TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
Julkaisu 728 • Publication 728

Eila Pajarre

Online Evaluations in Higher Education
Utilising Information Systems in Assessing the Validity of Student Evaluations

Tampere 2008

Tampereen teknillinen yliopisto. Julkaisu 728
Tampere University of Technology. Publication 728

Eila Pajarre

Online Evaluations in Higher Education

Utilising Information Systems in Assessing the Validity of Student Evaluations

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Festia Building, Auditorium Pieni Sali 1, at Tampere University of Technology, on the 11th of April 2008, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2008

ISBN 978-952-15-1945-1 (printed)
ISBN 978-952-15-2047-1 (PDF)
ISSN 1459-2045

Abstract

Practically all university students have responded to at least some student evaluations in the course of their studies. For their part, most lecturers also collect them in one way or another. The use of student evaluations is extensive, ranging from developing a single course to even salary and tenure decisions of the whole faculty. Because of the significance of the evaluation results, also the validity and reliability of student evaluations have become issues of heated debate. Many research findings have been published dealing with such topics, yet the debate still goes on unresolved.

During the last decade, many universities have replaced traditional paper-based evaluations with online evaluation systems, but in most cases this has simply meant converting the existing paper-based questionnaires into electronic format. Also research into online ratings has focused mainly on investigating the differences between online and paper-and-pencil ratings systems and on describing experiences of the developed online evaluation systems. What has not hitherto been recognised is the potential of information technology to implement and analyse the evaluations.

This study concentrates on assessing the additional value to be gained by connecting online student evaluations to universities' other databases, especially to analyse the validity of student ratings. The study approach follows the principles of decision-oriented research approach. After initial determination of the research problem, the first phase involved identifying the gaps and contradictions in the current knowledge in the field. Next, an online survey was developed and implemented, the results of which were linked up (with respondents' permission) to other university databases for further analysis. These were then compared with previous research findings.

The results of this study suggest that an interlinked system can aid in assessing the validity of student evaluations by enabling observations at an individual respondent level instead of using averages and estimations. Such a system also reduces the need for complex statistical analyses which demand expertise to be interpreted correctly. This study also gives examples of issues which either have been actively researched but unsatisfactorily resolved or which have earlier been impossible to analyse. For example, this study demonstrates that it is possible to analyse individual final course grades connected to the individual students' survey responses even at a single questionnaire item level. The effect of the order in which individual students undertake their courses connected to the same students' evaluation results and their individual final grades from the same courses, was an example of an issue not previously investigated at the individual level.

Keywords: Student Evaluations, Online Ratings, Validity, Higher Education

Acknowledgements

In this dissertation, the feedback gathered from students has played a crucial role. In higher education, as in life, we collect feedback when we want to improve something. Often in higher education that "something" is ultimately students' learning. Likewise, during this research process, feedback provided by my supervisor, colleagues and others, has been essential. Without their comments, help and encouragement I would have been overwhelmed by the sheer scope of the topic.

There are several people I wish to thank. My supervisor, Professor Erkki Uusi-Rauva, has encouraged and supported me in my research work from the outset and provided much valuable help, especially with methodological issues. Professor Asko Miettinen also provided support for which I am very grateful.

My warmest thanks are due to TUT system designer Jaakko Ruohtula, who implemented the online survey that was designed and studied in this thesis. In fact, the idea of the potential to connect university's online evaluation system to university's other data systems had already come up in our discussions while writing my M. Sc. thesis. This dissertation work has made it possible to examine the diverse potential of this connection.

I am also especially indebted to pre-examiners Dr. Pirkko-Liisa Vesterinen and Professor Risto Salminen for their constructive comments and guidance. Their help, especially during the final stages, was invaluable.

The support of the staff of the Department of Industrial Management and of the Faculty of Business and Technology Management has been important. I owe my thanks to Development Manager Jukka Annala and to all colleagues at the department. I would especially like to thank secretaries Sirpa Järvenpää, Annamaija Paunu-Virtanen and Marita Nikkanen, not only for their help with everyday work but also for being good friends. Their cheerfulness brightened up the working days. I also wish to thank Mrs. Hanna-Kaisa Desavelle for her supporting comments. And thanks to my colleagues Mika Ojala and Mikko Kaataja with whom I've shared a work room for the last few years. It has been a privilege to work in such an atmosphere!

For funding this thesis, I gratefully acknowledge the Finnish Graduate School of Industrial Engineering and Management and Tampere City Science Foundation. I would also like to thank Alan Thompson for proofreading the manuscript. I am also grateful to Virpi Hämäläinen for retrieving the official TUT student statistics needed in this study.

Most of all I wish to thank my mother and my family. My husband Eero provided vital help in writing the Perl script which was used when analysing the survey results of this thesis. Even more important has been his inexhaustible support and encouragement over the years. To my sons, Markus and Matias, go my heartfelt thanks for never letting me forget what is most important in life.

Tampere 19th March 2008

Eila Pajarre

Table of Contents

1 INTRODUCTION

> *There are probably more misconceptions about student ratings than facts about them, yet we do know quite a bit.*
>
> – John C. Ory (2001) –

For an issue claimed to be one of the most widely researched topics in the field of higher education (e.g., Marsh 1987; Cashin 1995; McKeachie & Kaplan 1996; Wilson 1998; Menges 2000; Franklin 2001), opinions on any elements of student evaluations are surprisingly controversial. Moreover, there is no consensus on the validity or reliability of the results of the ratings and there is no uniform view of how the results of the ratings should be utilised. It has been estimated that there are probably more studies on student ratings than on any other aspect of higher education (Cashin 1995; McKeachie 1996).

The history of student ratings, at least in their present form[1], dates back to the 1920s when student evaluation programs were introduced at Harvard and several other U.S. universities (Marsh 1987; Wachtel 1998; Kulik 2001; Parjanen 2003). The first research studies on student ratings were published in the 1920s (Marsh 1987; Kulik 2001; Dommeyer et al. 2002b) and since the 1950s there are more than 2000 scientific articles dealing solely with student ratings (McKeachie & Kaplan 1996; Wilson 1998; Centra 2003; Deasy 2004).

Today, student evaluations of teaching are commonplace in most universities. Both the dramatic growth in higher education and the public pressure towards increasing accountability have given rise to the use of various evaluation methods (Leathwood & Phillips 2000). Student evaluations have become one of the most frequently used performance indicators, not only for improvement use among faculty, but all more for administrative purposes, such as tenure decisions or even public comparisons between universities. (Mason et al. 1995; Jackson et al. 1999; Abrami 2001; Kulik 2001) The importance of student evaluations is emphasized by the fact that in many universities student ratings are used as one, sometimes only and the most influential, measure of teaching effectiveness (Seldin 1993; Kwan 1999). This sets strong demands on the validity and reliability on the evaluation methods being used.

[1] Several scholars (e.g., Marsh 1987; Platt 1993) have compared the differences between the feedback mechanisms of education in the ancient Greece and the prevailing practices in modern universities.

During the past decade, universities worldwide have moved increasingly from traditional paper-and-pencil based ratings towards online rating systems (e.g., Upcraft & Wortman 2000; Heerwegh & Loosveldt 2002; Romano & Himmelmann 2002; Zimitat & Crebert 2002). There is, however, a lack of comprehensive research into Internet based student surveys in higher education. Even though during the last few years online ratings have received increasingly more attention among scholars and within universities, there are still wide gaps concerning the possibilities of utilising modern technology in evaluating education. Most of the current online students rating systems are merely traditional course questionnaires transferred to an Internet context (see e.g., OnSET 2005). The potential for connecting online ratings into other university data systems has usually not been realised.

The validity of student ratings is still one of the most controversial issues among the researchers of evaluation in higher education. The purpose of the present study, therefore, is to introduce a new online rating method which will enable analyses of at least some of the validity issues that have so far remained unresolved.

1.1 Introduction to Feedback Systems in Higher Education Evaluation

Various assessment and evaluation methods have become a familiar part of university life worldwide. Pressure to adopt these evaluation methods has come both from within universities and from the public sector. Furthermore, the establishment of national evaluation institutes, such as the Finnish Higher Education Evaluation Council (FINHEEC) in Finland, as well as the growing number of university students have highlighted the need for assessing the quality of teaching and learning in higher education institutes (Lappalainen 1997). Not only has the actual number of students increased, but the students also come from all more heterogeneous backgrounds. As Beaty (2001, p. 75) observes, managing the teaching is becoming ever more difficult as "fewer students have predictable prior knowledge or predictable futures". Thus the vast increase in various evaluation methods is hardly surprising. It has been estimated that whereas only about 30 percent of universities collected student evaluations at the beginning of the 1970s, it is now commonplace at most higher education institutes (Wilson 1998).

The scope of the evaluations as well as the methods used in carrying out evaluations varies widely. Harman (1998, p. 335) divides the most common mechanisms of evaluation into "horizontal reviews" of disciplines and "vertical" institutional evaluations. Kellaghan et al. (2003) describe how the scope of evaluation ranges from the individual student evaluations carried out by their lecturers to "evaluations of schools and districts, to district-wide program evaluations, to national assessments and to the cross-national comparisons of student

achievement". According to the authors, educational evaluation encompasses a wide array of activities, such as student assessment, measurement, testing, program evaluation, personnel evaluation, curriculum evaluation as well as institute accreditation (Kellaghan et al. 2003). The Figure 1 below depicts the focus of evaluation at different levels.

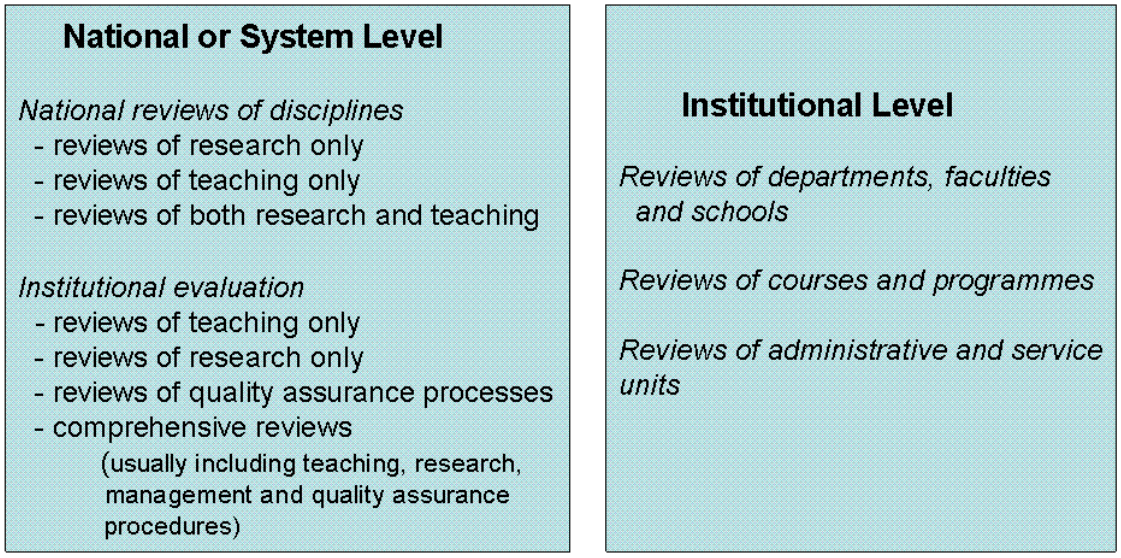


Figure 1. Focus of evaluation in different levels of higher education (Harman 1998, p. 334).

