decided that our format was fine; I simply needed to take more care in applying it to explain what we were covering and how different parts of the subject related."

"The moral of the story is, don't assume your students have the same sense of your logic that you do!"

12. **Begin each class period with a brief summary of the main points covered in the last meeting and then call for students' questions.**

The advantage of summarising and asking questions at the beginning of a class period is that, " students are fresher and after a brief recapitulation, they are more likely to realise and acknowledge if they have any problems," as one teacher puts it. A variation of this technique is to summarise and call for questions whenever they is a major transition from one topic to another within the same lecture.

13. **Use the blackboards to help you summarise.**

Several excellent teachers stressed that they plan their blackboard work carefully so that the most important concepts are still visible at the end of the hour and can be used in making a summary.

"I consciously attempt to write clearly and legible and to be sure that my board work is organised and is visible to everyone," one Engineering teacher says. "At the end of the class, I use this board work to go back over important theorems or equations, underlining and boxing in with coloured chalk important concepts and steps."

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they select no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 4: Group Interaction

Learning in institutionalised educational contexts is a social phenomenon. That is, except in rare cases of individual tuition, instruction is given to groups of students ranging from small to very large in size. This factor refers to verbal interaction in classrooms in the form of questions and answers facilitating the expression and sharing of ideas and knowledge. Higher ratings on items comprising this factor suggests that the motivational potential of social interaction with others in learning contexts is being capitalised upon and also that the classroom context is being exploited as a venue for activity in practising and testing ideas and obtaining feedback. As such the Group Interaction factor has a strong basis in principles of teaching and learning.

Teaching for Group Interaction

1. **Face openly reservations or fears you may have about encouraging group discussion and interaction.**

Many lecturers feel uncertain about the utility of class discussion or feel uncomfortable about implementing or keeping control of it.

"It is important to recognise potential problems and limitations involved in a discussion format, without rejecting it as an option," says an outstanding Economics lecturer. "Any strategy has potential weaknesses. The decision regarding whether to encourage more discussion with and between students should be based on a careful search for ways to maximise the utility of discussion, and to minimise any limitations – even if that does sound like a cold economist's approach!"

Potential concerns include covering less material, getting "side-tracked", and causing confusion among students who experience difficulties in understanding the purpose, relevance or flow of the discussion.

2. **Plan your sessions to incorporate time for discussion.**

Allowing time for discussion or other group activities often means that fewer points can be addressed in a given session. Acknowledging this in advance, a number of lecturers set out-of-class readings and assignments materials to cover additional content.

"I was initially reluctant to assign material which I had not specifically covered in lecturers, but I quickly found that when students know in advance what was required, and how the discussion related to the content, they were highly motivated and productive, both during the session and their readings and assignments," a lecturer in Physics revealed.

3. **Get to know your students: Where they are "at"; and what they relate to.**

Knowing your students is important for a number of reasons. Several outstanding lecturers stressed that new learning must begin from what students are already familiar with. "Otherwise they quickly become confused, disinterested or anxious," a lecturer in Education, explains. "Students will also open up more in class discussion if they feel a comfortable rapport with the lecturer."

Getting to know what gets the class "fired up", or what they relate to is a strategy that several lecturers have recommended as a means of generating a vibrant learning atmosphere. This requires establishing good rapport and making time to chat with students in non-teaching situations.

4. **Force yourself to lengthen your "wait time" after questions and after answers.**

A lecturer in Education points out that it is well worth waiting an extra few seconds before continuing to speak after asking a question or after a student contributes some to the class.

It takes a conscious effort to wait beyond the normal 'comfort zone', she warns, "but my experience has confirmed research which shows that both lecturers and students make great mental use of such a pause – at least three seconds – to work through what is being said or asked, leading to better quality questions, answers and discussion in general."

"Students in Mathematics are often not quite confident enough to offer answers, or even ask for clarification, for fear of sound ignorant," explained another outstanding lecturer. "By pausing for an extended period, and looking around at the faces in the class, students are more likely to decide to 'bite the bullet'. Sometimes it gives them a chance to formulate their question to their satisfaction, sometime they sense that I'm waiting to hear from them (which reinforces their confidence that their question is appropriate), and sometimes they sense that I won't go on until I've heard from someone. It can be a little uncomfortable for all of use occasionally, especially early on, but the gradual growth in confidence and competence shown by the students is definitely worth the effort".

5. **Use students' written assignments as the basis for discussion.**

An Engineering teacher identifies several key questions or issues, which he gives to his students a week or two before they are discussed. His students prepare written responses of no more than one typewritten doubled-spaced page. As a result of writing their answers, students come to class well prepared to discuss the material. Their written responses are turned in at the beginning of the period and are subsequently graded, as is their participation in the discussion of the topic.

A teacher of Business Administration uses the same approach throughout the term. Each week a "reaction" paper is due which requires his students to write one to three pages on a specific topic, typically responding to a controversial issue. The papers are graded and used as the basis for class discussion.

6. Reserve the last ten minutes of your class for questions.

A faculty member in the humanities wanted to provide opportunity for students questions during his lectures, but he was concerned that the questions might monopolise class time and take them off the topic. "I decided to reserved the last ten minutes of class for students questions," he says. I feel better knowing I will not be interrupted. My students feel better knowing they have an opportunity to clarify points they may not have understood."

7. Understand why students repeat the same questions.

"I used to be impatient with a student who asked a question which had been asked and answered earlier," says a teacher of Computer Science. "Only after several years did I come to understand that such students are not necessarily stupid or inattentive. I learned instead that a student can only ask a question after the material has registered with him, after it begins to make sense."

"Although one student may have asked the question the day before, other students may not have `heard' (i.e. understood) either the question or the answer. Only later, when the material `clicks,' does that same question become meaningful to for them. Indeed it appears as a `new' question for them and they are now receptive to the answer.

"I try to keep this in mind and patiently answer all relevant questions. I try to use different language or different examples, hoping that this will make it clear without boring those who grasped the idea a day or two earlier."

8. Begin discussion with questions based on common experiences.

Students often feel more comfortable talking about an experience they have in common: a field trip, a slide show, a demonstration, a film, a book, an exhibit, etc. A shared experience can stimulate good discussion because, as they exchange their observations, students frequently discover that they have different perceptions and reactions to the same event. The discussion can then focus on how and why perceptions vary.

An English and a History teacher both apply this technique to their courses. "I like to begin my discussions with a question all students can answer," explains the History teacher, "usually dealing with how students felt about the reading." The English teacher begins discussion by asking students' reactions to the novel.

9. Use brainstorming as a technique.

Brainstorming is a method that can be particularly effective in getting students to consider all of the possible causes, consequences, solutions, reasons or contributing factor to some phenomenon. The rules are very simple. Students are encouraged to contribute ideas rapidly and each idea is written down on the blackboard. During the formation of the list no idea is to be questioned or criticized by any member of the class. Spontaneity and inventiveness are to be encouraged. Only after a set period of time (ten minutes, for example) or when the group has pretty well exhausted its ideas is an analytical or critical discussion of the ideas permitted.

"Posting" is a variation on "brainstorming" in which two or more columns are labelled on the board. These might be "pros" and "cons" of an issue or "possible causes," "consequences" and "interactions" of a phenomenon or event. Again, criticism of ideas is postponed until a later period to encourage spontaneity and creativity.

10. Keep notes during discussion.

Some teachers find it useful to keep a clipboard handy during discussion so that they can jot down notes. As the class is discussing a topic, one Education teacher makes notes about important points, confusing concepts, or ideas that may have been overlooked in the discussion. At the end of the period, he makes a brief summary of the topics discussed, reinforcing the main points, and clarifying or elaborating as appropriate.

A teacher of Engineering employs a similar strategy, but he interjects his comments during the course of discussion. "I summarize and make remarks that will get the discussion back on track, or I shift the discussion from an issue that has already been adequately dealt with to a new one."

11. Divide the class period into blocks of time, one of which is a discussion segment.

"I found it boring when I was a student simply to listen to a teacher talk for an hour and a half," says a faculty member of Ethnic Studies. "So I try to vary the class activities by dividing the class period into three segments."

For the first 20 minutes of class time, he builds up to a discussion question by presenting evidence, facts or issues. The next 30-40 minutes is devoted to student discussion even though the class has several hundred students. The instructor asks students for possible explanations or interpretations of the facts or issues presented in the first part of lecture.

The last 20-30 minutes of class is spent analysing the discussion and bringing the topic to a conclusion. Finally he ends the period by posing a question which students are to think about before the next class meeting.

Incorporating discussion into large lecture classes takes careful preparation: the questions posed to students need to be identified in advance and their responses anticipated in order to ensure a productive discussion. Nevertheless, this approach is very effective for engaging students' interest and encouraging analytical thinking.

12. Having students complete a brief opinion questionnaire and using the results as a basis for discussion.

A faculty member of Business Administration has found this approach to be particularly effective. "The first seminar session generally begins with a questionnaire asking for opinions on a variety of issues that will be covered in the course. Each week we begin by analysing the questionnaire results on the relevant topic and talking about the views of political economy revealed by the students' answers," he explains.

This device is very effective in starting discussion and helping students (and the teacher) get to know one another's views.

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they select no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 5: Individual Rapport

Opportunities to provide for individual differences in capacity and to take account of learners' present knowledge and attitudes in higher education depend heavily upon individual contacts with instructors. Furthermore, individual tuition and guidance are available to the extent that instructors are interested in and accessible to individual students. Students who feel welcome also have greater access to motivationally significant opportunities such as face-to-face reinforcement and encouragement.

Establishing Individual Rapport

1. **Acknowledge the importance of getting to know your students: where they are "at"; and what they relate to.**

   Knowing your students is important for a number of reasons. Several outstanding lecturers stresses that new learning must begin from what students are already familiar with. "Otherwise they quickly become confused, disinterested or anxious," a lecturer in Education, explains.

   Getting to know what gets the class "fired up", or what they relate to is a strategy that several lecturers have recommended. This requires establishing good rapport and making time to chat with students in non-teaching situations.

2. **Consciously use your students' name whenever possible.**

   "I call roll several times during the beginning of the terms to connect faces and names as soon as possible," a teacher of Forestry says. "Later, if a student looks familiar but I can't remember his or her name, I simply admit it and ask the student to tell me again. Then I make a point of using the name right away to help me remember it the next time." A teacher of Entomology says, "in a class of 100, there are always three or four name that I don't seem to be able to learn. Nevertheless, my students greatly appreciate the effort."

   Another strategy is to walk around the class while your students are working on a quiz or problem and try to match faces with names. A Science teacher says that he circulates for 10 or 15 minutes and then goes back to his desk and tries to write everyone's name down. "This really reinforces my memory," he says.

3. **Keep the hour following a class open to talk with your students.**

   Make a habit of staying after class to talk with your students. "The biggest turn-off for students is for a faculty member immediately to gather up his notes and his briefcase and virtually beat his students to the door after class," a teacher of Public Health points out. "This suggests that he is too busy for students. I have developed a technique of loitering after class, very slowly erasing the boards and talking with my students as they leave. The result is that after the first few days of class, more and more of my students linger as well, and I get to know many of them in that way."

   If another class is scheduled in the room immediately following your class, then do as a Biochemistry teacher does and tell your students that you will stay in the hall for ten minutes following lecture to respond to students' short questions.

4. **Take care to communicate your genuine concern for students after class.**

   "Students can be very sensitive to non-verbal messages implying that you are not genuinely interested, and this can quickly turn them off a course," warns an outstanding Physics lecturer. "Some lecturers seem to fear that any further encouragement of their students to drop in would leave them inundated. It is actually quite easy to learn how to avoid negative rejection of students' requests, without devoting your entire day to them."

   She explained that shed made a point of never making students feel unwelcome. If a student dropped in at an inappropriate time, she would maintain a positive attitude, saying, for example, "I'd love to see you—how about 4:30?" rather than "I can't see you now, I'm busy—try again later."

5. **Go to class before it begins.**

   A Physics teacher makes a point of going to his classes a half-hour early (if the room is vacant) to erase the board, check out the equipment and the demonstrations he will be using and write a brief review on the board (e.g. pertinent equations, key phrases, topic areas).

   "This activity gets me in the teaching frame of mind and refreshed my students about the important points we covered the last time," he notes. "It also has the intended value of increasing the opportunities I have to talk informally with my students. Five or six students come early to the class each time to ask questions, share ideas, or just talk.

6. **Do some of your own work in your campus office.**

   Several teachers do non-teaching work in their campus office with an open-door policy. "I tell my students that if the door is open they should feel free to come in and ask whatever questions they have," one teacher of dramatic art says. "On the other hand, if the door is closed, it means either that I am not in or I prefer not be to be disturbed."

   An Engineering teacher follows the same policy. He tells his students that even outside formal office hours, "If you catch me in my office, I'm fair game. This is my number one job, so I'm around the office a lot."