In spite of the diversity of evaluation methods, student ratings are universally regarded as the major way of evaluating education in most universities (see e.g., Cannon 2001a; Saroyan & Amundsen 2001; Kember et al. 2002; Nasser & Fresko 2002). Originally taken into use in North America, various student rating forms are now also widely used in Canada, Australia, many Asian countries and in most countries in Europe (Ha et al. 1998; Husbands 1998; Theall & Franklin 2000; Harvey 2001; Hendry & Dean 2002; Bhola 2003; Chen & Hoshower 2003). However, there are country specific differences in the use of evaluation results, for example evaluation of teacher performance is common practice in the USA, but less common in the UK (Brennan & Williams 2004). Research into evaluation in higher education has been most active in North America, Australia and United Kingdom, in contrast to many countries in Europe, such as Austria and Italy, where research has been rather limited (Boffo & Moscati 1998; Greimel-Fuhrman & Geyer 2003).

Ory (2000, p. 13) has described the changes in the purpose and methodology of teaching evaluation in higher education over the past thirty years. According to this study, teaching evaluation in the 1960s was conducted primarily in response to student demands for public accountability whereas in the 1970s the focus of such evaluation was on helping faculty to improve and to develop. In the 1980s and 1990s student evaluations were motivated more by administrative than faculty or student needs, mostly because as a results of public pressure for greater university accountability (see also Cheng 2001). This pressure from the public

sector for evaluating quality by focusing on performance indicators such as student evaluations and other monitoring processes is an international phenomenon characteristic not only of the USA but all modern countries (for more information see e.g., Erwin 1991; Dill 1998; Harman 1998; Alexander 2000; Braskamp 2000; Välimaa 2000; McAlpine & Harris 2002; Ballantyne 2003; Harvey & Askling 2003; Ramsden 2003; Brennan & Williams 2004). As Barrie (2001) notes, formerly many academic units were concerned primarily with *improving* teaching and learning at the level of individual teachers or courses, whereas nowadays they are increasingly being called upon to *prove* teaching and learning quality at an institutional level.

In Finland the use of student evaluations within universities has become well-established in recent decades. In common with U.S., for example, the focus of evaluation has traditionally been at the course level (e.g., Lappalainen 1997, p. 26). However, unlike the U.S. where the focus of evaluation is mainly on the characteristics of the teacher (Brennan & Williams 2004), in Finland the questionnaires have aimed at providing a more holistic overview of the course characteristics (Lappalainen 1997, pp. 26–27). There have been attempts to build a uniform feedback system for universities nationally, such as KOLA-program (Takala 1993), but in practice each university seems to have its own particular feedback systems.[2]

In the present study the focus is on institutional student evaluations and especially on the question, what are the possibilities of utilizing modern technology in developing and analysing student ratings, not only at the course level, but also at a general level. Since most of the published research findings on student evaluations have concentrated on student evaluations of teaching effectiveness, the terms and definitions used in this study have been aligned with the established terminology of aforementioned articles. However, it should be noted that although these previous studies have mainly been focused on few courses or few lecturers, in the present study the target of student evaluations is broader than a single course or a single teacher.

1.2 Research Problem and Objectives of the Study

The purposes and uses of student evaluations vary markedly from one country to another[3] and even from one university to another. Despite this and the fact that student evaluations have been actively studied for as long as they have been used in their current form, the most

[2] The heterogeneity in gathering feedback was shown e.g., in a FINHEEC seminar "Opiskelija opetuksen laadunarvioinnissa" in 23rd January 2006 where several universities described their feedback systems.
[3] There are huge differences between different countries even within Europe. Boffo & Moscati (1998, p. 358) explain the resistance towards evaluation of the teaching activities resulting from "the absence of any real analysis of productivity related to the students' evaluation of courses". These differences will be discussed more thoroughly later in this study.

commonly posed question concerning student evaluations in higher education has prevailed for decades: Are student evaluations of education valid and reliable?

Several researchers have published studies attempting to prove the validity and reliability of student ratings (e.g., Marsh 1987; Ramsden 1991; Cashin 1995; Centra 2003). However, there are still a vast number of sceptics who question the usability of student evaluations or the students' ability to evaluate their teachers and the education they receive (Platt 1993; Becker 2000; Sproule 2000; Felton et al. 2004, to name but a few). Perhaps the longest ongoing debate on the validity of student evaluations centres on the grading satisfaction hypothesis, also known as grading leniency, i.e. the potential that course grades correlate positively with course evaluation ratings (Marsh 1987; Greenwald & Gillmore 1997a). Another question frequently raised by scholars and practitioners in the field of higher education is, whether the students who are doing poorly in their courses are able to judge their instructors objectively (see e.g., Adams 1997). A topic that has gained even more attention among researchers and practitioners in the field of higher education concerns the possibility that background characteristics or other external factors could bias student ratings.

The mere suspicion of any potential bias has provoked distrust in the validity of student ratings. Despite the extensive published research in the field examining these biases and underlying factors there is so far no consensus that would confirm or refute the validity of student ratings in terms of such potential biases (e.g., Nasser & Fresko 2006). Typical of these publications is that they deal with the student evaluations of single courses and single teachers, which is the predominant way of gathering feedback from students in most universities. In order to ensure an adequate sample of student evaluations data, one study has typically contained data from several course ratings and from thousands of students (e.g., Ramsden 1991; Centra 2003). There is, however a lack of research into the opinions of students concerning the whole curriculum, i.e. where all the courses completed for their degree, would be compared with the actual grades they received from the courses.

The main reason that it has been difficult to observe the relationships between students' responses and their background characteristics is largely because student ratings are typically collected anonymously. This has made it virtually impossible to analyse potential biasing factors at the individual level since the needed references such as students' grades have been available at the class-level only. Erwin (1991, p. 150) describes the problem as follows: "[T]his practice weakens assessment studies because such data cannot be linked to other available information about a particular student. For example, it is useful to compare test scores with courses completed, but such comparisons cannot be made unless individual test takers can be identified". More specifically, Carey (1993) has argued that there should be some way to correlate the evaluation of a given student with that student's overall grade-point average. Erwin's and Carey's comments are of extreme importance since they address the fundamental problem with existing student evaluation methods in explicit terms. Both

agree that without a straight linkage between individual students' responses and the same students' individual background data, the analyses of potential biases will always be based on inaccurate data such as class-level averages and estimates. And what makes the question of bias the subject of hundreds of research reports (see e.g., Parjanen 2003, 15) is that irrespective of whether potential biases actually impact student ratings, their utilization will be hindered as long as faculty *think* they are biased (Marsh 1987, p. 305).

The emphasis in this study is on examining the opportunities that new information technology offers for analysing the validity of student evaluations. To date, there has been very little research into the potential benefits of transferring traditional paper-based student ratings to an online environment. Partially this can be explained by the slow absorption rate of technological progress within universities. Turner & Stylianou (2004), for example, claim that the "use of technology in the infrastructure of academia, while much improved over the last 10 years, is less than it could be and perhaps less than it should be". What has been recognized, however, is that in many institutions more use could be made of feedback data (Brennan et al. 2003, p. ii). This observation is especially important because there has been a substantial increase in all kinds of data collection in recent decades which threatens students (a phenomenon known as "questionnaire burnout"), faculty and the administrative personnel. Heywood (1988, p. 65) warned already twenty years ago of the risk that "institutions become overburdened by the sheer volume of data they are supposed to collect, and that they will not make effective use of data they have." Instead of collecting more and more data Brennan et al. (2003) suggest that consideration should be given to collecting less data and analysing it more thoroughly.

It is worth noting that in theory the technical connections of online ratings to other data systems have been registered and discussed already some years ago. In 1999 Kelly & Marsh (1999), for example, described how the online-based evaluation system could be linked to other university held data systems. In practice, though, the possibilities of utilising the connection between online ratings and other data systems have largely been neglected. Research into the technical characteristics of student evaluations on teachings in university instruction has been confined to the usefulness of multiple sections on rating forms and the relationships between ratings and other factors (Algozzine et al. 2004, p. 135); the emphasis being on using traditional analysis methods rather than actively seeking new ones. Brennan et al. (2003, p. ii), for example, have criticised the narrow scope of utilising evaluation results. They claim that even when there is a central unit with responsible for student feedback, the analysis of the feedback data is often quite limited, with little use of comparative analysis or relating feedback data to other institutional datasets.

There are few studies where the possibility of connecting students' responses directly to other data has been utilised. Though there is a good deal of research work which attempts to demonstrate the interdependence between students' responses and the characteristics of the

respondents, there has been no way of comparing individual students' responses with their individual background data. This is due to the anonymous nature of student evaluations. Previous studies have mostly been based on class-averages (Toland & de Ayala 2005, p. 272) because of an absence of any link between students' responses and their individualized background information. Correlating class-average student evaluations with a class-average measure of background variable hypothesized to bias student ratings can provide a heuristic approach. However, despite having been the most frequently approach, "in isolation it can never be used to demonstrate a bias". (Marsh 1987, p. 311)

Kerridge & Mathews (1998) report a study, where students were specifically asked to record their identification number on the questionnaire. The results were later manually linked to a university-held database to analyse potential relationships between student ratings of modules and the students' individual module grade outputs, admission criteria and gender. On the topic of online ratings, one of the very few studies, if not the only one, which have taken advantage of the connection between online ratings results and other databases is reported by Thorpe (2002). In this study, the main objective was not to analyse the connection between students' responses and their individual background but on rather to determine if there were any significant differences existed in student responses to a course evaluation instrument based on a delivery method, in class by paper versus online. A subset of the results of this study was also linked to the university's database, to analyse demographic differences among students in order to explain a potential non-response bias among the online respondents. This was done by comparing the online responses of the particular course with four background characteristics gathered from the university database.[4] However, there are no reports of studies dealing specifically with the connection between students' individual online responses and their individual university held background data.