7. **Require all your students who do below pass level work on assignments or quizzes to see you.**

APPENDIX L

One teacher of Forestry does this in all his undergraduate courses. Another Forestry teacher writes a note, "Please see me" to students who score below 70 on his weekly quizzes. "It's important to find out why students score low," he explains. "If they are having difficulty understanding the material, I offer to help them. If it's a question of motivation or a student placing less priority on my class, that OK too. It helps me as a teacher to know the reasons for the poor performance. Show concern is also a powerful motivator for some students: they begin to do better."

A Zoology teacher concurs. "I call students in who get less than 50% on the biweekly quizzes," he says. "In a way, I play parent with them; I 'sit on' them a little. I think I understand better now than when I began teaching the need some students have for external motivation."

8. Schedule an individual appointment with each student.

An outstanding lecturer in Education stressed the importance of knowing and treating students as people, rather than simply as students. "This is central to making the material relevant, opening up discussion, and generally meeting their learning needs," she explained.

A Statistical teacher felt that he was not being successful in generating class discussion. At the end of the third week, still unable to encourage class participation, he decided to pass around a sheet of paper with a list of 10-minute blocks of time when he would be available for individual appointments.

Each of his students was required to sign up for one of the 10-minute appointments. They were told that the chief purpose was for him to get to know his student better and to listen to any complaints or suggestions they might have.

"I found that this was a real ice-breaker," he explains. "Even though most of our discussions were mainly chit-chat, some of my students used the opportunity to indicate problems they were having in the course or to make suggestions about course improvements. Perhaps the chief benefit was that it gave me an opportunity to get to know my students. As a result, they seemed to feel more comfortable asking and answering questions in class."

9. Arrive at class ten minutes early each day and talk informally with students.

"I try to target a different section each day," a History teacher says, "talking with students about the course or more general topics, getting to know their names and something about them as individuals. It helps me to remember a name if I can connect it with a place, an interest, a personality trait. An easy example would be Miss Baker from Bakersfield."

This teacher, like many others, believes that addressing students by name helps to break the excessive formality of a large lecture class and creates a more positive classroom environment.

10. Use index cards as mnemonic device.

As soon as you have a list of the students enrolled in your class, write each of their names on an index card. On the first day of class, call roll, laying the cards on your desk by seat and row to reflect where each students is sitting in class, refer to the index cards and use students' names whenever possible.

A faculty member who uses this technique finds it especially effective to return to her office immediately after class and lay out the cards in the same order and review the names. "I set a goal for myself of learning 5 names each time the class meets. With a class of 30, I find I can learn everyone's name within the first two weeks without any difficulty."

11. Ask students their names whenever possible.

Some faculty members find that learning students' names requires concentration and repetition. One Science teacher, for example, says, "I ask students their names at every opportunity: whenever one comes to visit me during office hours, whenever I see a familiar face in the hallways or crossing the campus, and whenever a student asks a question before or after class. Students are hungry for some recognition of their individuality, and they appreciate it enormously when I take time to learn their names."

12. Schedule specific topics for your office hours.

"I find it useful to identify in advance a specific topic for my office hours," says a Linguistics teacher. "I encourage my students who are having difficulty in that area to come for help." Based on past experience she knows which concepts and ideas cause problems and she schedules her office hours to provide further elaboration and discussion on these topics.

This way if one of my students misses a class or doesn't fully understand the topic, he or she has another chance at the material during office hours. My tutors are also encouraged to attend these sessions so that they better understand areas of student difficulty.

Another teacher uses one office hour a week in a similar fashion, although the specific topics are not necessarily ones covered by the course. "Sometimes they are enrichment topics; sometimes they are remedial, like how to do a term paper." He says.

As an added bonus, students and teacher get to know one another in a small informal setting.

13. Develop effective ways to encourage students to see you about their difficulties.

Lecturers who were rated by their students as most welcoming and accessible used cartoons, scribble boards, or notepads on door as a means of "attracting" reticent students to see them, and to providing a more relaxed, informal atmosphere. Such strategies can also help to promote a higher level of enthusiasm for the subject.

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they select no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 6: Breadth of Coverage

This factor reflects students' responses to items concerning the contrasting of implications of various theories, the provision of the backgrounds of ideas and concepts, the presentation of different points of view and the discussion of current developments. These all have to do with substantive qualities of instruction. Each would seem to have the potential to increase student knowledge and understanding through facilitating generalisation beyond the confines of the specific situation, to clarify the material to be learned and its meaningfulness to the learner.

Teaching for Breadth of Coverage

1. **Use your students' opinions to create a microcosm of society's attitudes on social, political, and economic issues.**

   At the beginning of the term, a teacher of Economics gives his students a questionnaire in which they are asked to agree or disagree with a series of controversial statements on the functioning of the economy. "Because the introductory course is so large (over 800 students), it is impossible to invite discussion even though many students enter the course with strong views about such matters as the causes and cures of inflation," he explains.

   "As a substitute for discussion, I use the survey results to introduce a variety of student viewpoints. Throughout he semester I reveal selected results from the survey as they related to new concepts or issues covered in readings and lectures.

   "This technique gives my students a sense of personal involvement in the subject matter. Students learn that some of their peers may share their viewpoint. They also learn that some of their peers don't share their viewpoint and why. Use of the students' data allows me to introduce most of the views currently reflected in the society as a whole."

2. **Present each of several competing theories as if you were an adherent of that position.**

   A teacher of Psychology introduces three major approaches or schools of thought in the field. "I discuss each one historically and contrast the basic elements and implications of each," he says.

   "I really don't have a point of view in this course. There is so little know with impartial certainty; I don't think one is justified in taking a position at this time. Therefore, I present the best case for each theory, then analyse each critically and comparatively."

   Even though they do not have a distinct point of view, several other excellent teachers report that they also present the best case for each of several competing theories before they reveal their own preferences.

3. **Encourage students to take an approach different from the one you have adopted.**

   A teacher of English uses this strategy in all of his literature courses. "I always approach literature from an historical point of view: history is a particular passion with me," he says. "At the same time, I point out that there are many other perspectives and encourage students to use alternative approaches, example, the psychoanalytic approach or that of the new literary criticism."

4. **Point out explicitly that there are alternative points of view.**

   "Indicate the polar principles which guide much of the research in the social sciences as well as much of our folk wisdom, e.g. 'opposites attract' versus 'birds of a feather flock together' or 'absence makes the heart grow fonder' versus 'familiarity breeds contempt'. In doing so I point out that they should be mindful that there may be good reasons to believe the opposite of what I say; that they should analyse all arguments of their opposites."

5. **Touch base repeatedly with the fundamentals or basics.**

   One Engineering teacher believes that too much of science and engineering is presented to students in a rote, plug-in-the numbers way.

   "There are thousands of formulae," he points out, "but all of these are variations of a limited number of basic ideas of theories." "These basic ideas are 'ideal theories' from which are derived all the 'approximate' or 'technical theories' which engineers use."

   "I try to teach my students how to judge when they can use an approximate theory with confidence and when they are obliged to go to a more rigorous level. In this way, I keep touching base with the fundamentals to reinforce students' understanding of them."

   Another Engineering teacher concurs. "Students typically are presented with 100 different equations in each course they take. They are exposed to 1100-1200 equations overall. Rote memorisation is futile; no one can remember that many equations. You have to point out over and over again that these 1200 equations are embedded in about 8 basic ones."

6. **Require your students to read journal articles.**

   "It's important for my students to be exposed to state-of-the-art ideas even in a lower division course," says one Political Science teacher. "I try to make sure that my reading list contains at least a few recent journals articles."

   "In some ways I find it easier to introduce recent development in the field to lower division students," says a faculty member in the biological sciences. "I do this by over-generalising. I translate the abstract of a journal article in layman terms. I present the basic findings in a narrative fashion, using little actual data."

I want my students to become excited by the open-ended nature of science. I want them to understand that what they are learning is not the final word."

7. **Require your students to read current newspapers or periodicals.**

A teacher of Economics assigns the Tuesday editorials of the *Wall Street Journal* each week. She uses them as a basis for discussion and for exam questions; she has her students compare them with textbook presentations on related topics.

8. **Tell your students about local events which will expand their understanding of your subject.**

"Every Monday I distribute a calendar announcing course-related events not only on the campus but in the area," one Social Science faculty member explains. "The events included dance troupes, plays, lectures, demonstrations, poetry readings and so forth. In this way the content of my course is expanded far beyond what I can actually cover in class. I also encourage my students to use these local resources in their research and writing assignments."

9. **Create opportunities for role playing.**

An Engineering teacher makes use of role playing to encourage his students to develop the broad range of skills they will need in their careers. "I give my students copies of an Engineering report, for example. Then one half of the class is asked to assume the role of the authors of that report and prepare an oral presentation for the client or funding agency. The other half of the class is assigned to act as representative of the client or funding agency and to prepare questions to be asked of the engineers.

"About a week later, during class time, I select certain students to actually enact these roles in front of the class. My students do not know ahead of time who will be called upon, so everyone has to be prepared. Those not called on join me in the role of observer. When the students have enacted the meeting, the rest of us give a critique of each side's performance."

10. **Assign provocative or controversial topics for papers.**

"I find that the quality of the papers I get often depends on the quality of the assignments I give," says a teacher of Business Administration. He tries to give provocative topics as paper assignments. For example, in a recent assignment he asked his students to respond to the question, "If you were working in a company that illegally pollutes the environment what would you do and why?" Giving provocative assignments not only challenges his students and makes for more interesting reading but also diminishes the changes that the papers will be plagiarised.

One lecturer who successfully engages students this way warns that it is important, even when deliberately trying to be provocative, or "realistic", to choose topics that the students are "ready" to deal with in the context of the material being covered.

The importance of getting to know what gets the class "fired up" and what they relate to, is emphasised by several lecturers in setting appropriately provocative assignments.

Lecturers also noted that such assignments are an excellent way to encourage students to consider a broader range of issues in their responses.

11. **Use real problems and have your student solve them.**

An Engineering teacher presents his students with problems based on real cases. "For example," he says, "my students are told that a ball bearing failure has occurred in an airplane. They are asked to outline what steps they would take in determining the cause and correcting it.

"They tell me what tests they would make and, using simulation techniques, I tell them what the results of those tests would be and ask what they would do next. This continues until my students have either solved the problem or are stumped. Then their results are compared with those from the actual case study.

"The value of this approach is to give my students experience solving the type of practical problems they will encounter as professionals," he explains. "Also, because the problems are based on actual cases, it gives my students a chance to compare their own problem-solving skills with those of practising engineers."

12. **Encourage group discussion to facilitate identification of related broader issues.**

A number of outstanding lecturers emphasised the importance of using group discussion or informal chats with student to identify links with broader issues which are of particular value or interest to the class.

"Questions or issues raised by students can help you identify points to emphasise or important links to make with other subjects being studied," a lecturer in Education explained.

13. **Invite guest speakers whose viewpoints differ from your own.**

A teacher of Education makes a point of doing this in his courses so that his students are exposed to a variety of positions. "I want them to understand what the different points of view are," he says, " and one of the best ways I have found to do this is to invite a colleagues or practitioner whom I know to be an adherent to each view to make a presentation to the class."

"I always take detailed notes during a guest lecture," says a teacher in the biological sciences, "In this way I am able to answer student questions about the material during later sessions and may learn something new myself!"

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training: and Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley.

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they select no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 7: Assessment/Grading

The instructional value of assessment and grading lies in the quality of the feedback and in the stimulus to study they provide. The items comprising this factor apply specifically to feedback and less specifically to motivational issues. Students' perceptions of fairness and relevance of assessment procedures are probably associated with their motivation to learn. However, this factor's main basis in principles of learning is reinforcement in the form of knowledge of results and effective consequence of that knowledge.

Teaching for Examinations/Grading

1. **Prepare test questions that are similar to those used in your quizzes, homework, or discussion.**

   "I try to generate exam problems that are similar to my homework problems so there are no surprises," comments a Mathematics teacher. "I also try to include problems everyone should be able to do (some very easy ones) as well as questions that require more thought and really make my students go beyond the material."

   "Questions on midterm and final exams should not take a form radically different from those which you use in quizzes, homework assignments, lecture or discussion."

   Several staff members stress the importance of showing exam questions to tutors or other colleagues before the tests are administered. One explains "tutors are very helpful in identifying test questions which may be too difficult for my students. They often see things that don't when I make up the exams."

   "When students can see a link between the things they are asked to do during their private study time, and the things they will be asked on the exam, they are more motivated to make the effort", a lecturer in Psychology said. "Particularly when there is a need to reduce the amount of assessable work, such an incentive ensures that students will appreciate the value of recommended readings or practice questions. Students also seem to rate the exams as being fairer and more appropriate when this link is clearly established."

2. **Balance the difficulty of test items.**

   A teacher of Business Administration distributes test items as follows: about 25% are reasonably easy questions that nearly everyone gets correct. About 50% of the questions require a little more sophistication but can be answered by students who have kept up with the course material. About 25% of the items are quite challenging and generally are answered correctly only by the 5-10% of the class.

   "A balanced test with easy, moderate, and difficult items gives my students an opportunity to show whether they have mastered the fundamentals of my course or have gone beyond the minimum," explains this staff member. "I try to give my students a feeling of satisfaction at the end of a course by providing them with an opportunity to express what they have learned, rather than frustrating them because what they have studied does not appear on the exam."

3. **Give students frequent assignments and make extensive, constructive comment on them.**

   "Students need to know what they are doing will in addition to what they need to improve," says one teacher of History. "I am always careful to praise their strengths and to be constructive and helpful as possible in pointing out their weaknesses."

   "I make a point of writing extensively on my students' papers," says a teacher of Architecture. "I make comments in the margin as I am reading and then append lists of strengths and suggestions for improvement."

4. **Give your students at least one assignment which consists of several options.**

   One teacher of English requires every student to write two essays on assigned topics. His third assignment, however, sets forth five or six options from which his students may choose the one which sounds interesting and most allows them to do their best.

   Examples of the options which he offers include: a piece of creative writing; a dramatic representation to be performed in front of the class (which can be a small group or team project); an original videotape to be shown to the class (which can also be a team effort); or a third essay (a "safe" option generally selected by his more conventional students). In addition, with his permission, students can create an option of their own if they wish.

   "More than five or six options tend to confuse some students; it becomes too difficult to decide, "he believes. "Two few options, on the other hand, restricts unduly my more creative and daring students." Although optional assignments must be related to the subject matter of the course, he encourages his students to take an interdisciplinary approach and to link content and skills from other courses.

5. **Ask students to analyse an essay or journal article and to write a critique of it.**

   One teacher of English assigns the work of literary critic and then asks his students to write an essay taking an adversary position. "If my assignments are provocative," he days, "I get better results. I stress the importance of their presenting a personal point of view. They should enjoy doing the paper; it should provide them with a personal learning experience."

   A Psychology teacher asks his students to write an evaluation or critique of a paper by a professional psychologist. "The process of analysis and evaluation captures what I am trying to do in the course," he explains.

APPENDIX L

6. Include peer-editing of student assignments (papers, computer programs, or design projects) in your course.

"In my upper division courses, I have my student submit two copies of each computer program they write," one staff member explains. "One copy goes to me and the readers and the other copy is assigned to another student in the class to evaluate and edit."

He believes that learning to program is like learning to write short stories; you learn not only by doing it but by reading programs other people have written. He has his students read and analyse exemplary programs, much as they might read excellent short stories. He believes that peer-editing also gives his students yet another opportunity to demonstrate understanding.

A teacher of Architecture uses the same strategy with student papers. He has students exchange papers to take home and edit. "The final paper is submitted along with a copy of the first draft with its edited corrections in read," he explains. "Each paper then receives two grades, one for the author and one for the editor."

In this way, students receive prompt informal feedback from a peer, followed by a grade and a formal critique by the staff member. This technique helps students acquire good editing as well as good writing skills.

7. Return a "perfect" exam to your students along with their own corrected exams.

A teacher of Business Administration likes to provide a great deal of feedback to his students after exams as a way of re-emphasising the themes of the course.

"I generally spend about half the class period walking my students through a 'perfect' midterm that I distribute to them along with their own corrected exams. I try to explain the ways in which most of their responses differ from what I consider to be a perfect answer or solution. I also hope that it helps them to do better on the second exam."

8. Have your students do research and write reports for specific "real world" clients.

Some teachers select or simulate a problem in their field and then have their students design a research project, gather the relevant data, and write up the results in a form appropriate for the "client".

Still other teachers find real clients for their students. For example, a teacher of Natural Resources has his students participate in all phases of the research, report writing, and oral presentation to client agencies for environmental impact studies in a particular area. Similarly, a Social Welfare teacher has her student's help agencies define their needs and write grant proposals for submission to foundations and federal agencies.

An Education teacher frequently has his students meet with top level university administrators to define current evaluation or information needs on the campus. Each of his students then designs and conducts a small-scale evaluation project on the campus and writes a report for the client-administrator in lieu of a standard term paper. He notes, "You get better results from your students if they feel there is a real audience for their ideas."

9. Give frequent quizzes.

One excellent Science teacher gives students practice quizzes (of 10 to 15 minutes duration) throughout the term. "I don't grade the quizzes," he explains, but I do read them and review materials with which a large number of students seem to have difficulty. I also seek out students who are having real problems understanding the material and spend more time with them in my office or in the departmental course centre."

10. Provide self-instructional materials or "modules" which relate to basic principles and skills needed to succeed in your course.

A staff member in Biochemistry had prepared a computer-assisted instructional unit for review by his students whose science and math backgrounds were weak. "I give a short diagnostic test the beginning of the course to help identify students who need this kind of review in order to keep up with my course," he explains.

A Physics teacher also gives his students a review module covering basic algebra during the first week of class. "Students who are unable to pass a quiz after reviewing this unit are not allowed to continue in my course," he says, "because there is no way they could succeed without understanding the fundamentals of algebra." Such students are advised to take an Algebra review course before enrolling in Introductory Physics.

11. Give your students frequent homework assignments and return them at the next class meeting.

"When I schedule students assignments, I block out my own time or grade them immediately following class," one Engineering teacher says. "This is important for two reasons. First, the quick turn around time ensures that my students are still thinking about the assignment. Thus any criticism or feedback is likely to have a stronger impact than if it were delayed a week or more. Second, prompt feedback indicates to my student the importance of what they are doing and my concern for their learning the material."

An English teacher agrees. "The impact is enormous when you return assignments at the next class session. Students are still anxious to know how they have done. That's a tremendous advantage in maximising the impact of feedback on their learning."

12. Hold review sessions before the midterm and the final exam.

Many excellent teachers hold reviews in all of their courses, but it is especially important in lower division courses where many students are still unsure about the performance levels expected of them.

"Many freshmen and transfer students have not really developed good study skills," says one humanities teacher. "Furthermore, because many of them realise or suspect this, their anxiety level is especially high when they enter the University. I try to help by giving them study questions for reviewing the content of my courses and by reviewing these questions in the last session of class."

These ideas and strategies have been adapted and used with permission from ***Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and*** Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley..

APPENDIX L

TEACHING IDEAS AND PRACTICAL STRATEGIES

The following ideas are suggested and used by outstanding university lecturers across a range of institutions and disciplines. Lecturers have found these strategies most beneficial when, after considering all the ideas, they select no more than three or four which appeared potentially most profitable and made a commitment to apply or adapt them to enhance their teaching.

SEEQ Factor 8: Assignments/Readings

Student work in higher education especially is largely oriented to the completion of assignments, including required readings. Thus, positive SETs of the texts and supplementary readings and of other assignments probably indicate that activity in learning was found to be valuable and that the learning experiences involved were meaningful. Assignments provide students with opportunities to practice new knowledge and skills. Furthermore, learning tasks that constitute assignments are often presented in learnable units even if they are not always completed in an appropriately paced sequence. The Assignments factor too seems consistent with sound principles of learning.

Teaching for Assignments/Readings

1. **Give a brief early assignment that allows your students to build on knowledge and skills acquired in previous courses.**

   One teacher of Architecture does this in his studio courses. "Beginning with a problem that my students can easily master increases their self-confidence and creates a relaxed, non-threatening atmosphere for the course," he explains. "My first assignment always calls for my students to use skills learned in prior courses and to apply them to an elementary design problem."

2. **Have your students keep a journal of their learning experiences during the course.**

   A journal can be a very effective way to facilitate students' reflection on their own learning during a course, leading to greater understanding and appreciation of the subject. It is important, however, to ensure that student have been familiarised with the process of journal writing and the benefits they can expect from it.

   This may be required as an assessable project, or simply recommended as effective preparation for class discussions, presentations, or examinations (particularly if the exam is designated as "open-journal" rather than "open-book," as one lecturer suggests).

3. **Use a structured process to help your students choose topics and groups.**

   One Public Health class, students' work in small groups on a major project throughout the term the teacher has developed procedures to help his students choose topics and groups. First, all possible project ideas are listed on the board using a brainstorming technique. The question posed to students is "What topics or areas would you like to explore?"

   Enough topics are generated so that each is taken on by a group of four to six students. The small groups meet around their selected topic of interest and students explore in detail the nature of their project. At the end of the first period, student indicate on an index card their name, address, phone number, group and whether their decision is firm. This list is typed and distributed at the next class meeting when needed changes are made.

   This procedure gives students a chance to identify appropriate topics and explore in preliminary fashion how they might proceed. It gets students working on their term projects early and has the added benefit of providing each student with a list of everyone in the class and their project interests.

4. **Use case studies to give your students practice at answering practice questions.**

   A teacher of Anthropology carefully prepares case study assignments to give her lower division students exposure to primary research techniques and strategies. Students are presented with a collection of photos, maps, and narrative information which depicts a site as an archaeologist would see it. Students must answer a series of questions, example, "What changes in eating habits can you infer from the artefacts found at two different levels?"

5. **Use the Socratic method to lead students through the steps involved in applying a particular concept.**

   For example, taking a concept like "licensing" as a public policy tool, a Political Science teacher guides his students through the steps involved in creating a regulatory commission, example, to license prostitution. "What would such a commission look like?" he asks. "Who would want to serve on it? What problems would it encounter? I force my students to apply abstract concepts and principles from their readings to new situations," he explains.

   Later in the term, he has his students actually stimulate the workings of a particular regulatory commission and engage in debates on the pros and cons of particular policy solutions.

6. **Make assignments which give your students field experience.**

   A Political Science teacher always includes at least one experiential assignment in his courses. A recent example was to require his students to interview a local politician as well as his or her spouse, children, staff members, and several constituents in order to get a better understanding of the daily life of a politician and the issues and problems he or she faces.

   "My students were then asked to tell their class about their experiences so that generalisations could be drawn. They compared their own conclusions with those presented by both the theoretical and the popular conceptions of politicians represented in their reading assignments."

   "My students are so experience-poor and theory-rich," he explains, "that I find as many ways as possible to get them to use the local area as a laboratory for enriching their understanding of course concepts and theories."

APPENDIX L

My students are also so competitive, that I try to give them a few non-competitive assignments where each student has his or her own unique experience which can be pooled with those of others in the class in a way that enriches everyone's understanding.

7. **Have your students keep a logbook of their work.**

The logbook should not be graded on its aesthetics or its organisation. It is intended to be a work in progress, not a final document. Following is an excerpt from a faculty member's course syllabus explaining the procedures to be followed in keeping their logbook.

This term you are being requested to maintain a "Logbook". "Your book should be organised along the following principles:

a. Include your notes and thought on all design problems, lectures, readings – or anything that bears on this course.

b. After each project is complete, include a photograph or sketch of it in the book.

c. After each review, comment on what was said about your project and indicate how you would modify your scheme if you were to continue on work on it.

d. At the end of the term, reread all the materials in the book, making new comments from your advanced perspective.

e. The logbook will be reviewed at mid-semester; and will be due on the last day of class."

8. **Prepare answers to exams and quizzes to hand out as soon as your students turn in their work.**

One Chemistry teacher prepares a handout of correct answers which he gives to students as they turn in their answer sheets and leave the room. "There is no point in making students wait several days or weeks to find out how they did, "this teacher explains. "They are most interested in the results at the time of the examinations, and it is at the time of the examinations that the greatest reinforcement of the learning can take place.

Note that this method gives students immediate feedback even though it may be a week or more before the assignments can be returned with comments or grades.

9. **Give your students frequent homework assignments and return them at the next class meeting.**

"When I schedule students assignments, I block out my own time or grade them immediately following class," one Engineering teacher says. "This is important for two reasons. First, the quick turn around time ensures that my students are still thinking about the assignment. Thus any criticism or feedback is likely to have a stronger impact than if it were delayed a week or more. Second, prompt feedback indicates to my student the importance of what they are doing and my concern for their learning the material."

An English teacher agrees. "The impact is enormous when you return assignments at the next class session. Students are still anxious to know how they have done. That's a tremendous advantage in maximising the impact of feedback on their learning."

10. **Have students solve problems at the board.**

A faculty member who teaches quantitative methods calls on students to come up to the board to solve problems. Each student is permitted to bring a fellow student as a "coach" so that he or she is not put on the spot. At the beginning of the term the problems are based on homework assignments. Toward the end of the term they are based on impromptu examples.

This method increases student discussion and interaction and encourages students to pay close attention in class.

11. **Assign "thought problems" which are typical of the problems faced by professionals in the field.**

A Forestry teacher assigns weekly "thought problems" which are of the same type he includes on his two midterms and final examination. "These `thought problems,'" he explains, "serve two functions: to expose students to my kind of exam and to get them to really think through the material covered each week. They are not graded, but every Monday I go over them in class so that students can see how well they are doing."

He goes on to say, "I can best define `thought problems' in terms of the type of questions professional foresters are asked, such as, `What is killing that tree?' not `Name six factors which can kill trees.'"

12. **Give assignments which require students to visualize problems and make approximations.**

"I try to get students to see things visually," one Engineering teacher says. "I try to get them to understand that there is more than one way to solve a problem."