It is, of course, possible to gather very detailed individual feedback from students in other ways than formal feedback questionnaires, by using personal in-depth interviews and so on. These kinds of interviews also have the advantage over structured surveys in that they can reveal important issues of which the developers of the survey have not been aware of. In the quest for potential biases, however, the existing research literature has relied on the *volume* of gathered responses. For example, in their work Marsh & Hocevar (1991) include a total of 24158 courses, the findings of Ramsden (1991) were based on 3372 student responses, and the research article of Centra (2003) was based on responses collected from 55 000 classes.[5] When the evaluation is aimed at thousands of students, it is evident that no other evaluation method except a questionnaire is possible. As discussed later in this study (Chapter 3.2.), the

[4] The results of Thorpe's (2002) findings suggested that no significant differences in student survey responses were found based on survey method (paper or online). It is worth noting is that in reporting the advantages of online survey technologies in the summary section there is no mention of the usefulness of the linkage between student responses and the background data. This is despite the fact that the data needed for analysing the nonresponses bias in the study was gathered using that linkage.

[5] The exact amount of classes varied from 46 687 to 55 155 depending on the evaluated items.

most useful way to gather information would be to use multiple sources of evaluation. The use of multiple evaluation methods, nevertheless, does not diminish the importance of survey results being as accurate and valid as possible.

This study concentrates on assessing the additional value to be gained by connecting online ratings to universities' other databases, especially in analysing the validity of student ratings. The objective is thus not limited only to presenting student survey results. The emphasis is at least equally or more on assessing the utility of the developed and implemented interlinked evaluation system. This approach might also be regarded as an "evaluation of an evaluation".[6] The purpose is to give new insights into the way the validity of student evaluations is observed. In contrast to many research reports (especially those conducted in the United States) the scope of the executed survey in this study is not limited to nor focused on an assessment of *teaching* effectiveness. On the contrary, the focus of interest here concerns the actual courses as well as the students' whole curricula and the responding students themselves, not the teachers and their characteristics.

With contemporary student evaluation methods, the most often frequently asked question concerns the validity and reliability of the evaluation results (discussed more thoroughly in Chapter 3). To prove the existence or non-existence of potential biases in student ratings, hundreds if not thousands of studies have been conducted (Marsh 1987, Centra 2003, Emery et. al. 2003). Due to the absence of a direct connection from students' responses to the same respondents' background information, the analyses have typically been – and still are – based on information such as class averages and estimated average course grades which have been analysed by factor analyses, multiple regressions and other statistical algorithms. As such these statistical algorithms are very efficient at identifying the factors the survey instrument is designed to measure (Marsh 1997, p. 266). However, basing factor analyses or other methods on averages and estimations can cause two kinds of problems. First, the background information gathered originally as a form of averages can never produce results as accurate as those gained by data collected at an actual individual level. Second, operating with statistical procedures requires an understanding of how the mathematical formulas work. If the researcher is not sufficiently familiar with the underlying principles, there is a serious risk of gaining misleading or erroneous results (see e.g., Dieks 1992; Ruskai 1997; Toivonen 1999). Moreover, even if the results (received from this average data) are mathematically correct, interpreting them also requires some familiarity with statistics on the part of the faculty (Aleamoni 1999; Franklin 2001).

[6] Evaluation of an evaluation is among practitioners of the field also known as meta-evaluation. Typically meta-evaluation is done by applying a certain evaluation-specific checklist (see e.g., Scriven 1991, pp. 228–229). In the present study, however, the aim is simply to assess the developed evaluation system, and thus it can not be called as meta-evaluation.

The purpose of this study is to examine the possibilities which linking online evaluations to other data systems bring to the assessment of the validity of student evaluations and especially the existence of potential biases, e.g., the grading leniency. The most important background factors to be compared with the students' responses are the students' actual individual course grades and the completion dates for every course which the individual students had completed.

In the present study, the research questions are as follows:

1. What kind of new information can be achieved by linking an online ratings system together with the university database and comparing individual students' responses to their individual background information, such as their actual course grades or the order in which they have proceeded in their studies?

2. How can the linking of an online student evaluation to the other university's databases enhance the assessment of the validity of student evaluations, especially in search of potential biases such as the grading leniency?

To respond to the above presented research questions, the research objectives have been formulated. The purpose is as follows:

I) To identify gaps and conflicts in the current knowledge on the validity of student evaluation results

II) To develop and build an online based student evaluation system connected with the university held database to gather responses from students linked with their individual background information

III) To analyse the gathered student responses together with their individual background information to assess the validity of students' responses

IV) To assess the value of received new research results, the usability of the research results and the developed online survey system at a more general level.

In the aforementioned stages III and IV the emphasis will be on analysing the problematics in showing the existence of potential grading leniency by comparing individual survey responses with individual actual course grades; analysing the effect of the implementation order of courses on the students' satisfaction and analysing the connection between students' individual course grades, individual opinions regarding the individual courses usefulness and how demanding the course has been.

It should be noted that the present study does not attempt to analyse the role of different learning models, e.g. objectivist and cognitive models, in developing the content of the student evaluation instrument. Several researchers, for example Serva & Fuller (2004) have described the deficiencies of evaluating learning by student evaluations employing traditional objectivist teaching and learning approaches. According to these authors, typical of many such objectivist evaluation measures is that they focus on the instructor's ability to present the reality and knowledge as a form of rote learning. However, in this study the focus is on the possibilities for utilising new technology to gather feedback from students and to assess the validity and reliability of the collected feedback. Thus the study is not tied to any particular learning approach, but addresses the utility of student background information, regardless of the models underlying the evaluation.

The technical details of existing online survey systems are not specifically dealt with in any depth in this study. This is because the emphasis is on presenting the potential value of utilising the results gained from an online survey linked up with the information from other databases, rather than on how such systems are technically implemented. Typically the implemented surveys are basically quite similar consisting of a customised HTML web page from which the survey results are recorded into a dedicated database in a university held server (see e.g., Andrews & Feinberg 1999; McGourty et al. 2002a). The software and hardware requirements needed to implement an online survey are discussed in greater detail elsewhere, for example by Schmidt (1997).

It is worth noting that new technology does not necessarily entail improvements, or as Lieberman et al. (2001) remark, that "new is not always better". Instead, this study has been conducted with their subsequent view in mind: "Electronic assessment and evaluation methods should be considered where they are either more efficient or perhaps more reliable than the traditional paper-and-pencil approach".

1.3 Research Design

Until few years ago there were very few doctoral studies in universities of technology dealing with educational issues. However, during the last years the need for investigating teaching and learning also in engineering has been recognized and several doctoral theses devoted to such issues in engineering have been published (e.g., Kolari 2003; Savander-Ranne 2003; Ala-Mutka 2005). Positioning research work in the field of engineering yet concentrating on aspects that have been traditionally associated mostly with pedagogics is no simple matter. Since this study was carried out in Tampere University of Technology at the Institute of Industrial Management and focuses on examining the possible added value of online ratings in assessing the validity of student ratings, it also comprises technological, management and

educational elements. Thus selecting an appropriate framework for this study was initially challenging. To give a more comprehensive picture of the challenges connected with the methodology selection, the subject is next discussed more extensively.

According to Uusitalo (1995, p. 42) any problem can be observed from more than merely one viewpoint and the function of a framework is to define the viewpoint for a particular research work. Scientific research in the field of industrial management is typically classified as belonging to the applied sciences (Olkkonen 1993). According to Olkkonen (1993, p. 19) the function of the applied sciences is to act as a bridge between the basic sciences and technology (practice), the value and results of the applied sciences can be measured by both epistemic and practical utilities, and their criteria are both truth and utility. Educational research, on the other hand, is often classified under the social sciences, where the particular term "social science" refers to the scientific study of human behaviour (e.g., Punch 2005).

In the case of the present study, even though the research subject can be approached both from the perspective of applied sciences and from the perspective of social sciences, the principles of applied sciences, especially of those in the field of industrial management, have guided the present research from the very beginning of this work. However, it has also been noted that all research has to be related to some scientific discussion (Kaikkonen 1996, p. 30). In this research study most of the published scientific reports concerning student evaluations come from social sciences. This means that even though the methodological approach in this work is dominated by an approach of applied sciences, the theoretical part of this work, which gives an overview of the researched phenomenon in general, is largely based on research findings reported by social scientists.

Several rules have been presented in the literature to guide the selection of research approaches and methods. These typically underline the importance of a structured and systematic approach. According to Punch[7] (2005, p. 24) a research project will be difficult to understand, and will lack credibility without structure in its research questions, its design (especially in its data) and also in its report. Gummesson (1993, p. 15) emphasises that the selection and use of methods must be related to the researcher's scientific paradigm, i.e. the platform on which one operates mentally. Uusitalo (1995, p. 50) stresses this by arguing that the research problem should be the dominant element when selecting the research setting. He also states that the research problem determines the kind of research material needed and how this should be analysed. Kauranen et al. (1992, p. 30) and Olkkonen (1993, p. 64–65) also highlight the importance of the research approach in the selection of the research method, especially in how information is obtained and processed. Kauranen et al. (1992, p. 29–30) add that the selection of research methods is also founded on the research approach prevailing in the researcher's scientific community.

[7] Although depicting the research process in social sciences, Punch's views concerning the process can be applied in other sciences as well.

In positioning this study, the premise that research work should be evaluated in terms of the context it in which it is carried out, has meant that the orientation of this work is towards the field of applied sciences. On the other hand, it has also been argued that the models of hermeneutic approaches in the field of business economics, for example, originate from humanities such as behavioural sciences (see e.g. Olkkonen 1993, p. 52). As a result the differences in approaches and methods between the applied and social sciences are often more in the terminology used rather than in their context. This viewpoint is supported by Kallio (2006, p. 511) who argues that the methodological problematics in management and organisational research do not significantly differ from the methodology used in the social sciences and pedagogics. Eskola & Suoranta (2000, p. 22) also observe that the differences between research methods are ultimately very minor. They suggest that the "hard" data obtained from survey research have originally been gathered by "soft" approaches, communicating linguistically, either orally or in written form.

In the field of business economics and industrial management little methodological research has been undertaken and the classification of research methods of Neilimo & Näsi (1980) remains one of the most fundamental (Kallio 2006, p. 511–515). The classification of Neilimo & Näsi (1980, p. 31), has been complemented by a fifth approach by Kasanen et al. (1993). Figure 2 shows the arrangements of these five approaches based on the positioning of Kasanen et al (1993, p. 257).