He notes that with the advent of computers, students are inclined to take even very poor data to five decimal places. "They have little feel for approximation, little experience using a rule of thumb. Without taxing them mathematically, I give them assignments which require them to think visually and to make approximations without resort to a computer."

13. **Be very specific in the questions you ask.**

"As beginning essay writers, undergraduates need focused test questions," one history teacher explains. Problem oriented exams can elicit more meaningful responses than broad, vaguely worded questions. For example, it is difficult for a student to respond to a question like, "Discuss the implications of the Monroe Doctrine." Students have no sense of boundaries or when they have completed the topic. On the other hand, a question such as, "Illustrate how the Monroe Doctrine might be involved in a Russian-American incident," is likely to generate good responses from students.

These ideas and strategies have been adapted and used with permission from Marsh, H. W. & Roche, L. A. (1994), The Use of Students' Evaluations of University Teaching to Improve Teaching Effectiveness. Canberra: Department of Employment, Education & Training; and Davis, B. G., Wood, L., & Wilson, R. (1993). Berkeley Compendium of Suggestions for Teaching with Excellence. University of California, Berkeley.

APPENDIX M

INTERVIEW GUIDE

Name: _____

1. What factors affected your participation in the group meetings?

2. How have you benefited from talking about your student ratings in groups?

3. How comfortable are in talking about your student ratings with peers?

4. Were you able to make any changes to your teaching as a result of receiving the ratings and having discussions with your peers?

5. Did the fact that you were examining the ratings with colleagues led to any action that you normally would not have taken?

6. What would you recommend to make the intervention successful for faculty professional development?

7. Which aspect of the intervention was most helpful to you?

8. In what ways did the intervention motivate you to improve your teaching?

9. Was the intervention worth your time and effort?

10. What would you say are the strengths and drawbacks of the intervention?

APPENDIX N: Letter for Control Group

November 2002

Dear Colleague

Please find enclosed your ratings report from the mid-semester ratings collected in October. The following items are also enclosed:

Interpretation Guide
Copy of the questionnaire
TIPS sheets for the factors on the questionnaire

As indicated earlier a lunchtime discussion session will be held on Thursday, November 28 and Friday, November 29, at 12:15 p.m. in the meeting room of the Academic Affairs unit. You are required to attend only one session.

Please do make every effort to attend this session to allow us to explore together how student ratings feedback might be effectively used for personal and professional development. Should you have any questions please contact me at telephone number 941-0286 or via e-mail at arpenny@hotmail.com.

Sincerely

Angela Penny

APPENDIX O: Letter of Appreciation

December 2002

Dear _____

Thank you for your voluntary participation in the research project that examined how student ratings feedback can facilitate personal and professional development. This study is part of my doctoral studies being done at the University of Durham, England.

You giving me the opportunity to administer rating forms on your teaching effectiveness twice in the semester has made it possible for data to be collected to draw conclusions on how student ratings feedback can provide a base for improving the quality of teaching and learning in higher education.

I do hope for you and your other colleagues at _____ who participated in the study that this will be the beginning of further action research activities that will reap significant benefits to your own professional development, to the wider university community, and to the present and future students of the University.

Best regards

Angela Penny

Appendix P

Correlations of SEEQ Factors by Students N = 2370

	LEARN	ENTHUS	ORGAN	GROUP	RAPPORT	COVER	GRADING	READING	OVERALL
LEARNING	–								
ENTHUSIASM	.659	–							
ORGANISATION	.664	.681	–						
GROUP	.538	.596	.551	–					
RAPPORT	.504	.645	.580	.606	–				
COVERAGE	.524	.580	.644	.570	.560	–			
GRADING	.498	.539	.566	.483	.551	.500	–		
READINGS	.476	.432	.474	.446	.361	.480	.415	–	
OVERALL	.670	.728	.700	.565	.625	.563	.539	.436	–

All correlations are significant at the 0.01 level (2-tailed).

Appendix P

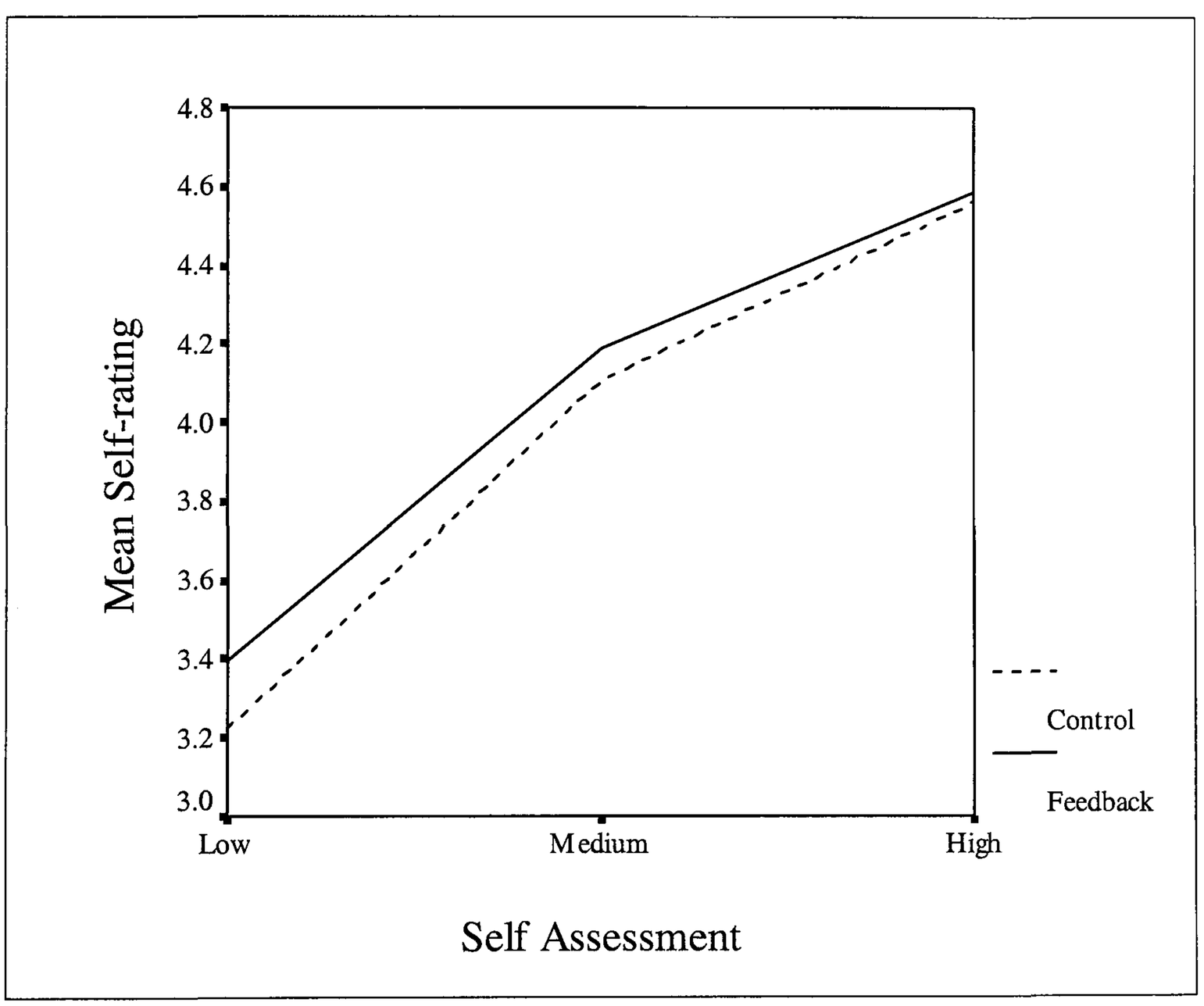
Correlations of SEEQ Factors by Classes N = 150

	LEARN	ENTHUS	ORGAN	GROUP	RAPPORT	COVER	GRADING	READING	OVERALL
LEARNING	–								
ENTHUSIASM	.830	–							
ORGANISATION	.871	.828	–						
GROUP	.759	.747	.738	–					
RAPPORT	.718	.827	.765	.787	–				
COVERAGE	.785	.767	.819	.804	.728	–			
GRADING	.692	.745	.726	.631	.714	.669	–		
READING	.696	.621	.655	.712	.580	.726	.634	–	
OVERALL	.891	.911	.884	.773	.812	.770	.772	.675	–

All correlation is significant at the 0.01 level (2-tailed).

APPENDIX Q

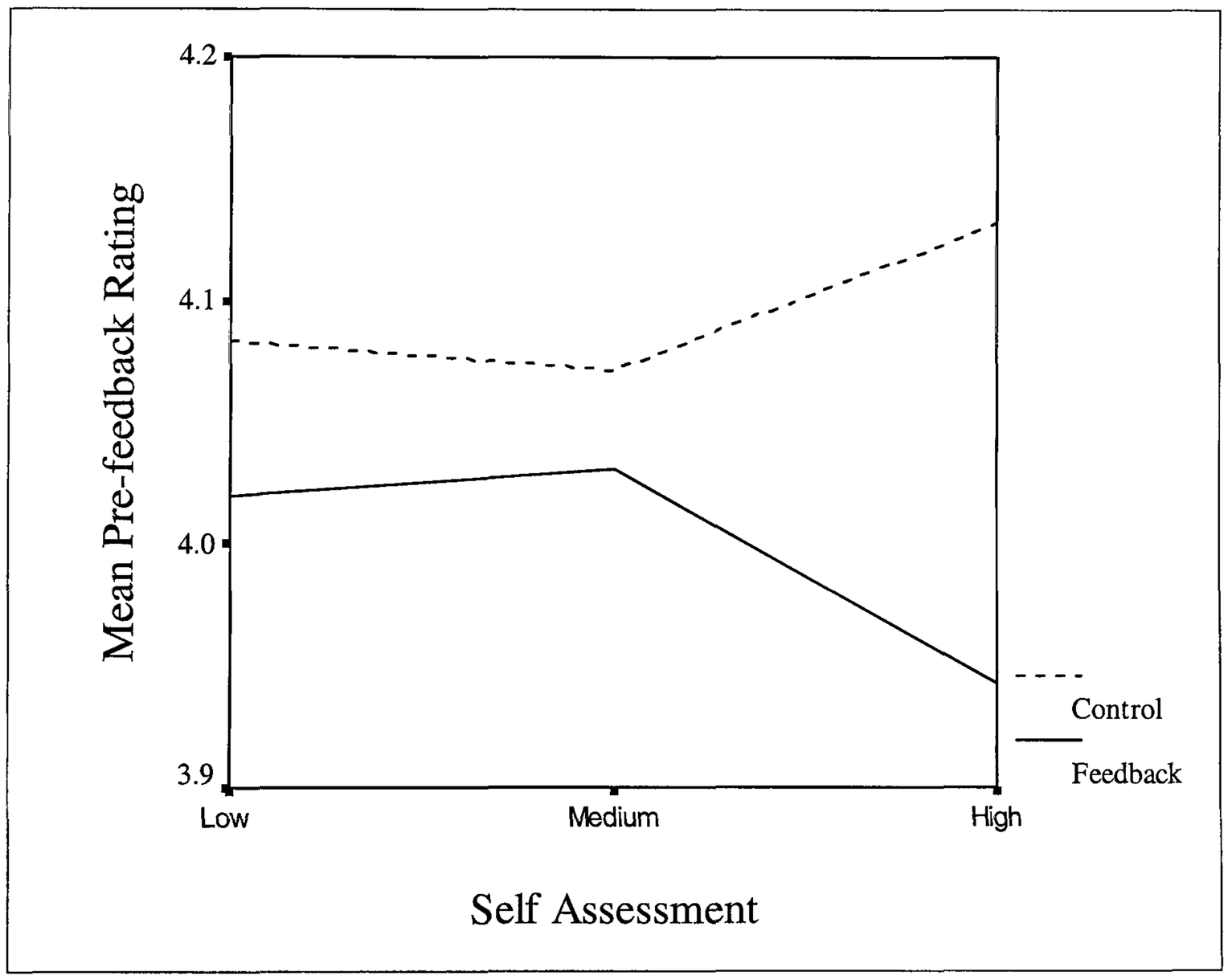
Teacher Self-Rating by Intervention Group



Note: Low = 20th percentile
Medium = 20th to 80th percentile
High = above 80th percentile

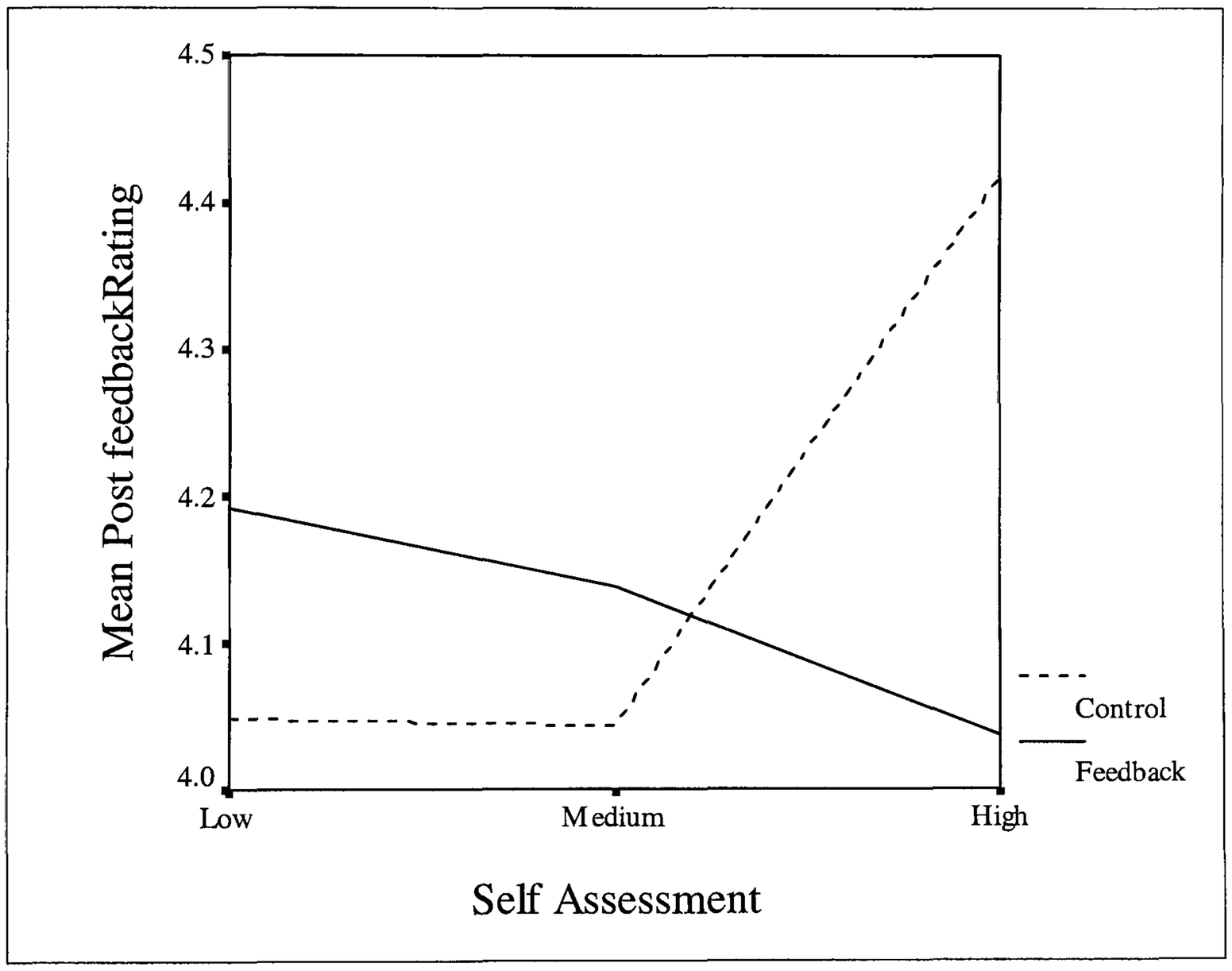
Appendix Q

Teacher Self-Rating on Pre-feedback Ratings



APPENDIX Q

Teacher Self-Rating on Post feedback Ratings



References

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30*, 221-270.

Abrami, P. C. (2001). Improving judgements about teaching effectiveness using student ratings forms. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them*? New Directions for Institutional Research, (No. 109, pp. 59-87). San Francisco: Jossey-Bass.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not know. *Journal of Educational Psychology, 82* (2), 219-231.

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. *Review of Educational Research, 52*, 446-464.

Adams, J. V. (1997). Student evaluations: The ratings game. *Inquiry 1* (2), 10-16.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149* (1), 1-43.

Aleamoni, L. M. (1978). The usefulness of student evaluations in improving college teaching. *Instructional Science, 7*, 95-105.

Aleamoni, L. M. (1996). Why we do need norms of student ratings to evaluate faculty: Reaction to McKeachie. *Instructional Evaluation and Faculty Development, 16* (1), 18-19.

Annis, L. F. (1989). Partners in teaching improvement, *Journal of Staff, Programme and Organizational Development*, *7*, 7-13.

Argyris, C., & Schön, D. A. (1974). *Theory in practice: Increasing professional effectiveness*. San Francisco: Jossey-Bass.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker Publishing.

Atchison, E. C. (1987). *Effects of student feedback on instructional improvement when augmented with consultation and a developmental teaching model*. (Doctoral dissertation, George Washington University, USA).

Atherton, J. S. (2002). *Learning and Teaching: Learning from Experience*. Retrieved January 10, 2004 from http://www.dmu.ac.uk/~jamesa/learning/experience.htm.

Bandura, A. (1982). Self-efficacy mechanisms in personal agency. *American Psychologist*, *37*, 122-148.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, *28* (2), 117-148.

Barnett, R. (2000). *Realizing the university in an age of supercomplexity*. Buckingham: SRHE /Open University Press.

Barr, R. B., & Tagg, J. (1995). From teaching to learning: A new paradigm for undergraduate education. *Change*, Nov/Dec, 13-25. Retrieved January 20, 2004 from http://critical.tamucc.edu/~blalock/readings/tch2learn.htm

Baxter, E. P. (1991). The TEVAL experience, 1983-88: The impact of a student evaluation of teaching scheme on university teachers. *Studies in Higher Education*, *16*, 151-179.

Beaty, L. (1999). Consultation through action learning. In C. Knapper, & S. Piccinin (Eds.), *Using consultants to improve teaching*. New Directions for Teaching and Learning, (No. 79, pp. 51-58). San Francisco: Jossey-Bass.

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher Education*, *19* (2), 151-61.

Becher, T., & Trowler, P. R. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines* (2nd ed.). Buckingham: SRHE/Open University.

Bergquist, W. H., & Phillips, S. R. (1975). Components of an effective faculty development program. *Journal of Higher Education*, *46* (2), 177-211.

Bernstein, D. J., Jonson, J., & Smith, K. (2000). An examination of the implementation of peer review of teaching. In K. E. Ryan (Ed.), *Evaluating teaching in higher education: A vision for the future*. New Directions for Teaching and Learning, (No. 83, pp. 73-86), San Francisco: Jossey-Bass.

Bigger, J. T. (2003, September). Issues in subgroup analyses and meta-analyses of clinical trials. *Journal of Cardiovascular Electrophysiology*, *14* (9), Supplement S6-S8.

Biggs, J. (2001). The reflective institution: Assuring and enhancing the quality of teaching and learning. *Higher Education*, *41*, 221-238.

Biglan, A. (1973a). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, *57*, 195-203.

Biglan, A. (1973b). Relationship between subject matter characteristics and the structure and output of university departments. *Journal of Applied Psychology*, *57*, 204-213.

Border, L. L. (1997). The creative art of effective consultation. In K. T. Brinko & R. J. Menges (Eds.), *Practically speaking: A sourcebook for instructional consultants in higher education* (pp. 17-21). Stillwater, OK: New Forums Press.

Boruch, R., de Moya, D., & Snyder, B. (2002). Importance of randomized field trials in education and related areas. In F. Mosteller, & R. Boruch (Eds.), *Evidence matters: Randomised trials in education research.* Washington D.C.: Brookings Institution.

Boud, D. (1999). Situating academic developments in professional work: Using peer learning. *International Journal for Academic Development*, *4* (1), 3-10.

Boud, D., & McDonald, R. (1981). *Educational development through consultancy.* Surrey, England: Society for Research into Higher Education.

Boud, D., Cohen, R., & Walker, D. (Eds.). (1993). *Using experience for learning.* Buckingham: SRHE/Open University.

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning*. New York: Kogan Page.

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate.* Princeton, NJ: The Carnegie Foundation for the Advancement of Teaching.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work.* San Francisco: Jossey-Bass.

Bray, J. H., & Howard, G. S. (1980). Methodological considerations in the evaluation of a teacher training programme. *Journal of Educational Psychology*, *72*, (1), 62-70.

Brinko, K. T. (1990). Instructional consultation with feedback in higher education. *Journal of Higher Education*, *61* (1), 65-83.

Brinko, K. T. (1993). The practice of giving feedback to improve teaching: What is effective? *Journal of Higher Education, 64* (5), 574-593.

Brinko, K. T. (1997). The interactions of teaching improvement. In K. T. Brinko, & R. J. Menges (Eds.), *Practically speaking: A sourcebook for instructional consultants in higher education.* Stillwater OK: New Forum Press. (Reprinted from *New Directions for Teaching and Learning*, No. 48, 1991).

Brinko, K. T., & Menges, R. J. (Eds.). (1997). *Practically speaking: A sourcebook for instructional consultants in higher education.* Stillwater OK: New Forum Press.

Brookes, S. T., Whitley, E., Peters, T. J. Mulheran, P. A., Egger, M., & Smith, G. D. (2001). Subgroup analysis in randomised controlled trials: Quantifying the risks of false-positives and false negatives. *Health Technology Assessment,* 5 (33).

Brookfield, S. (1995). *Becoming a critically reflective teacher.* San Francisco: Jossey-Bass.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* London: Houghton Mifflin.

Caribbean Community Secretariat (1997). *Report of the Eighteenth Meeting of the Conference of Heads of Government of the Caribbean Community*, Montego Bay, Jamaica.

Carroll, J. G. (1981). Faculty self-evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation.* California: Sage Publications.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48* (3), 378-399.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited.* IDEA Paper No. 32. Manhattan, KS: Centre for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University. Retrieved May, 31, 2001 from http://idea.ksu.edu.

Cashin, W. E. (1996). Should student ratings be interpreted absolutely or relatively? Reaction to McKeachie (1996). *Instructional Evaluation and Faculty Development*, *16* (2), 14-19.

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin & Associates, *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 25-44). Bolton, MA: Anker Publishing.

Cashin, W. E., Downey, R. G., & Sixbury, G. R. (1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: A reply to Marsh (1994). *Journal of Educational Psychology*, *86*, 649-657.

Centra, J. A. (1973). Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology*, *65*, 395-401.

Centra, J. A. (1978). Types of faculty development programs. *Journal of Higher Education*, *49* (2), 151-162.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco: Jossey-Bass.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *71* (1), 17-33.

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education*, *21* (1), 68-75.

Clement, M., & Vandenberghe, R. (2000). Teachers' professional development: A solitary or collegial (ad)venture? *Teaching and Teacher Education*, *16*, 81-101.

Coe, R. (2001). *What is an 'effect size'? A guide for users*. (Unpublished manuscript, University of Durham, Durham, UK).

Coe, R. (1998). Can feedback improve teaching? A review of the social science literature with a view to identifying the conditions under which giving feedback to teachers will result in the improved performance. *Research Papers in Education*, *13* (1), 43-66.

Coffey, M., & Gibbs, G. (2001). The evaluation of the student evaluation of educational quality questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education*, *26* (1), 89-93.

Cohen, J. (1977). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cohen, L., & Manion, L. (1994). *Research methods in education*. London: Routledge.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college teaching: A meta-analysis of findings. *Research in Higher Education*, *13*, 321-341.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research*, *51* (3), 281-309.

Cohen, P. A. (1983). Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education*, *54* (4), 448-456.

Cohen, P. A. (1987). *A critical analysis and re-analysis of the multisection validity meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (ERIC Reproduction Service Document, No. ED283876).

Cohen, P. A. (1991). Effectiveness of student ratings feedback and consultation for improving instruction in dental school. *Journal of Dental Education*, *55* (2), 145-150.

Cohen, P. A., & Herr, G. (1982). Using an interactive feedback procedure to improve college teaching, *Teaching of Psychology*, *9* (3), 138-140.

Convery, A. (2001). A teacher's response to 'reflection-in-action'. In J. Soler, A. Craft, & H. Burgess (Eds.), *Teacher development: Exploring our own practice*. London: Paul Chapman/Open University.

Cook, T. D. (2002). Randomised experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24* (3), 175-199.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional methods for summarising research findings. *Psychological Bulletin*, *87*, 442-449.

Cooper, H., & Dorr, N. (1995). Race comparisons on need for achievement: A meta-analytic alternative to Graham's narrative review. *Review of Educational Research*, *65* (4), 483-508.

Cooper, H., & Hedges, L. V. (Ed.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, H., Valentine, J. C., & Charlton, K. (2000). The methodology of meta-analysis. In R. Gersten, E. P. Schiller, & S. Vaughn (Eds.), *Contemporary special education research: Synthesis of the knowledge based on critical instructional issues.* London: Lawrence Erlbaum Associates.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41* (5), 511-535.

Cowan, J. (2001). Developing skills, abilities or capabilities: Implications for educational developers. *Educational Developments, 2* (3), 1-4.

Cranton, P., & Carusetta, E. (2002). Reflecting on teaching: The influence of context. *International Journal for Academic Development, 7* (2), 167-176.

Curriculum Evaluation and Management Centre (2000). A culture of evidence. Retrieved May 25, 2002 from http://cem.dur.ac.uk/ebeuk/culture.htm.