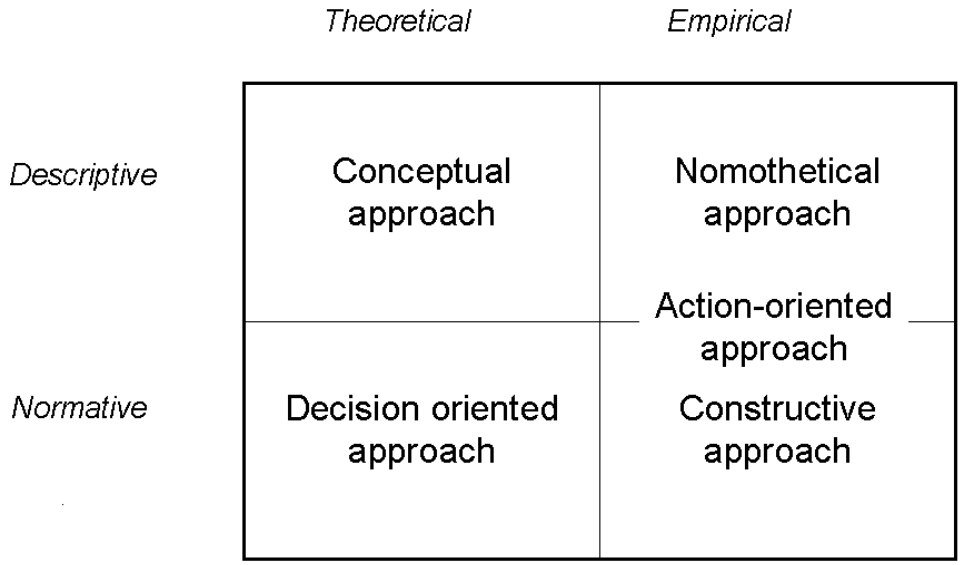


Figure 2. Positioning of typical research approaches in management accounting (Kasanen et al. 1993, p. 257).

Of these five approaches, both the constructive approach and the decision oriented approach have pragmatic starting points (Kasanen et al. 1993). According to Neilimo & Näsi (1980, p.

30) the focus of a decision oriented approach is *a solution to a certain problem*. Olkkonen (1993, p. 70) has stated that decision oriented research is typically used for developing models (often based on mathematics) that can be of assistance in a decision making process. The results are often in the form of mathematical or simulation models. These decision-making support models are usually meant for organisations to be used as such or to be customised within each organisation. (Olkkonen 1993) Both Kasanen et al. (1993, p. 256) and Kallio (2006, p. 516) describe the nature of decision oriented research as normative and the results as pragmative, meant to help management in running the firm (Kasanen et al. 1993; Kallio 2006).

In a decision oriented approach, data is usually a compilation of previous knowledge on dependencies. The scientific contribution is assessed through the utility and creativity of the solution, in a form of such as solution of a previously unsolved problem, or a faster or a more economic solution. Typically, the developed solution is tested in one or a few practical cases. (Kasanen et al. (1993, p. 256)

Kallio (2006, p. 533) mentions problem-oriented, instrumentalism, constructionism and practicality as the key concepts of both decision oriented and constructive research approach. In common with a decision oriented research, a constructive research seeks to produce a solution to a particular pragmatic problem (Kasanen et al. 1993, p. 244). However, an essential feature of a constructivist research process is showing the evidence in a practical application and testing the functionality of the application in practice (Kasanen et al. 1993; Olkkonen 1993; Kaikkonen 1996). In assessing the results of normative research, such as the decision oriented research, the role of an empirical part is to serve as an example (Neilimo & Näsi 1980). The evidence is usually shown by demonstrating that the developed solution will give better results than the earlier solutions have done (Olkkonen 1993, p. 55).

Olkkonen (1993, p. 71) has set out the principles and progress of a decision oriented research process and this is presented in Figure 3 below.

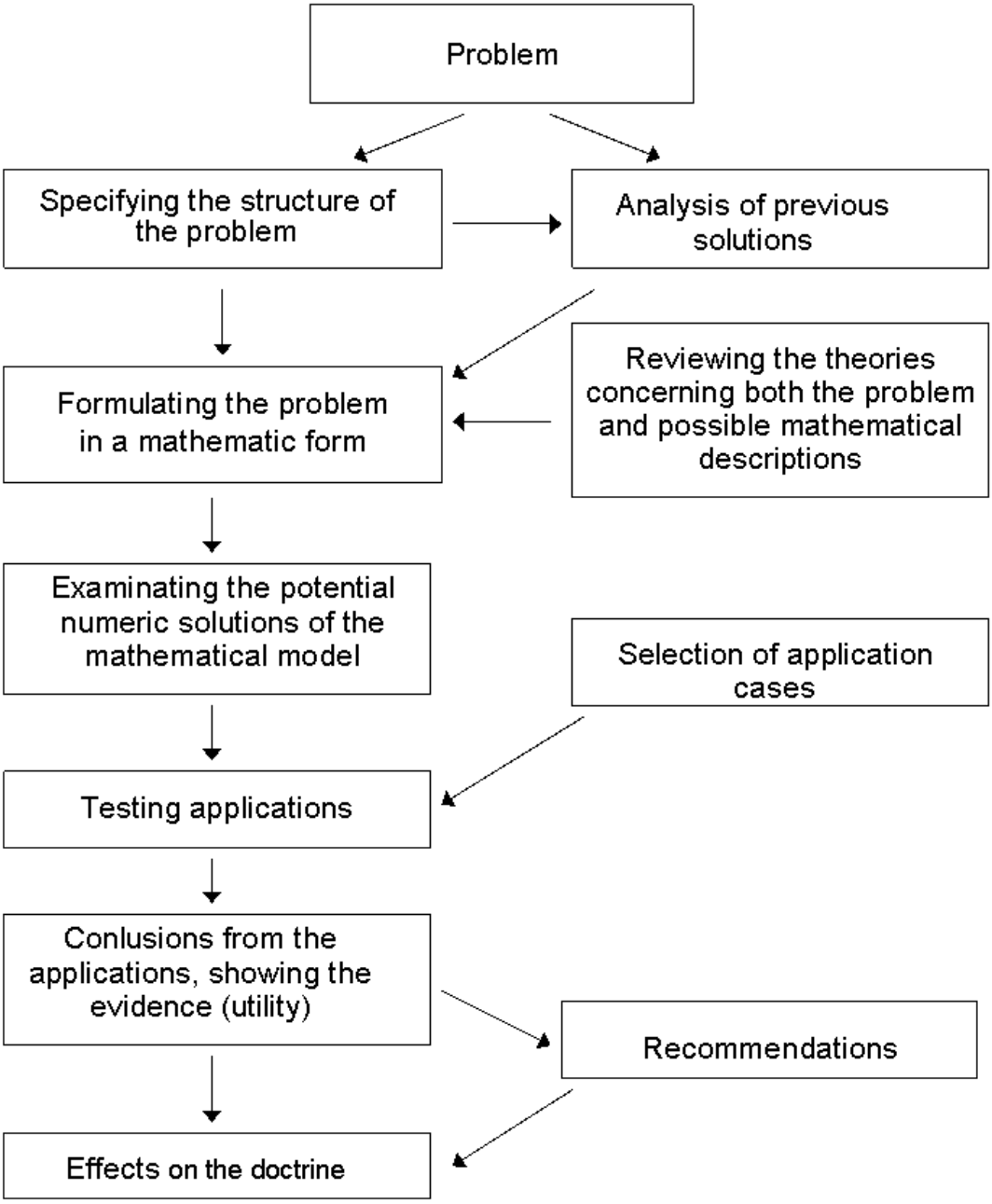


Figure 3. Decision oriented research process (adapted from Olkkonen 1993, p. 71).

This study follows the principles of a decision making research process. The starting point of the present study was pragmatic[8]: the idea for this study originated during earlier research from an insight that conducting student evaluations online provides an unutilised opportunity to compare the evaluations given by students directly with their personal course grades and

[8] According to Wicks & Freeman (1998) pragmatism allows researchers to "develop research that is focused on serving human purposes – i.e. both morally rich and useful to organizations and the communities in which they operate".

other personal background information, instead of the traditional way of comparing evaluation results with class-average values and expected grades.

Utilising earlier knowledge and experience in the area of student evaluations and a thorough exploration of previous solutions documented in the research literature, the main focus of the research was directed at the following question: does transferring student ratings system online and connecting it with other data systems provide an opportunity to yield more valid and more reliable ratings results? To be able to construct a model for answering this question and to determine which issues to analyse and how, it was necessary to be fully acquainted with existing student evaluation systems and the recognised and documented problems and pitfalls in using them. Even though the actual technical implementation of the interlinked survey was simple, it can – and as this study attempts to show, is of use to – be implemented in practically any university, it required a detailed knowledge in the field of student evaluations, to be able to design the interlinked system, to decide which issues to examine and to analyse and present the results and conclusions in this study in such a way that the possibilities of using systems of this kind to assess the validity and reliability of student evaluations are best elicited.

In the present study the gathered results of the developed interlinked evaluation system are presented and analysed to demonstrate the various aspects which this new evaluation system makes visible and which have not been possible to analyse with previous evaluation methods. The gathered survey data which comprises the empirical part of this study consists of both qualitative and quantitative elements. The numerical results have been obtained using statistical procedures typically used in quantitative research, though analysing and interpreting the results also required a good understanding of the M.Sc. study process. The observation of Eskola & Suoranta (2000, p. 20) that data gathered in research enhances rather than dampens the research thinking process, depicts the analysis process of this research work well. The results gained from the constructed solution are intentionally presented as straightforward as possible, keeping the use of more complicated statistical methods to a minimum in order to demonstrate the full potential of the interlinked system. Furthermore, interpretation of the results avoids extensive sociological issues since the focus of this study is on showing the utility of the solution, rather than on assessing the pedagogic aspects received from the results.

The utility and theoretical novelty of the solution developed in this study are both discussed in Chapter 6, *Results of the Empirical Study* and in Chapter 7, *Conclusions and Discussion.* Here, the most important findings are presented and analysed linking the results to earlier research findings in the area. Chapter 6 also provides examples of issues which are, or can be, received with this kind of model, which with earlier evaluation tools have remained unattainable. The purpose is to demonstrate the theoretical novelty of the developed

construction and to describe its potential areas of use. Figure 4 illustrates the structure of the research process.