Currie, J., Harris, P., & Thiele, B. (2000). Sacrifices in greedy universities: Are they gendered? *Gender and Education, 12* (3), 269-291.

d'Apollonia, S., & Abrami, P. C. (1997a). Navigating student ratings of instruction. *American Psychologist, 52* (11), 1198-1208.

d'Apollonia, S., & Abrami, P. C. (1997b). "In Response". *Change, 29* (5), 18-19.

Daly, P., & Defty, N. (2003). *Extension of single-sex public school provision: Evidential concerns.* Paper presented at the Evidence-Based Policies and Indicator Systems Conference. London.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47* (2), 108-121.

Davis, B. G., Wood, L., & Wilson, R. C. (1983). *Berkeley Compendium of Suggestions for Teaching with Excellence*. University of California, Berkeley. Retrieved December 10, 2001 from http://teaching.berkeley.edu/compendium/.

Day, C. (1997). In-service teacher education in Europe: Conditions and themes for development in the 2lst century. *British Journal of In-Service Education*, *23* (1), 39-54.

Day, C. (1999). *Developing teachers: The challenges of lifelong learning*. London: Falmer Press.

Day, C. (2000). Teachers in the twenty-first century: Time to renew the vision. *Teachers and Teaching: Theory and Practice*, *6* (1), 101-115.

Deming, W. E. (1986). *Out of crisis*. Cambridge, MA: Center for Advanced Engineering Study.

Denzin, N. K. (1978). *The research act* (2nd ed.). New York: McGraw Hill.

Denzin, N. K., & Lincoln Y. S. (Eds). (2000). *Handbook of qualitative research.* London: Sage Publications.

Department for Education and Skills (2003). *The future of higher education*. Retrieved February 17, 2003 from http://www.dfes.gov.uk./highereducation/hestrategy.

DeZure, D. (1999). Evaluating teaching through peer classroom observation. In P. Seldin & Associates. *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 70-96). Bolton, MA: Anker Publishing.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Chicago: D.C. Heath.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student rating of teaching. *Journal of Higher Education*, *53*, 51-62.

Dowling, W. C. (2002, February). All lose when education is sold. *Daily Targum*. Retrieved February 14, 2002 from http://www.bus.isu.edu.

DuCette, J., & Kennedy, J. (1982). Do grading standards affect student evaluations of teaching? Some evidence on an old question. *Journal of Educational Psychology*, *74*, 308-314.

Dunkin, M. J. (1996). Types of errors in synthesizing research in education. *Review of Educational Research*, *66* (2), 87-97.

Easthope, C., & Easthope, G. (2000). Intensification, extension and complexity of teachers' workload. *British Journal of Sociology of Education*, *21* (1), 43-58.

Edwards, T. (2000). 'All the evidence shows...': Reasonable expectations of educational research. *Oxford Review of Education*, *26* (3 & 4), 299-311.

Egger, M., & Smith, G. D. (1997). Meta-analysis: Potentials and promise. *British Medical Journal*, *315*, 1371-1374.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43* (4), 483-501.

Elliott, J. (1991). *Action research for educational change*. Buckingham: Open University Press.

Elton, L. (1998). Dimensions of excellence in university. *International Journal for Academic Development*, *3*, 3-11.

Erickson, G. R., & Erickson, B. L. (1979). Improving college teaching: An evaluation of a teaching consultation procedure. *Journal of Higher Education*, 50 (5), 670-683.

Erickson, G. R., & Sheehan, D. S. (1976). *An evaluation of a teaching improvement process for university faculty*. Paper presented at annual meeting of the American Educational Research Association, San Francisco.

Evans, L. (2001). Delving deeper into morale, job satisfaction and motivation among education professionals: Re-examining the leadership dimension. *Educational Leadership and Administration*, *29* (3), 291-306.

Eysenck, H. J. (1994). Systematic reviews: Meta-analysis and its problems. *British Medical Journal*, *309*, 789-792.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education*, *9*, 199-242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, *10*, 149-172.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education 18* (1), 3-124.

Feldman, K. A. (1984). Class size and college students' evaluations of teachers and course: A closer look. *Research in Higher Education*, *21* (4), 45 –116.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities. *Research in Higher Education*, *28*, 291-344.

Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers. *Research in Higher Education*, *30*, 137-194.

Feldman, K. A. (1989b). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30* (6), 583-645.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In. R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368-395). New York: Agathon Press.

Ferman, T. (2002). Academic professional development practice: What lecturers find valuable. *International Journal for Academic Development*, *7* (2), 146-158.

Ferrer, R. L. (1998). Graphical methods for detecting bias in meta-analysis. *Family Medicine*, *30* (8), 579-583.

Fitz-Gibbon, C. T. (1984). Meta-analysis: An explication. *British Educational Research Journal*, *10* (2), 135-144.

Fitz-Gibbon, C. T. (1985). The implications of meta-analysis for education research. *British Education Research Journal*, *11* (1), 45-49.

Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. In S. W. Raudenbush, & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies and schooling from a multilevel perspective*. London: Academic Press.

Fitz-Gibbon, C. T. (1996). *Monitoring education: Indicators, quality, and effectiveness*. London: Cassell.

Fitz-Gibbon, C. T. (1997). *The value added national project: Final report feasibility studies for a national system of value added indicators*. London: Curriculum and Assessment Authority.

Fitz-Gibbon, C. T., & Morris, L. L. (1987). *How to analyse data*. London: Sage Publications.

Fontana A., & Frey, J. H. (2000). The interview: from structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 645-672). London: Sage Publications.

Franklin, J., & Theall, M. (1989). *Who reads ratings: Knowledge, attitude and practice of users of student ratings of instruction*. Paper presented at the 70th annual meeting of the American Educational Research Association, San Francisco. (ERIC Reproduction Production Service ED306241).

Freebody, P. (2003). *Qualitative research in education: Interaction and practice*. London: Sage Publications.

Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). London: RoutledgeFalmer.

Fullan, M., & Hargreaves, A. (1992). *What's worth fighting for in your school? Working together for improvement*. Buckingham: Open University Press.

Fuller, F. F., & Manning, B.A. (1973). Self-confrontation reviewed: A conceptualisation for video playback in teacher education. *Review of Educational Research*, *43* (4), 469-528.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research: An introduction. New York: Longman.

Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and application*. Upper Saddle River, New Jersey: Prentice-Hall.

Gillmore, G. M. (1996). Chair's corner: Like devils, the instructor you know is better than the one you do not know. *Instructional Evaluation and Faculty Development*, *16* (1), 10.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research.* London: Sage Publications.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper, & L. V. Hedges (Eds.). *The handbook of research synthesis* (pp. 339-355). New York: Russell Sage Foundation.

Golby, M., & Appleby, R. (1995). Reflective practice through critical friendship: some possibilities. *Cambridge Journal of Education*, *25*, 149-160.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN*. London, University of London, Institute of Education.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A multilevel analysis of examination results. *Oxford Review of Education*, *19* (4), 425-433.

Gorard, S. (2002a). Political control: A way forward for educational research. *British Journal of Educational Studies*, *50* (3), 378-389.

Gorard, S. (2002b). Let's keep it simple: The multilevel modelling debate. *Research Intelligence*, *81*, 24-25.

Gordon, G. (2002). The roles of leadership and ownership in building an effective quality culture. *Quality in Higher Education*, *8* (1), 97-106.

Graham, S. (1995). Narrative versus Meta-analytic reviews of race differences in motivation: A comment on Cooper and Dorr. *Review of Educational Research*, *65* (4), 509-514.

Greene, J. C. (2000). Understanding social programs through evaluation. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 981-999). Thousand Oaks, CA: Sage Publications.

Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383-398). New York: Russell Sage Foundation.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52* (11), 1182-1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist*, *52* (11), 1209-1217.

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology*, *26*, 534-552.

Hammersley, M. (1997). Educational research and a response to David Hargreaves. *British Educational Research Journal*, *23* (2), 141-161.

Hampton, S. (2001). Student ratings feedback and consultation on instructional practices: Effects on teaching, learning, and motivation. *Dissertation Abstracts International* (UMI No. 3001662).

Handal, G. (1999). Consultation using critical friends. In C. Knapper & S. Piccinin (Eds.), *Using consultants to improve teaching*. New Directions for Teaching and Learning, (No. 79, pp. 59-70). San Francisco: Jossey-Bass.

Hargreaves, A. (1997). Cultures of teaching and educational change. In B. J. Biddle, T. L. Good, & I. F. Goodison (Eds.) *International handbook of teachers and teaching*, (Vol. 2, pp. 1297-1319). London: Kluwer Academic.

Hargreaves, A. (2001). The emotional geographies of teachers' relations with colleagues. *International Journal of Educational Research, 35*, 503-527.

Hargreaves, D. H. (1999). The knowledge-creating school. *British Journal of Educational Studies, 47* (2), 122-144.

Hargreaves, D. H. (2000). Teaching as a research-based profession: Possibilities and prospects. In B. Moon, J. Butcher, & E. Bird (Eds.), *Leading Professional Development in Education.* London: RoutledgeFalmer. (Reprinted from Teaching Training Agency Lecture, 1996, London: Teacher Training Agency).

Harrison, R. (1996). Action Learning: Route or barrier to the learning organisation? *Journal of Workplace Learning, 8* (6), 27-38.

Harvey, L. (2002). The end of quality. *Quality in Higher Education, 8* (1), 5-22.

Harvey, L. (2001). *Student Feedback: A report to the higher education funding council for England.* Birmingham: Centre for Research into Quality, University of Central England, Birmingham.

Hativa, N. (1996). University instructors' ratings profiles: Stability over time and disciplinary differences. *Research in Higher Education, 37* (3), 341-365.

Hatti, J., & Marsh, H. W. (1996). The relationship between research and teaching: A meta-analysis. *Review of Educational Research, 66*, 507-542.

Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*, 33-49.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.

Hicks, O. (1999a). Integration of central and departmental development - reflections from Australian universities. *International Journal of Academic Development*, *4* (1), 43-51.

Hicks, O. (1999b). A conceptual framework for instructional consultation. In C. Knapper, & S. Piccinin (Eds.), *Using consultation to improve teaching*. New Directions for Teaching and Learning, (No. 79, pp. 9-18). San Francisco: Jossey-Bass.

Ho, A., Watkins, D., & Kelly, M. (2001). The conceptual change approach to improving teaching and learning: An evaluation of a Hong Kong staff development programme. *Higher Education*, *42*, 143-169.

Hopkins, K. D., Hopkins, B. R., & Glass, G. V. (1996). *Basic statistics for the behavioural sciences*. Boston: Allyn & Bacon.

Hox, J. J. (1995). Applied Multilevel Analysis. Amsterdam: TT-Publikaties. Retrieved March 22, 2003 from http://www.fss.uu.nl/ms/jh/publist/amaboek.pdf.

Hoyt, D. P., & Howard, G. S. (1978). The evaluation of faculty development programs. *Research in Higher Education*, *8*, 25-38.

Humes, W., & Bryce, T. (2001). Scholarship, research and the evidential basis on policy development in education. *British Journal of Educational Studies*, *49* (3), 329-352.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications.

Husbands, C. T., & Fosh, P. (1993). Students' evaluations of teaching in higher education: Experiences from four European countries and some implications of the practices (1). *Assessment &Evaluation in Higher Education*, *18* (2), 95-114.

Hutchings, P. (2004, January). Building a better conversation about learning. *Carnegie Perspectives*. Retrieved February 19, 2004 from http://www.carnegiefoundation.org/perspectives/perspectives2004.Jan.htm

Hutchings, P., & Shulman, L. S. (2000). The scholarship of teaching: New elaborations, new developments. In D. DeZure (Ed.) *Learning from change: Landmarks in teaching and learning in higher education from change magazine, 1969-1999*, London: Kogan Page. (Reprinted from *Change*, *31* (5), 10-15).

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression.* London: Sage Publications.

Jenkins, A. (1996). Discipline-based educational development. *International Journal for Academic Development*, *1* (1), 50-62.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press.

Johnson, T. D., & Ryan, K. E. (2000). A comprehensive approach to the evaluation of college teaching. In K. E. Ryan (Ed.), *Evaluating teaching in higher education: A vision for the future*. New Directions for Teaching and Learning, (No. 83, pp. 109-123). San Francisco: Jossey-Bass.

Kember, D. (1997). A reconceptualisation of the research into university academics' conceptions of teaching. *Learning and Instruction*, *7* (3), 255-275.

Kember, D. (2000). *Action learning and action research: Improving the quality of teaching and learning*. London: Kogan Page.

Kember, D. (2003). To control or not to control: The question of whether experimental designs are appropriate for evaluating teaching innovations in higher education. *Assessment & Evaluation in Higher Education*, *28* (1), 89-101.

Kember, D., & Gow, L. (1994). Orientations to teaching and their effect on the quality of student learning. *Journal of Higher Education 65* (1), 58-74.

Kember, D., & McKay, J. (1996). Action research into the quality of student learning: A paradigm for faculty development. *Journal of Higher Education 67* (5), 528-554.

Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perception of good and poor teaching. *Higher Education*, *40*, 69-97.

Kember, D., Leung, D. Y. P, & Kwan, K. P. (2002). Does the use of student feedback questionnaire improve the overall quality of teaching? *Assessment & Evaluation in Higher Education*, *27* (5), 411-425.

Kemmis, S., & McTaggart, R. (2000). Participatory action research. In N. K. Denzin, & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 567-606). Thousand Oaks, CA: Sage Publications.

Kerridge, J. R., & Matthews, B. P. (1998). Student ratings of courses in HE: Further challenges and opportunities. *Assessment & Evaluation in Higher Education*, *23* (1), 71-82.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Knapper, C., & Piccinin, S. (1999). Consulting about teaching: An overview. In C. Knapper & S. Piccinin (Eds.), *Using consultants to improve teaching*. New Directions for Teaching and Learning, (No. 79, pp. 3-8). San Francisco: Jossey-Bass.

Knight, P. (2002a). A systematic approach to professional development: Learning as practice. *Teaching and Teacher Education*, *18*, 229-241.

Knight, P. (2002b). *Being a teacher in higher education*. Buckingham: SRHE/Open University.

Knight, P., & Trowler, P. (2000). Department-level cultures and the improvement of learning and teaching. *Studies in Higher Education*, *25* (1), 69-83.

Knowles, M. S. (1980). *The modern practice of adult education: From pedagogy to andragogy*. New York: Association Press.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliff, NJ: Prentice Hall.

Koon, J., & Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education*, *66* (1), 61-81.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modelling*. London: Sage.

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of research in education*. Itasca, IL: F. E. Peacock Publishers.

Kulik, J. A. (2001). Student ratings: validity, utility, and controversy. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* New Directions for Institutional Research, (No. 109, pp. 9-26). San Francisco: Jossey-Bass.

Kwan, K. P. (1999). How fair are student ratings in assessing the teaching performance of university teachers? *Assessment & Evaluation in Higher Education*, *24* (2), 181-194.

Kwan, K. P. (2000). How University Students Rate their Teachers: A study of the rating attitudes and behaviours of university students in teaching evaluations. (Doctoral dissertation, University of Durham, Durham, UK).

L'Hommedieu, R. L., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology*, *82* (2), 232-241.

Lafferty, G., & Fleming, J. (2000). The restructuring of academic work in Australia: Power, management and gender. *British Journal of Sociology of Education*, *21* (2), 257-267.

Larrivee, B. (2000). Transforming teaching practice becoming a critical reflective teacher, *Reflective Practice*, *1* (3) 293-307.

Lave, J. (1993). The practice of learning. In S. Chaiklin, & J. Lave. *Understanding practice: Perspectives on activity and practice*. Cambridge: Cambridge University Press.

Lave, J., & Wenger E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Lenze, L. F. (1996). Instructional development: What works? *Update*, *2* (4), 1-4. National Education Association.

Levinson-Rose, J., & Menges, R. J. (1981). Improving college teaching: A critical review of research. *Review of Educational Research*, *51*, 403-34.

Lewin, K. (1951). *Field theory in social science; selected theoretical papers*. In D. Cartwright (Ed.). New York: Harper & Row.

Light, G., & Cox, R. (2001). *Learning and teaching in higher education: The reflective professional*. London: Paul Chapman.

Light, R. J., Singer, J. D., & Willett, J. B. (1994). The visual presentation and interpretation of meta-analyses. In H. Cooper, and L. V. Hedges (Eds.) *The Handbook of Research Synthesis* (pp. 439-453). New York: Russell Sage Foundation.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. London: Sage Publications.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioural treatment: Confirmation from meta-Analysis. *American Psychologist, 48* (12), 1181-1209.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. London: Sage Publications.

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record, 91* (4), 509-536.

Little, J. W. (2003). Inside teacher community: Representations of classroom practice. *Teachers College Record, 105* (6), 913-945.

Litzleman, D. K., Stratos, G. A., Marriott, D. J., Lazaridis, E. N., & Skeff, K. M. (1998). Beneficial and harmful effects of augmented feedback on physicians' clinical teaching performances. *Academic Medicine, 73* (3), 324-332.

Locke, E. A., & Latham, G. P. (1984). *Goal Setting: A motivational technique that works*! Englewood Cliffs, NJ: Prentice Hall.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance.* Englewood Cliffs, NJ: Prentice Hall.

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin & Associates, *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 45-69). Bolton, MA: Anker Publishing.

Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77-95.

Marsh, H. W. (1987). Students' evaluation of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*, 253-388.

Marsh, H. W. (1991). Multidimensional students' evaluation on teaching effectiveness: A test of alternative higher order structures. *Journal of Educational Psychology*, *83*, 285-296.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education*, *64* (1), 1-18.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.) *Higher Education: Handbook of Theory and Research* (Vol. 8, pp. 143-232). New York: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991a). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level and course level. *Teaching and Teacher Education*, *7* (1), 9-18.

Marsh, H. W., & Hocevar, D. (1991b). Students' evaluations of teaching effectiveness: The stability of mean rating of the same teachers over a 13-year period. *Teaching and Teacher Education*, *7* (4), 303-314.

Marsh, H. W., & Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and effective criteria. *Journal of Educational Psychology*, *72*, 468-475.

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, *30* (1), 217-251.

Marsh, H. W., & Roche, L. A. (1994). *The use of students' evaluations of university teaching to improve teaching effectiveness*. Camberra: Department of Employment, Education and Training. Retrieved November 24, 2001 from http://cea.curtin.edu.au/seeq/target.html.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52* (11), 1187-1197.

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology*, *92* (1), 202-228.

Marsh, H. W., Rowe, K. J., & Martin, A. (2002). PhD's Students' Evaluation of Research Supervision: Issues, complexities, and challenges in a nationwide Australian experiment in Benchmarking Universities. *Journal of Higher Education*, *73* (3), 313-348.

ton, F., & Säljö, R. (1976). On qualitative differences in learning: Outcomes and process. *British Journal of Educational Psychology*, *46*, 4-11 and 115-127.

ton, F., & Säljö, R. (1997). Approaches to learning. In F. Marton, D. Hounsell, & N. Entwistle (Eds.), *The Experience of Learning*. Edinburgh: Scottish Academic Press.

ɨill, I. & Beaty, L. (1995). Action learning: A guide for professional management and educational development (2nd ed.). London: Kogan Page.

nnis, C. (2000). Changing academic work roles: The everyday realities challenging quality in teaching. *Quality in Higher Education*, *6* (2), 143-152.

:eachie, W. J. (1996). Do we need norms of student ratings to evaluate faculty? *Instructional Evaluation and Faculty Development*, *16* (1), 14-17.

:eachie, W. J. (1997a). Good teaching makes a difference – and we know what it is. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 396-408). New York: Agathon Press.

:eachie, W. J. (1997b) Student ratings: The validity of use. *American Psychologist*, *52*, 1218-1225.

:eachie, W., Lin, Y., Daugherty, M., Moffett, M., Neigler, C., Nork, J., Walz, M., & Baldwin, R. (1980). Using student ratings and consultation to improve instruction. *British Journal of Educational Psychology*, *50*, 168-174.

:h, L. R. (2000a). The stateless art of teaching evaluation. D. DeZure (Ed.) *Learning from change: Landmarks in teaching and learning in higher education from Change magazine, 1969-1999* (p. 347-349). London: Kogan Page.

(Reprinted from *Change*, 1976).

Meeth, L. R. (2000b). What is improved teaching? In D. DeZure (Ed.), L*earning from change: Landmarks in teaching and learning in higher education from Change magazine, 1969-1999*, (pp. 349-351). London: Kogan Page. (Reprinted from *Change*, 1977)

Menges, R. J. (1997). Foreword and acknowledgement. In K. T. Brinko, & R. J. Menges (Eds.), *Practically speaking: A sourcebook for instructional consultants in higher education*. Stillwater, OK.: New Forums Press.

Menges, R. J. (2000). Shortcomings of research on evaluating and improving teaching in higher education. In K. E. Ryan (Ed.), *Evaluating teaching in higher education: A vision for the future*. New Directions for Teaching and Learning (No. 83, pp. 5-11). San Francisco: Jossey-Bass.

Menges, R. J., & Austin, A. E. (2001). Teaching in higher education. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1122-1156). Washington, DC: American Educational Research Association.

Menges, R. J., & Brinko, K. T. (1986). *Effects of student evaluation feedback: A meta-analysis of higher education research*. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Mezirow, J. (1991). Transformative dimensions of adult learning. San Fransisco: Jossey-Bass.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. London: Sage.

Miller, M. T. (1971). Instructor attitudes toward, and their use of student ratings of teachers. *Journal of Educational Psychology*, *62*, 235-239.

Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457-483). New York: Russell Sage Foundation.

Morgan, G. A., & Griego, O. V. (1998). *Easy use and interpretation of SPSS for windows: Answering research questions with statistics*. Mahwah, MJ: Lawrence Erlbaum.

Morrison, D. E. (1995). Opening doors to better teaching: The role of the peer-based instructional consultation. *Dissertation Abstracts International* 56, 2134A (UMI No. 9535493).

Morrison, D. E. (1997). Overview of Instructional Consultation in North America in K. T. Brinko & R. J. Menges (Eds.), *Practically speaking: A sourcebook for instructional consultants in higher education*. Stillwater OK: New Forum Press.

Morrison, K. (2001). Randomised controlled trials for evidence-based education: Some problems in judging 'what works'. *Evaluation and Research in Education*, *15* (2), 69-83.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *International Journal for Academic Development*, *2* (1), 8-23.

Murray, H. G., Rushton, P. J., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology*, *82*, 250-261.

Neumann Y., & Neumann, L. (1981). Determinants of students' satisfaction with coursework: An international comparison between two universities. *Research in Higher Education*, *14* (4), 321-333.

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education*, *26* (2), 135-146.

North, J. D. (1999). Administrative courage to evaluate the complexities of teaching. In P. Seldin & Associates, *Changing practices in evaluating teaching*: *A practical guide to improve faculty performance and promotion/tenure decisions* (pp. 183-193). Bolton, MA: Anker Publishing.

Oakley, A. (1998). Experimentation and social interventions: A forgotten but important history. *British Medical Journal*, *317*, 1239-1242.

Ory, J. C. (2000). Teaching evaluation: Past, present, and future. In K. E. Ryan (Ed.), *Evaluating teaching in higher education: A vision for the future*. New Directions for Teaching and Learning (No. 83, pp. 13-18). San Francisco: Jossey-Bass.

Ory, J. C., & Ryan, K. (2001). How student ratings measure up to a new validity framework? In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* New Directions for Institutional Research (No. 109, pp. 27-44). San Francisco: Jossey-Bass.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students cognitive and affective outcomes. *Journal of Educational Psychology*, *71* (6), 856-865.

Pambookian, H. S. (1974). Initial level of student evaluation of instruction as a source of influence on change after feedback. *Journal of Educational Psychology*, *66*, 52-56.

Paterson, L., & Goldstein, H. (1991). New statistical methods for analysing social structures: An introduction to multilevel models. *British Educational Research Journal*, *17* (4), 387-393.

Patrick, J., & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: The emergence of three critical factors. *Assessment & Evaluation in Higher Education*, *23* (2), 165-178.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). London: Sage Publications.

Payne, D. A., & Hobbs, A. M. (1979). The effect of college course evaluation feedback on instructor and student perceptions of instructional climate and effectiveness. *Higher Education*, *8*, 525-533.

Piccinin, S. (1999). How individual consultation affects teaching. In C. Knapper & S. Piccinin (Eds.), *Using consultants to improve teaching*. New Directions for Teaching and Learning (No. 79, pp. 71-84). San Francisco: Jossey-Bass.

Piccinin, S., & Moore, J. (2002). The impact of individual consultation on the teaching of younger versus old faculty. *The International Journal for Academic Development*, *7* (2), 123-134.

Piccinin, S., Cristi, C., & McCoy, M. (1999). The impact of individual consultation on student ratings of teaching. *International Journal of Academic Development*, *4* (2) 75-88.

Plewis, I. (1998). Multilevel models social research. *Update*, 23, University of Surrey, Department of Sociology, UK.

Powney, J., & Hall, S. (1998). *Closing the loop: The impact of student feedback on students' subsequent learning*. Edinburgh: Scottish Council for Research in Education.

Rader, B. R. (1995). The effect of consultation on nursing educators' student ratings of instruction. Dissertation Abstracts International (UMI No. 9606244) Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Ramsden, P. (1992). *Learning to teach in higher education*. London: Routledge.

Ramsden, P. (1998), *Learning to lead*. London: Routledge.

Ramsden, P., Margetson, D., Martin, E., & Clarke, S. (1995). Recognizing and rewarding good teaching in Australian Higher Education. Canberra: Committee for the Advancement of University Teaching. Retrieved November 11, 2001 from http://www.autc.gov.au/caut/rrgt/Chapter2.html.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). London: Sage Publications.