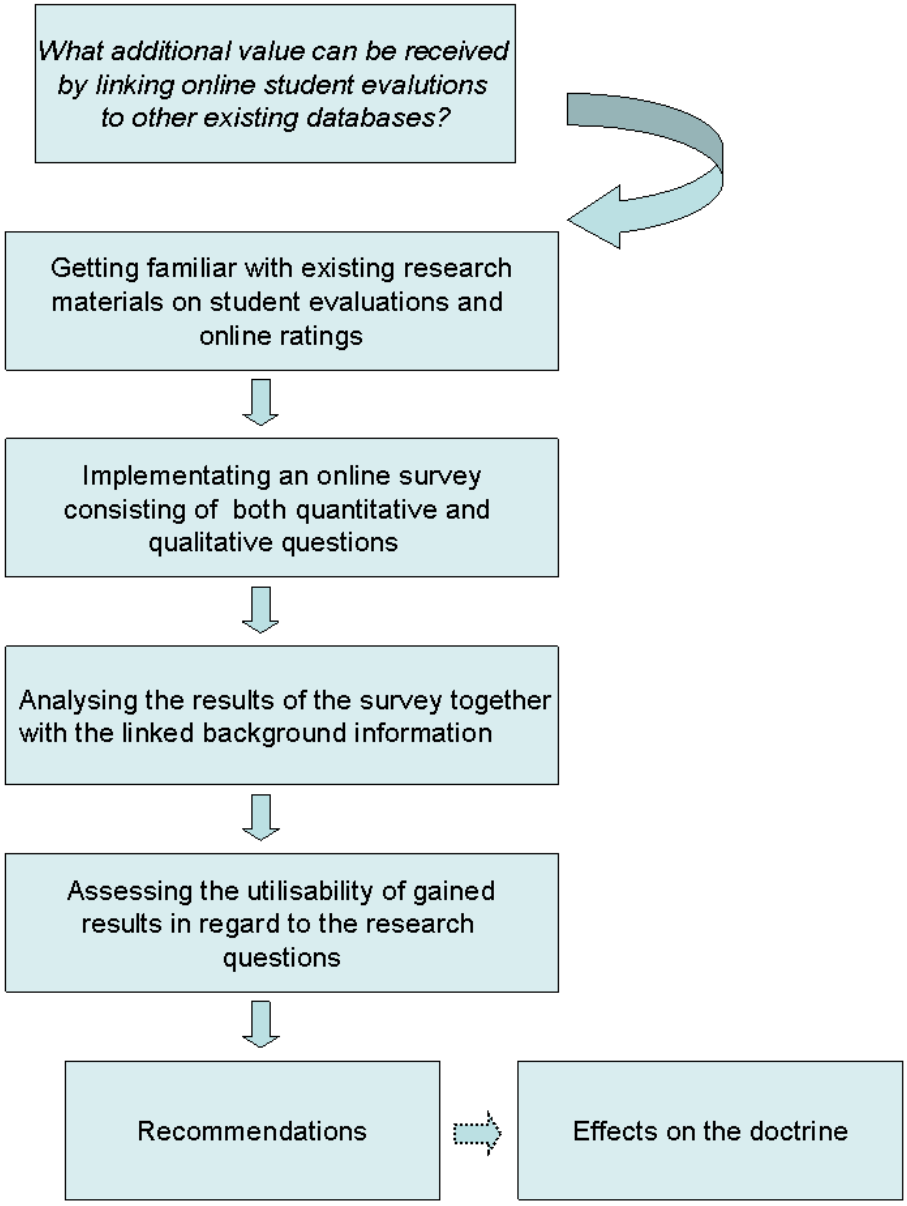


Figure 4. Structure of the research process of this study.

In the present study, the purpose is to confirm the utility of the connection between online ratings and university's other databases in assessing the validity and reliability of ratings results. Thus the main interest is not in proving the existence of any particular potential biases in students' behaviour but to demonstrate the usability of an interlinked online evaluation system in assessing the existence of potential biases.

1.4 Structure of the Study

In order to fully analyse the potential of online ratings it is necessary to be familiar with the existing knowledge of student evaluations in general and with the characteristics of implemented online evaluation systems. Therefore, this study begins with an introduction to the role of student evaluations in higher education in general and to the key concepts being used in the field. The study then examines the validity and reliability issues connected with student evaluations as well as other possible problems concerning evaluations and evaluation results. Most of the problems concerning the validity and reliability issues are not dependent on the instrument by which the survey responses are collected – paper or online – but on the respondents' ability and willingness to respond honestly and objectively to the presented questions,. As a result, this part of the study is largely based on research results from data collected from paper-based surveys. The reason for this is clearly because most existing in-depth studies in the field still rely on paper-based surveys, although the amount of published online-based research studies is steadily increasing.

Next the study presents the current uses of online ratings, the perceived advantages and limitations in the use of implemented online systems and observed gaps in the current research. This is followed by an empirical part which begins with a description of the development of the experimental online system. After this follows the results part where the data collected by the system are analysed to provide examples of ways how interlinked evaluation and background data results can be utilised. These examples are also compared when possible with previous research findings. Finally, the conclusions are presented.

This study is divided into seven chapters. This first chapter gives an overview of the research area, introducing the trajectory of evaluation practices within universities during the last century, the field of various feedback systems in higher education and the significance of student evaluations within universities. The first chapter also presents the research question, the research approach and the methodology of the research. The key concepts and their use are introduced in the second chapter. The third chapter describes the theoretical knowledge on student evaluations starting from the history of student ratings. It deals with the various purposes of collecting student information, the traditional assessment methods and the possible problems in the use of student ratings. The fourth chapter describes the main focus of this study, the online student ratings systems, the present knowledge on the use and research of online student evaluation systems within universities and the advantages and limitations associated with online evaluations in higher education. The fourth chapter also identifies the gaps in knowledge on the potential for using online evaluations in a broader context. The empirical part of this study is introduced in the fifth chapter, starting with a description of the development and implementation of a networked student evaluation survey

connected to university-held database. In the sixth chapter the key findings of the study are introduced and analysed. The limitations of the survey results and the generalisability of the results are also discussed. In Chapter 7, the final chapter, the conclusions of the study are presented along with an assessment of the value of the results obtained and suggestions for future research.

2 ANALYSIS OF CONCEPTS

> *When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative.*
>
> – Bob Stake –

Despite the fact that in recent decades education evaluation has received the utmost attention of many researchers there still exists no clearly established terminology in the field. Even many of the most basic terms can be ambiguous. For example, the term *assessment* is sometimes used as a synonym for evaluation (Braskamp & Ory 1994, p. 13), sometimes it is understood as feedback gained from students and, perhaps, most often, it is interpreted as the process by which students are graded, passed or failed. This chapter introduces the prevailing interpretations of the most typical terms used in the context of traditional and online student ratings and, more broadly, in the field of educational evaluation.

2.1 Evaluation

Definitions and the use for the term *evaluation* vary widely even when the subject is narrowed to concern only the field of higher education[9]. What makes the interpretation of the term especially demanding is that it can be understood in both broad and narrow terms, depending on the context. In its widest sense, evaluation in higher education has sometimes been described as an umbrella term covering a wide array of methods and techniques to assess extensive projects, programs and services (Soininen 1997, pp. 29–31). In practice, however, the term the evaluation has also been understood much more concisely, often in connection with teaching and especially with the validity and reliability of student ratings (Knapper 2001, p. 8).

[9] The study of evaluation developed especially during the 1950s and 1960s (Henkel 1998, p. 285) and has been accepted in the U.S. as an autonomous discipline (Scriven 1991, p. 141). However, over time, the development of evaluation and evaluation studies have become more complicated (Henkel 1998). Even the purpose of evaluation has received no uniform agreement amongst scholars. Whether it is possible to attain some kind of "objective truth" or whether the existence of an objective reality is ontologically completely denied as Guba & Lincoln (1989) suggest, affect the nature of the outcomes of the evaluation.

In general, evaluation can be defined as "both judgment on the worth of a program, procedure, or individual and the process whereby that judgment is made" (Dressel 1976, p. 1). This approach has also evoked criticism, for example Scriven (2001) has argued that "evaluation does not always involve judgment; it may only involve measurement against established standards". Instead, Scriven (1991; 2003) defines the term evaluation as the process of determining the merit, worth or value of something or the product of that process. According to Scriven (1991, p. 143) what "distinguishes evaluation from other applied research is at most that it leads to evaluative conclusions, and to get them requires identifying standards and performance data and the integration of the two". According to Raivola (2000, p. 66) evaluation typically contains a comparison between expressed goals and attained results, between the observed program and other programs, between earlier and new achievements or between alternative procedures. Stufflebeam (2003, p. 31) describes evaluation in a similar way to Scriven as "assessing something's merit and worth", but unlike Scriven, who emphasizes the need for coming to a certain conclusion, Stufflebeam emphasizes the improvement aspect by stating that "the most important purpose of evaluation is not to prove, but to improve".

Chelimsky (1997, p. 10) distinguishes the different purposes of evaluation into three general perspectives: evaluation for accountability (for example the measurement of results or efficiency), evaluation for development (e.g., the provision of evaluative help to strengthen institutions) and evaluation for knowledge (e.g., the acquisition of a more profound understanding in some specific are or field). According to the author, the major purpose of the knowledge perspective is "to increase understanding about the factors underlying public problems, about the "fit" between these factors and the policy or program solutions proposed, and about the theory and logic (or their lack) that lie behind an implemented intervention". All these elements are also relevant when the subject is analysed from an educational point of view.

In the context of higher education, Vuorenmaa (2001, p. 220) defines the purpose of evaluation as to "collect information needed in determining the value and usefulness of education and an educational reform and to use this information in the decision-making, monitoring, development and guidance of education". Hendry et al. (2001, p. 327) have described evaluation primarily as a means of achieving quality improvement in higher education. Cannon (2001b) concentrates on the learning aspect when he describes the process of evaluating as gathering information about learning and making judgments based on that information, emphasizing the importance of the latter element. Henkel (1998) defines the functions of evaluation in higher education as being the appraisal of new knowledge, certification of students, ranking of students and academics (e.g., candidates for posts or promotion and the allocation of rewards), the maintenance of common standards within a higher education system and scholastic improvement. In addition to the above publications,

there are also various other kinds of evaluation presented in the higher education literature, such as empowerment evaluation presented by Fetterman (1997).

For Stake & Cisneros-Cohernour (2000), evaluation is "usually thought of as based on criteria, with certain descriptive scores indicating more or less of something and drawing forth a judgment of goodness or badness". According to the authors, whether the judgment precedes or follows the explication of the criterion is not always clear, but criteria can usually be found when everything is evaluated. Beaty (2001, p. 84) has identified the same target by stating that evaluation "requires a set of criteria by which we judge whether or not something is good or effective". However, in clear contrast to Stake and Cisneros-Cohernour, Beaty depicts the magnitude which the absence of criteria has induced in university settings. He notes that "over the decades during which the evaluation of teaching has remained a contentious issue in higher education, the absence of such criteria has been a roadblock in our thinking".

Vuorenmaa (2001, p. 220) claims that although evaluation [in higher education] was originally understood as a means of assessing the results and effects of a programme, in the present context it is seen as a broader concept, addressing all the phases of educational decision-making. She suggests that the research viewpoint expands its scope to even further and regards evaluation as an action of a decentralised administration of education. Kellaghan et al. (2003, p. 3) have identified three distinctive features which separate educational evaluation from other types of evaluation: "First, it has been strongly shaped by its roots in testing and student assessment, on one hand, and curriculum and program evaluation on the other. … Second, education is the predominant social service in most societies. Unlike business and industry, or other social services such as health and welfare, education affects, or aspires to affect, almost every member of society. Thus, public involvement and the concerns of evaluation audiences and stakeholders are of special significance in educational evaluation, compared to evaluation in other social services, and even more so compared to evaluation in business and industry. Third, teachers play very important roles in educational evaluation as evaluators, as evaluation objects, and as stakeholders."