Raudenbush, S. W., & Willms, J. D. (Eds.). (1991). *Schools, classroom, and pupils: International studies of schooling from a multilevel perspective*. London: Academic Press.

Redding, R. E. (1998). Students' evaluation of teaching fuel grade inflation. *American Psychologist, 53* (11), 1227-1228.

Richardson, V. & Placier, P. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 905-947). Washington, DC: American Educational Research Association.

Richlin, L. (2001). Scholarly teaching and the scholarship of teaching. In C. Kreber (Ed.), *Scholarship revisited: Perspectives on the scholarship of teaching*. New Directions for Teaching and Learning (No. 88, pp. 57-68). San Francisco: Jossey-Bass.

Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education*, *42* (4), 377-399.

Robinson, V. (1992). Why doesn't educational research solve educational problems? *Educational Philosophy and Theory*, *24* (2), 8-28.

Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self-concept. *Instructional Science*, *28*, 439-468.

Rosenberg, M.S., Adams, D. C., & Gurevitch, J. (2000). *MetaWin: Statistical Software for Meta-analysis*. Version 2.0. Sunderland, MA: Sinauer Associates.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results, *Psychological Bulletin*, 86 (3), 638-641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Applied Social Research Method Series, Vol. 6. Thousand Oaks, CA: Sage.

Rosenthal, R. (1994). Statistically describing and combining studies. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Russell Sage Foundation.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118* (2), 183-192.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59-82.

Rotem, A. (1978). The effects of feedback from students to university instructors: An experimental study. *Research in Higher Education*, *9*, 303-318.

Rotem, A., & Glasman, N. S. (1979). On the effectiveness of students' evaluative feedback to university instructors. *Review of Educational Research, 49*, 497-511.

Rowland, S. (2000). *The enquiring university teacher*. Buckingham: SRHE/Open University.

Rowland, S. (2001). Surface learning about teaching in higher education: The need for more critical conversations. *International Journal for Academic Development*, *6* (2), 162-167.

Rozeman, J. E., & Kerwin, M. A. (1991). Evaluating the effectiveness of a teaching consultation program on changing student ratings of teaching behaviours. *Journal of Staff, Program, and Organizational Development*, *9* (4), 223-230.

Sachs, J. (2000). The activist professional. *Journal of Educational Change*, *1*, 77-95.

Samuelowicz, K., & Bain, J. D. (2001). Revisiting academics' beliefs about teaching and leaning. *Higher Education*, *41*, 299-325.

Santhanam, E., & Hicks, O. (2002). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education*, *7* (1), 17-31.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, *2* (4), 419-436.

Schön, D. A. (1983). *The reflective practitioner*. New York: Basic Books.

Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Schulz, K. F., & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: Chance not choice. *The Lancet*, *359*, 515-519.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research and Evaluation*, *4* (7). Retrieved October 15, 2001 from http://ericae.net/pare/getvn.asp.

Seldin, P. (1993, July 21). The use and abuse of student ratings of professors. *The Chronicle of Higher Education*, p. A40. Retrieved from July 10, 2001 from http://trc.ucdavis.edu/trc/ta/TAdevel/seldin.pdf.

Seldin, P. (1999a). Current practices-good and bad-nationally. In P. Seldin & Associates, *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions*. Bolton, MA: Anker Publishing.

Seldin, P. (1999b). Self-evaluation: what works? What doesn't? In P. Seldin & Associates, *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions*. Bolton, MA: Anker Publishing.

Seldin, P. (2000). Evaluating college teaching: Myth and reality. In D. DeZure (Ed.), *Learning from Change: Landmarks in Teaching and Learning in Higher Education from Change Magazine, 1969-1999*. London: Kogan Page. (Reprinted from *Change*, 2000).

Senge, P. M. (1993). *The fifth discipline: The art and practice of the learning organisation*. London: Century Business (Original work published 1990).

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-280). New York: Russell Sage Foundation.

Shapiro, E. G. (1990). Effect of instructor characteristics on students' class evaluations. *Research in Higher Education*, *31* (2), 135-148.

Shelvin, M., Banyard, P., Davies, M., Griffith, M. (2000). The validity of student evaluations of teaching in higher education: Love me, love my lectures. *Assessment & Evaluation in Higher Education*, *25* (4), 397-405.

Shulman, L. S. (2000). Teaching as community property: Putting an end to pedagogical solitude. In D. DeZure (Ed.), *Learning from change: Landmarks in teaching and learning in higher education from Change magazine, 1969-1999*. London: Kogan Page. (Reprinted from Change, Nov/Dec, 1993)

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, *22* (3), 199-213.

Smeby, J. (1996). Disciplinary differences in university teaching. *Studies in Higher Education*, *21* (1), 69-79.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Sage Publications.

Sorcinelli, M. D. (1997). The teaching improvement process. In K. T. Brinko & R. J. Menges (Eds.), *Practically speaking: A sourcebook for instructional consultants in higher education* (pp. 157-158). Stillwater OK: New Forum.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8* (50). Retrieved August 25, 2001 from http://epaa.asu.edu/epaa/v8n50.html.

Stake, R. E. & Cisneros-Cohernour, E. J. (2000). Situational evaluation of teaching on campus. In K. E. Ryan (Ed.) *Evaluating teaching in higher education: A vision for the future.* New Directions for Teaching and Learning (No. 83, pp. 51-72). San Francisco: Jossey-Bass.

Stanley, C. A., Porter, M. E., & Szabó, B. L. (1997). An exploratory study of the faculty developer-client relationship. *Journal of Staff, Program and Organizational Development*, *14* (3), 115-126.

Stevens, J. J., & Aleamoni, L. M. (1985). The use of evaluative feedback for instructional improvement: A longitudinal perspective. *Instructional Science*, *13*, 285-304.

Stone, J. E. (1995). Inflated grades, inflated enrolment, and inflated budgets: an analysis and call for review at the state level. *Education Policy Analysis Archives*, *3* (11). Retrieved November 15, 2001 from http://olam.edu.asu.edu/epaa/arch.html.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. (3rd ed.). New York, NY: Harper Collins.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oakes, CA: Sage Publications.

Taylor, P. G. (1999). *Making sense of academic life: Academics, universities and change*. Buckingham: SRHE/Open University Press.

Theall, M. (1996). "Who is norm, and what does he have to do with student ratings? A reaction to McKeachie. *Instructional Evaluation and Faculty Development*, *16* (1), 7-9.

Theall, M., & Franklin, J. (1997). Collecting information using student ratings. In K. T. Brinko & R. J. Menges (Eds.), *Practically speaking: A sourcebook for instructional consultants in higher education* (pp. 73-92). Stillwater OK: New Forum.

Theall, M., & Franklin, J. (2000). Creating responsive student ratings systems to improve evaluation practice. In K. E. Ryan (Ed.) *Evaluating teaching in higher education: A vision for the future*. New Directions for Teaching and Learning (No. 83, pp. 95-108). San Francisco: Jossey-Bass.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of teaching. In M. Theall, P. C. Abrami, & L. A. Mets (Eds.), *The student ratings debate: Are they valid? How can we best use them?* New Directions for Institutional Research (No. 109, pp. 45-56). San Francisco: Jossey-Bass.

Tiberius, R. G. (1995). From shaping performances to dynamic interaction: The quiet revolution in teaching improvement programs. In W. A. Wright & Associates, *Teaching improvement practices: Successful strategies for higher education* (pp. 180-203). Bolton, MA: Anker Publishing.

Tiberius, R. G., Sackin, H. D., Slingerland, J. M., Jubas, K., Bell, M., & Matlow, A. (1989). The influence of student evaluative feedback on the improvement of clinical teaching. *Journal of Higher Education*, *60* (6), 665-681.

Torgerson, C. J., & Torgerson. D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, *49* (3), 316-328.

Trigwell, K. (2001). Judging university teaching. *The International Journal for Academic Development*, *6* (1), 65-73.

Trigwell, K., & Prosser, M. (1996). Changing approaches to teaching: a relational perspective. *Studies in Higher Education*, *21* (3), 275-284.

Trout, P. (2000, July/August). Flunking the test: The dismal record of student evaluations. *Academe*, *86* (4). Retrieved on February 16, 2002 from www.aaup/org/publications/Academe/ooja/JA00/Trou.htm.

Tukey, J. W. (1977). Exploratory data analysis. London: Addison-Wesley.

Tymms, P. (1993). Accountability—can it be fair? *Oxford Review of Education*, *19* (3), 291-299.

Tymms, P. B. (1999). *Baseline assessment and monitoring in primary schools: Achievement, attitudes and value-added indicators*. London: David Fulton.

UNESCO (1998). *Higher education in the 21st century: Vision and action*. World Conference on Higher Education. Paris, France,

UNESCO (2002). *Gross enrollment ratio at tertiary level for academic years 1998/1999, 1999/2000 and 2000/2001*. UNESCO Institute for Statistics. Retrieved from December 12, 2003 from http://www.uis.unesco.org.

Van Manen, M. (1995). On the epistemology of reflective practice. *Teachers and Teaching*, *1* (1), 33-50

Villaescusa, T. K., Franklin, J., & Aleamoni, L. M. (1997). *Improving the interpretation and use of student ratings: A training approach*. Paper presented at 78th annual meeting of the American Educational Research Association, Chicago.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*, 191-212.

Watkins, D. (1994). Student evaluations of teaching effectiveness: A cross-cultural perspective. *Research in Higher Education*, *35*, 251-266.

Webb, G. (1996). *Understanding staff development*. Buckingham: SRHE/Open University.

Weimer, M. & Lenze, L. F. (1997). Instructional interventions: A review of the literature on efforts to improve instruction. In R. P. Perry, & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 205-250). New York: Agathon Press. (Reprinted from *Handbook of theory and research*, Vol. 7, 1991).

Weimer, M. (1997). Exploring the Implications: From research to practice. In R. P. Perry and J.C. Smart (Eds.) *Effective teaching in higher education: Research and practice*. New York: Agathon Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Wenger, E. C. (2000). Communities of practice and social learning systems. *Organisation*, *7* (2), 225-256.

Wenger, E. C., & Synder, W. M. (2000). Communities of practice: The organisational frontier. *Harvard Business Review*, *78* (1), 139-145.

Wildman, T., & Niles, J. (1987). Reflective teachers: Tensions between abstractions and realities. *Journal of Teacher Education*, *38* (4), 25-31.

Williams, W. M., & Ceci, S. J. (1997). "How'm I doing?": Problems with student ratings of instructors and courses. *Change, 29* (5), 13-23.

Wilson, R. (1998). New research cast doubt on value of student evaluations of professors! *The Chronicle of Higher Education: Colloquy*. Retrieved June 18, 2002 from http://chronicle.com/colloquy/98/evaluation/background.htm.

Wilson, R. C. (1986). Improving faculty teaching: Use of student evaluations and consultants. *Journal of Higher Education, 57* (2), 196-211.

Wilson, S. J., & Lipsey, M. W. (2000). Wilderness challenge programs for delinquent youth: A meta-analysis of outcome evaluations. *Evaluation and Programme Planning, 23*, 1-12.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. London: Sage Publications.

Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education, 27* (1), 49-64.

Worthington, A. G., & Wong, P. T. P (1979). Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology, 71*, 764-775.

Wragg, E. C., Haynes, G. S., Wragg, C. M., & Chamberlin, R. P. (2000). Failing teachers? London: Routledge.

Wright, W. A., & O'Neil, M. C. (1995). Teaching improvement practices: International Perspectives. In W. A. Wright & Associates, *Teaching improvement practices: successful strategies for higher education* (pp. 1-57). Bolton, MA: Anker.

Young, S., & Shaw, D. G. (1997). Effective teaching: The effect of student status and gender on teacher ratings. *Newsletter of the American Educational Research Association Special Interest Group in Faculty Teaching, Evaluation and Development (SIGFTED)*. Retrieved November 10, 2002 from http://www.uis.edu/~ctl/specialtopics/sigfted/s97htm.html.

Young, S., & Shaw, D. G. (1999). Profile of effective college and university teachers. *Journal of Higher Education, 70* (6), 670-686.

Zeichner, K. (2001). Educational action research. In P. Reason & H. Bradbury (Eds.), *Handbook of action research: Participatory inquiry and practice*. London: Sage Publications.

Zuber-Skerrit, O. (Ed.). (1996). *New Directions in Action Research*. London: Falmer Press.

Achievement

"I do my work". He who in faith can say

That simple phrase, is set upon the way

To bend the will of Fortune to his will.

The world makes place for him whose strength and skill

Rebel at doubt and rankle at delay.

The visions that hold true, the dreams that stay

Are wrought by those who labour, come what may.

Their slogan--be their fortune good or ill--

"I do my work".

Kingdoms may fall and empires lose their sway,

But on their wreck and out of their decay

The toiler shall erect new wonders still,

Urged by an impulse time nor fate can kill,

With this his only vaunt from day to day--

"I do my work".

Published in: Adventure Magazine - April 18, 1921