In higher education, a common application of the evaluation process is institutional evaluation. This, again, is a use of the term that can be understood in many different ways. According to Vartiainen (2005, p. 374), in Finland, institutional evaluations have been carried out by analysing management, the decision-making process, quality and the university's opportunities for conducting development work. Implementation of the Finnish institutional evaluation processes consists of three parts: institutional self-evaluation, external peer review and published report. This is similar to the process employed in other countries such as England. Unique to Finnish institutional evaluation processes, however, is that universities in Finland evaluate their functions in the basis of their *own* needs and interests, which means that the focus of institutional evaluation is different for each university.

(Vartiainen 2005) Leathwood & Phillips (2000, p. 318) emphasise this issue of ownership of evaluation by stating that for the success of an evaluation strategy, different constituents, especially teaching personnel and course teams, need to have a genuine stake in the process.

One of the well-known dichotomies in evaluation is its separation into formative and summative evaluations. Scriven (1991; 1996) defines formative evaluation as evaluation conducted *during* the development or improvement of a program for the in-house staff with the intent to improve. Summative evaluation, instead, is evaluation of a program conducted *after* the completion or stabilisation of the program for the benefit of some external audience or decision-maker. Summative evaluation thus serves as an aid for both decision-support and knowledge (i.e. research or inquiry) support. (ibid.)

Manwaring & Calverley (1998, p. 9) approach the distinction between formative and summative evaluation by stating that formative evaluation is "an ongoing process used to gauge overall progress and areas needing some attention or change, helping to mould the final article". They claim that within formative evaluation, information can be transferred back into the original work to both strengthen and move it forward, whereas summative evaluation "provides a fixed point or reference, and it may provide a measure of success or otherwise against original objectives or planned outcomes". Patton (1997, p. 69) on the other hand distinguishes between formative and summative evaluation by stating that formative evaluation typically connotes collecting data for a specific period of time, usually during the start-up or pilot phase of a project, to improve implementation, solve unanticipated problems, and make sure that participants are progressing toward desired outcomes. However, according to Patton (1997), improvement-oriented evaluation more generally includes using information systems to monitor program efforts and outcomes regularly over time to provide feedback for fine-tuning a well-established program. The problematics concerning the acquisition and use of summative and formative evaluation results within higher education are discussed in greater detail in Chapter 3.1, *Different Purposes of Obtaining and Using Feedback*.

In the present study, the focus of evaluation is concentrated on student evaluation of teaching. According to Yao (2001), this is defined as "any procedure by which students evaluate their instructor's teaching performance. Usually it takes the form of ratings by students in the middle or end of a term". Marsh (1987, p. 259) has described the purposes of student evaluation of teaching effectiveness as follows:

- diagnostic feedback to faculty about the effectiveness of their teaching that will be useful for the improvement of teaching.

- a measure of teaching effectiveness to be used in administrative decision making

- information for students to use in the selection of courses and instructors

- a measure of the quality of the course, to be used in course improvement and curriculum development

- an outcome or a process description for research on teaching.

Even though the previous purposes are uniformly accepted by most scholars, there are huge differences in opinion regarding the suitability of student evaluations for the aforementioned purposes (e.g., Johnson & Ryan 2000). Such differenced are particularly marked with regard to the validity and reliability of student evaluations (for more information see e.g., Greenwald 1997; Wachtel 1998; Kulik 2001). These are discussed in greater detail later in this study.

2.2 Assessment

The term a*ssessment*, in the context of higher education, although sometimes used interchangeably or overlapping with the term evaluation (Cannon 2001b), is most often referred to as the means used to measure the outcomes of education. Dietel et al. (1991) define assessment as "any method used to better understand the current knowledge that a student possesses". Johnston (2004) distinguishes the cultural differences in terminology by noting that in the American literature the terms 'assessment' and 'evaluation' are used interchangeably whereas in the UK 'assessment' refers to assessment of students and 'evaluation' to evaluation of institutes and programmes.[10]

The Quality Assurance Agency for Higher Education in UK, QAA (2000) defines assessment as "a generic term for a set of procedures that measure the outcomes of students' learning, in terms of knowledge acquired, understanding developed and skills gained". According to Yorke (1998) there are three main purposes of assessment: assisting in the process of learning; determining what learning has occurred; and providing evidence regarding the success of the program in question. Pellegrino et al. (2001, p. 2) presented a parallel definition for assessment stressing that "every assessment, regardless of its purpose, rests on three pillars: a model of how students represent knowledge and develop competence in the subject domain, tasks or situations that allow one to observe students' performance and an interpretation method for drawing inferences from the performance evidence thus obtained". Kerka & Wonacott (2000) in considering assessment results, note that assessment is usually intended to provide both instructors and students with information on progress and to measure the achievement of learning goals. Mutch (2002, p. 165) has a broader view, which includes the often-competing purposes of assessment: to confer an award, to rank students and to aid learning. Fenwick (2001) classes assessment into two levels: micro-level indicators

[10] The term "programme evaluation" can also be understood in several different ways. Konttinen (1996, p. 13) notes that evaluating operations is often called "programme evaluation" to distinguish it from continuous follow-up such as e.g., monitoring.

or assessments conducted in class and macro-level indicators or assessment conducted across programs and institutions.

There are not only cultural, but also historical and contextual differences in the use of the term assessment. Braskamp & Ory (1994, p. 12–13) explain the word *assess* in terms of its Latin root *assidere*, meaning to "sit beside". According to Ory (2000) "sitting beside" implies dialogue and discourse, with one person trying to understand the other person's perspective before making value judgments. Scriven (1991) has noted that for many years in the twentieth century the term "evaluation" referred only to student evaluation and in particular only to the evaluation of student work done in tests. Thus, the interpretation for either the word "evaluation" or "assessment" is not self-evident but situationally dependent. While most often associated with the outcomes of individual students, there are also wider interpretations of assessment. For example, Harvey & Askling (2003) argue that assessment may also benchmark against other institutions, national norms or against oneself over time. Similarly Mutch (2002, p. 170) observes that "at a programme level, the key strategic question is the way in which assessment tests the learning outcomes for the programme as a whole". Harman (1998, p. 334) notes that the approaches adopted for assessments to vary considerably, the most important methodologies being:

- self study or self-evaluation
- peer review, usually including the use of at least some external visitors
- analysis of statistical information and/or use of performance indicators
- surveys of students, graduates, employers, professional bodies, and
- testing of students' knowledge, skills and competencies.

In the field of higher education, Braskamp & Ory (1994, p. 19) see two main uses of assessment: individual and institutional. In the individual use, the primary function is career development and the primary types of information communicated are documentation of strengths and suggestions for improvement. In the institutional use, the primary function is institutional accountability and the primary types of information communicated are judgments of merit and worth to institution and society. (ibid.)

Usually the term assessment is quite narrowly used to describe student-outcome assessment but recently its meaning has expanded somewhat among academics. For example, Kurz & Banta (2004, p. 85) widen the concept of assessment to include both the outcome of the process and its starting point by "providing an impetus for a faculty member to define a bottleneck to students' learning and devise ways of helping students move past that bottleneck". Brakke & Brown (2002) define assessment as a process of ongoing formative

and summative evaluation, with a workable structure of an assessment program consisting of the following components:

- Moves the focus from teaching to learning
- Considers goals and measurable outcomes for general education and major programs
- Is based on findings of cognitive research
- Reflects clear institutional purpose and vision for skills and abilities expected of all students
- Gives attention to the cognitive development of the student
- Gives attention to the attitudes and preconceptions of students
- Effectively demonstrates the impact of the teaching practice and other experiences on the learning of students.

It has been claimed that as the assessment movement has matured over time, the prevailing emphasis has changed. There has been a shift from a limited, short-term focus on measurement of student outcomes with standardised, norm-referenced tests for the purpose of accountability. The most recent approaches in assessment are a much broader, with a long-term focus on a range of techniques to gather information for instructional, programmatic, and institutional change. (Gray 1991, p. 62) Ory takes a holistic view of the subject in his article (2001), arguing that "assessment is more than counting, measuring, recording, or accounting. It promotes teaching evaluation not as a scientific endeavour, with absolute truth as its goal, but rather as a form of argument where the faculty use their data to makes a case for their teaching".

Hubball et al. (2004) describe assessment as the "systematic gathering of information about component parts of the thing being evaluated". They stress the importance of authentic assessment, which they claim is designed to address learning that is meaningful to the learner and also the skills and abilities needed to perform real world tasks. Hubball et al. prefer authentic assessment over traditional assessment, criticising the latter for measuring learning in ways that are important to or convenient for faculty and institutions but often irrelevant to the real-world application of learning. Rust (2001) notes the crucial role of assessment in the education process. Assessment, he suggests, determines much of the work students undertake, affects their approach to learning and is an indication of which aspects of the course are valued more highly. According to Rust, assessment should be an intrinsic part of the learning process and thus should be seen as a vital part of the initial design of the course or module.

One of the latest research topics in the area of assessment has been computer assisted assessment (CAA), which is usually understood in broad terms to include a range of activities

such as, the collation, analysis and transmission of examination grades across networks, as well as the use of computer based assessment, where students complete assessments at workstations and their answers are automatically marked (e.g., Brown et al. 1997; Stephens et al. 1998; Bull 1999; Bull & Danson 2004). Another, possibly even more challenging method, made possible by advances in modern technology is reasoning from complex assessments (e.g., Mislevy et al. 2002). This construct-centred approach uses a student model which takes the form of a fragment of a Bayesian net to manage knowledge and uncertainty in assessment. Xenos (2004) has also used Bayesian networks for modelling the behaviour of the students in a computer course that deploys distance educational methods.

2.3 Performance Measurement

A third term frequently used in the context of evaluating higher education is *performance measurement*. Though mainly used in the management accounting research (for more details, see, e.g., Lönnqvist 2004), the term is also associated with accountability in higher education, where performance measurement is often seen as a way for of public management to compare and rank institutional performance (e.g., Dill 1998; Alexander 2000). Ranki (1999) states the function of measuring as to make things visible and gain information which can otherwise be difficult to acquire, such as competence and the level of knowledge. According to Ranki, measurement is used to gain factual information on the state of the knowledge. Ranki also notes that measurement can be used to show the importance of certain matters and to direct investments and learning. Toegel & Conger (2003) argue that the popularity of performance measurement is because it informs both employees and organisations of their effectiveness in getting results and achieving goals.

According to Wholey (1996), for many policymakers and managers, performance measurement is evaluation, although it does not typically provide information on the net impacts of policies or programs (for details of tensions in performance measurement see, e.g., Ryan 2002). Wholey describes performance measurement as a bridge to more sophisticated formative and summative evaluating studies. This is made possible because performance measurement can facilitate consensus of valid performance indicators and provide time series data on program outcomes. It can also identify appropriate opportunities to use qualitative evaluation to explore factors contributing to performance variations over time or performance variations among subordinate units. (Wholey 1996) Table 1 below presents some evaluative approaches which Cannon (2001a) considers to act as performance indicators:

Table 1. Typical performance indicators in higher education (Cannon 2001a, p. 89).

Focus of Evaluation	*Indicators of Teaching Performance*
Individual instructor or teacher	Student evaluation
Teaching teams	Peer evaluation
Course, unit, or program of study	Course Experience Questionnaire
Academic department	Portfolios Audits Benchmarking
Institution	Portfolios Benchmarking Ranking

Cheng (2001, p. 526) mentions student test scores on aptitude, grade point averages and retention/persistence/graduation rates as among the most popular performance indicators adopted. However, he criticises these indicators for not measuring certain aspects of an institution's effectiveness. According to the author they are unable to provide meaningful information on students' intellectual and personal development as the outcomes of their collegiate experience. To overcome this problem several research groups have adopted survey instruments which include items of students' self-perceived intellectual, social and personal gains. (Cheng 2001)

Several research articles emphasize the importance of student evaluations, the focus of the present thesis, as a performance indicator. For example Parjanen (2003, p. 14) has even argued that actually the student ratings are the only actual performance indicators that have generally been proven to be valid. According to Ballantyne et al. (2000, p. 221) the use of student evaluations as a performance indicator of teaching effectiveness in higher education has increased in many countries. Porter & Whitcomb (2003, p. 389) have also noted that student survey data have grown increasingly important for institutions of higher education, both in terms of internal assessment and their use in external performance indicators.

Typical users of performance indicators in higher education are governments, who use them to inform policy and to allocate resources and the media, who use them to construct rankings of universities (Cannon 2001a; Leathwood & Phillips 2000). Ramsden (1991) actually claims

that the use of performance indicators is a direct result of efforts by national governments to increase universities' and colleges' accountability to their paymasters. However, the sole reliance on performance indicators is also questioned by other authors such as Hoey & Gardner (1999, p. 43) who warn against solutions that are overly simplified, noting that "what is easy to measure may not be what is most meaningful in measuring program performance". Similarly, Cheng (2001) states that "being able to form outcome measures does not necessarily mean that an institution has found the answer to the critical questions of what is excellence in higher education and how it can be attained and assessed".

In higher education, performance indicators entail the collection of data at different levels of aggregation to aid managerial judgments, which may be made either within institutions or at the level of the higher education system as a whole (Ramsden 1991). Alexander (2000, p. 426) has claimed that performance funding and budgeting policies have increased the tension between policymakers and higher education administrators and faculty because of divergent objectives: government authorities prefer to use indicators that measure institutional efficiency, consumer satisfaction, job placement, and value for resources. According to the author, such policies also favour the use of performance measurements as a means of comparing institutional productivity and performance. On the other hand, university administrators and faculty favour performance measures that reflect the quality of the educational experience in a manner that promotes their own institutional mission(s). University leaders also advocate the adoption of performance measures, but only when the results are to be used in a non-competitive way to improve individual performance. (Alexander 2000)

Ramsden (1991) says that performance indicators are not about rating individual staff, but rather about the performance of the units in which they work and so the level of analysis should be defined as the collection of information about programmes rather than individual teachers' performance. Braskamp (2000, pp. 25–26) also argues that if we focus too much on using student ratings of teaching to measure the quality of teaching, the overall quality of teaching on a campus will arguably be less effective.

2.4 Feedback

Unlike many other terms used in the discipline, the term *feedback* in higher education is largely understood as collecting information from students about their experience of higher education (e.g., Braskamp & Ory 1994; Harvey 2001; Brennan et al. 2003). Gathering feedback is one of the commonest ways of acquiring information in higher education institutes. Harvey (2001) divides the purposes for collecting feedback into two categories: gaining internal information to guide improvement and external information for potential

students and other stakeholders. Brennan et al. (2003, p. 45) define student feedback as obtaining information about (1) student satisfaction with specific programme/unit or services, (2) student views on whether their objectives have been met and (3) student accounts of their learning and study methods. However, as Brennan & Williams (2004) note, whether students' personal objectives have been met is not necessarily the same as feedback on satisfaction with the teaching and learning processes of specific programmes, modules or services.

Menges (1991, p. 29) has listed the feedback sources in terms of the following categories:

- Classroom information (teacher perceptions during class, audio- and videotapes)
- Student opinion (student surveys, student informal comments, student interviews)
- Alumni opinion (informal alumni comments, alumni surveys)
- Student learning (student course examinations, student papers and other graded work, student standardised test scores)
- Collegial information (supervisor visit, teaching consultant visit, colleague assessment of teaching materials, colleague informal comments).

According to Geis (1991) the information is typically gathered as a part of an evaluation process and the feedback[11] can be of various sorts: it can be a pure reflection on an event such as a teaching episode or it can be information in some translated form such as student ratings of a class. He states that feedback can refer to the performance itself or to its outcome or impact, i.e., the teaching process as such or examination scores or graduate placements. Braskamp & Ory (1994, p. 121) highlight the adjustive nature of feedback, claiming that feedback prompts the receivers to make adjustment and alter their behaviours; it is meant to trigger action either on work activities (e.g., lecturing or conducting surveys) or outcomes (e.g., students learning a skill).

As noted earlier, one of the contentious issues concerning student feedback is the validity and reliability of students' responses. Since this issues forms the main focus of this thesis, it will be discussed more extensively in Chapter 3 and also examined in the empirical part of this study.

[11] According to Geis (1991) the word "feedback" is originally associated with the field of cybernetics in N. Wiener's book Cybernetics: Control and Communication in the Animal and the Machine, MIT Press, 1948, but the concept is largely derived from physiology, where it denotes self-regulating systems of a living body. It refers to a process by which information about the effect of a system is fed back into that system, providing potential for adjustment.

2.5 Rating

There are several different definitions of the term *rating*. Webster's Encyclopedic Unabridged Dictionary (1994) gives seven different definitions of the term, the first which being "classification according to grade or rank".

Scriven (1991) defines rating as "usually same as grading" in his book Evaluation Thesaurus, whilst in a higher education evaluation setting, the term is more often associated with the various rankings given *by* students than the grades given *to* students. Cashin (1995) has described the difference between evaluation and rating as follows: "evaluation" has a definitive and terminal connotation; it suggests that we have an answer; "rating" implies that we have data which need to be interpreted. He adds that using the term "rating" rather than "evaluation" helps to distinguish between the people who provide the information (sources of data) and the people who interpret it in combination with other sources of data (evaluators). In the context of higher education, "rating" can also relate to "ranking" which refers to a classification of the order of superiority of universities according to a certain criteria. There are numerous journal and newspaper articles such as "Oxbridge closes on Harvard in rankings" (Ince 2006) as well as Internet sites listing the "top 100 universities in US" or "World top 500 universities".

In the research literature of higher education, the term *student ratings* is often used as a synonym for instructor and instruction evaluations (see e.g., Aleamoni 1999), whereas some scholars consider ratings by students as types of evaluation within an all-round assessment perspective (e.g., Husbands 1998). There may also be cultural and geographical differences in the use of the term "student ratings". For example Berk (2005), quoting previous research findings, says that in the United States student ratings have become synonymous with *faculty* evaluation.

In the field of higher education research, surveys are often used as a synonym for ratings, both referring to a way of gathering information from students. There may be seen a subtle tendency towards calling traditional paper-based student evaluations as ratings and corresponding Internet-based evaluation as Web surveys. However, there is so far no well-established uniform terminology in the field. Couper (2002) points out that though the terms survey and questionnaire are used interchangeably today, before the growth of the Internet, surveys were considered to be administered face-to-face interviews whereas questionnaires were self-administered.

In this study, the terms student ratings, students' evaluations, student evaluations and student surveys are all used interchangeably to mean questionnaires aimed at students in order to gather opinions regarding their studies for many kinds of development use, not solely or

specifically faculty evaluation. The purposes as well as possibilities and limitations of student ratings are discussed more thoroughly later in this study.

2.6 Bias

The following are some of the many definitions of the term *bias*: "a personal and sometimes unreasoned judgment"; prejudice**:** "an instance of such prejudice", "deviation of the expected value of a statistical estimate from the quantity it estimates" and "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others" (Merriam-Webster Online 2007). According to Scriven (1991, p. 67) in the evaluative sense, bias means much the same as prejudice, and as its antonyms he gives the terms "objectivity", "fairness" and impartiality". In educational settings, all the above interpretations are strongly represented whenever when the existences of potential biases are discussed.

The concern over the effect of bias is the possibility that background characteristics or other factors that are not related to the effective teaching itself can affect the results of student ratings (Worthington 2002, p. 49). The importance of this topic within higher education is well illustrated by the fact that in 1987 an entire issue of *International Journal of Educational Research* was devoted to Marsh's monograph "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research". This made six claims, one which was that student ratings are "relatively uncontaminated by many variables often seen as sources of potential bias".

In the article Marsh (1987, p. 310) also argues that "an important problem in research that examines the effect of potential biases to students' evaluations is that adequate definitions of bias have not been formulated. The existence of a significant correlation between students' evaluations and some background characteristics should not be interpreted as support for a bias hypothesis". Another article involved in the debate on the existence of bias is Aleamoni's (2004) "Student Rating Myths Versus Research Facts from 1924 to 1998", where the author refutes sixteen of the most common claims regarding student ratings (e.g., "The size of class affects Student Ratings" or "The gender of the student and the gender of the instructor affect student ratings").

Several years, and many other research articles later, there is still no consensus to confirm or reject the validity of student ratings with regard to the potential biases (e.g., Nasser & Fresko 2006).

2.7 Grading Leniency

Perhaps the most commonly expressed suspicion of bias in student ratings relates to the possibility that the students' course grades correlate positively with the student ratings of the same courses (Marsh 1987; Greenwald & Gillmore 1997a). In other words, students tend to give higher ratings when they expect higher grades in the course (Huemer 2007).

In addition to the numerous articles discussing bias in general, there are many research articles dealing specifically with grading leniency (e.g., Greenwald & Gillmore 1997a; Wachtel 1998; Olivares 2001; Griffin 2004). It has been widely believed that the positive correlation between the expected grade and student ratings of instruction is a reflection of valid measurement of student ratings since better instruction should result in more learning, better grades and better ratings (Griffin 2004. p. 411). However, as a counter argument the grading leniency hypothesis still persists, i.e. instructions are rewarded with higher ratings for assigning higher grades as a result of lenient grading practices or conversely penalized with lower ratings for assigning lower grades due to grading harshness (ibid.).

This ongoing unsolved debate over the potential bias in student ratings caused by grading leniency has been a key factor motivating the present study. A possible explanation for the chronic lack of consensus on the existence of grading leniency is that there have been no tools for linking students' course evaluations with their actual course grades. This study sets out to demonstrate new solutions.

2.8 Summary of Key Concepts

Table 2 draws together the key concepts presented in this chapter with their definitions as they are used in the following chapters. The relationships between the terms can be described as follows: Evaluation, in general is defined as a process of determining the merit or value of something. In higher education, evaluation covers almost all aspects of educational decision making. Thus, performance measurement within higher education can be interpreted as one form of evaluation: as a means to compare and rank institutional performance. Student evaluation is one way of gathering feedback from students and it is also a commonly used performance indicator within performance measurement. Student evaluations are usually carried out using various forms of survey which are also called as student ratings. Whether the student evaluations are biased, especially in terms of grading leniency, is an issue which has been widely researched but not yet satisfactory resolved.

Table 2. Summary of key concepts.

Concept	**Definition**
Evaluation	Understood in general as the process of determining the merit, worth or value of something. Within higher education it is often focused on an entire institution. On the other hand the term can also be used in a narrower context, such as student evaluation, which typically means gathered students' opinions concerning the teaching or the overall education they are given.
Assessment	A term commonly used to describe the measurement of student outcomes; however, it is sometimes also used as a synonym for the term evaluation.
Performance measurement	A management tool for determining performance in a certain measurement situation. Within higher education, usually defined as the means for comparing institutional productivity and performance.
Feedback	Information collected from students about their experience of certain elements of higher education.
Rating	Any classification aiming to grade or rank. In higher education the terms student ratings, student evaluations and surveys are often used interchangeably.
Bias	A tendency to favour a certain set of values. Within student evaluations, the term is used mean that student ratings are affected by factors unrelated to effective teaching or teaching performance.
Grading Leniency	Hypothesis claiming that students tend to give higher ratings when they expect higher grades in the course and vice versa.

As the table above demonstrates, all the above terms can be interpreted in various ways according to the context in which they are used. In this study the terms "evaluation" and "rating" refer mainly to the process, while the term "feedback" refers to the results of the information gathered from university students.

3 APPROACHES TO STUDENT EVALUATIONS IN HIGHER EDUCATION

If all the educational assessment devices were stacked on top of each other,
it would make quite a pile.
If we removed from that pile all of the paper-and-pencil tests,
then the remaining pile would be quite small.

– W. James Popham –

As noted in the first chapter, student evaluations are probably the most researched topic in the field of higher education. Despite the popularity of the topic, or perhaps because of it, there are still major differences of opinion among the scholars about even the most fundamental elements of the evaluations (for details see, e.g., Johnson & Ryan 2000, Emery et al. 2003). Even though the main focus of this work is online evaluations, the basic elements concerning the purpose and utilisation of evaluations are the same, regardless of how the information is collected. This chapter presents an overview of the scope and the role of different evaluation methods. It introduces the current state of research knowledge on student evaluations, especially the various purposes of obtaining feedback, the different ways of gathering information and the limitations of ratings. These are topics which have long been discussed by researchers and practitioners in the field.

3.1 Different Purposes of Obtaining and Using Feedback

Systematic feedback collecting from university students started in USA in the 1920s and the first research reports of student evaluations were published in 1927 by Brandenburg & Remmers (fore details, see, e.g., Marsh 1987). For the next five decades the use of student ratings increased slowly, for example, Wilson (1998) has estimated that only 30 percent of colleges and universities had set up student evaluations on lecturers in 1973. However, in recent decades student ratings have become commonplace in most universities throughout the world (see Centra 2000; Cranton 2001). As the use of ratings has become more general, so too has the significance of ratings results become increasingly important for institutions of higher education (Porter & Whitcomb 2003, p. 389).

There are several geographical and cultural differences in both the purposes for and uses of collecting student feedback in addition to research into student evaluations. Perhaps the most fundamental of them relates to whether the focus of evaluation is on teaching only or education as a whole. If the focus is solely on teaching or teaching effectiveness, then only the lecturers' personal teaching skills are being evaluated. However, if the focus is on education, other factors, such as the course material, contents of the course, schedule and even the students' personal conception of what has been learned are also being evaluated. Marsh (1987, p. 259) addresses this difference, stating that student ratings of teaching effectiveness are "primarily a function of the instructor who teaches the course rather than the course that is being taught". According to him, this also means that student ratings collected for the purposes of evaluating teaching effectiveness are not the appropriate source of student input into questions of course evaluations.

In U.S.A. most student evaluations are aimed at assessing individual lecturers' teaching effectiveness on individual courses (see, e.g., Thorpe 2001, Emery et. al. 2003) but worldwide there are also examples of larger scale evaluations. In Australia, for example, the focus is on assessing the whole curriculum and there are specific methods such as the CEQ[12] method developed and widely used for the purpose (Cannon 2001a). In Europe most countries perform student evaluations in order to improve single courses or modules (e.g., Cannon 2001a). Shevlin et al. (2000, p. 398) have described the differences in the use of student ratings by stating that, whereas in the USA the student ratings are collected largely for faculty decisions such as salary and promotion, in the UK, for example, the ratings results are used more often as a guide for potential chances, in the content of course material and the method of delivery (see also Cannon 2001a; Coffey & Gibbs 2001). In Finland feedback is often gathered from individual courses for improvement purposes (e.g., Lappalainen 1997, p. 15). The research literature in USA has concentrated on students' evaluations on their teachers whereas the predominant Australian and British research literature has targeted students' perceptions of the quality of their programme (Brennan et al. 2003, p. 9).

There are various reasons for collecting student feedback. As early as the 1960s Remmers et al. (1966, p. 9) stated that "we evaluate because we must always be concerned with whether we are reaching the goals of our teaching efforts. By analyzing the methods and results we hope to find ways of improving them. Evaluation is not an extra chore imposed upon instruction – a distasteful task to be completed as quickly as possible; it is an integral part of what a good teacher does to make his teaching more effective." As Brennan et al. (2003, p. 47) note, student feedback is collected at a number of levels and the use of student feedback at different institutional levels may reflect differences in purpose. Among the most commonly cited reasons for collecting student feedback are (1) student guidance in choices

[12] According to Ramsden (1991) the original principle in designing the Course Experience Questionnaire (CEQ) was to produce "quantitative data which permit ordinal ranking of units in different institutions, with comparable subject areas, in terms of perceived teaching quality".

of courses and instructors, (2) diagnostic feedback to faculty for the improvement of teaching, (3) evaluating teaching for use in personnel decisions; e.g., tenure or merit salary increases and (4) a measure of the quality of the course for use in course improvement and curriculum development (e.g., Marsh 1987; Braskamp & Ory 1994; Scriven 1995; McKeachie 1996; McKone 1999; Johnson & Ryan 2000; Kulik 2001; Kember et al. 2002; Spencer & Schmelkin 2002; Conn & Norris 2004; Turner 2004). Marsh (1987, p. 259) also adds "an outcome or a process description for research on teaching" as a fifth major reason for collecting feedback.

Brennan et al. (2003) list the following levels where feedback from students is collected: an individual teacher or class, a module or unit, a semester or year of study, a programme of study, a subject, a department and a faculty. In the USA most of the research reports have focused on the students' evaluations of teaching and teachers on single courses, whereas outside USA greater emphasis has been given to the development of modules or programmes. However, the focus of evaluation may be broadening. For example, Cannon (2001a) has argued that the focus of evaluation is already shifting from individual teachers towards the evaluation of teaching in its wider context.

Harvey (2001) classifies the various types of student feedback surveys being used in UK into following categories: institution-level satisfaction surveys, faculty-level surveys, programme level surveys and teacher appraisals by students. According to the author, the characteristics of the institution-level surveys are that they

- are almost always based on questionnaires mainly consisting of questions with pre-coded answers augmented by one or two open questions
- are undertaken by a dedicated unit with expertise in undertaking surveys producing results to schedule
- are not to be confused with standardised institutional forms seeking feedback at the programme or module level and
- seek to collect data that provide either management information designed to encourage action for improvement or a descriptive overview of student opinion which can be reported as part of an appropriate accountability process.

Faculty-level surveys, as described by Harvey, are based on pre-coded questionnaires and tend to focus on those aspects of the experience that the faculty controls or can directly influence. Programme-level surveys, on the other hand, are not always based on questionnaires (although Harvey notes that most tend to be). In some cases, feedback on programmes can also be solicited through qualitative discussion sessions. Harvey describes informal feedback on programmes as a continuous part of the dialogue between students and lecturers and as an important source of information at this level for improvement. Module-

level feedback, according to Harvey, provides an important element of continuous improvement. It tends to focus on the specific learning and teaching associated with the module, along with some indication of the problems of accessing module-specific learning resources. Yet, Harvey seems to take a rather critical view towards the use of module-level feedback in UK. He argues that module-level questionnaire feedback is usually superficial, results in little information on what would improve the learning situation, and because of questionnaire-processing delays, rarely benefits the students who provide the feedback. According to the author student appraisal of teacher performance also has a limited function, which is actually ritualistic rather than improvement-oriented. (Harvey 2001).

Despite of the widespread use of student ratings within universities, opinions concerning the use of ratings results among teachers are at least as contentious as they are among researchers of the field. Whereas some teachers view ratings as reliable and valid measures bringing scientific accuracy to the evaluation of teaching, others suspect that students use ratings as a means of exercising their power or fear that the ratings become a personality contest instead of an evaluation teaching effectiveness (Kulik 2001). In practice, however, the lack of consensus among teachers and researchers has not prevented the significance and magnitude of student evaluations from increasing steadily during the last few decades. On the contrary, universities are devoting considerable resources to obtaining feedback from their students (Brennan et al. 2003).

The research literature on educational evaluation adduces two dichotomies with regard to the purposes of student feedback: the accountability/developmental purposes of feedback and the summative/formative nature of evaluations. According to Johnson & Ryan (2000, p. 115), two pressing demands on teaching evaluations are the need for information to aid administrator decision making (accountability purposes) and the need for information to help instructors improve their teaching (developmental purposes). In the former, faculty focus on their strengths and accomplishments; in the latter, they focus on identifying and understanding the problem areas and discovering ways to improve in these areas (Johnson & Ryan 2000). Although frequently cited this dichotomy is also sometimes questioned. For example, MacLeod (2001) argues that the reality is that every university's student feedback procedure serves the dual functions of staff development and staff appraisal. What is problematic, however, as MacLeod observes, is that the function (i.e., development or appraisal) selected by the university as the purpose of student feedback will have a major impact on the characteristics of the system to be constructed. Thus, the effort to imbue the system with characteristics that enhance development will simultaneously compromise the capacity of the system to adequately serve appraisal, and vice versa (MacLeod 2001). Brennan et al. (2003) agree, arguing that there is no evidence that obtaining feedback at one level would be useful or effective in monitoring or improving quality at another level. Stake & Cisneros-Cohernour (2000, p. 52) express their concern by warning that "data used for one

purpose may unethically be used for another purpose, for example, data volunteered for staff improvement may wrongly be used for accountability."

When examining the purposes of student evaluations the differences between formative and summative evaluation are made explicit in the work of Theall & Franklin (2001a, p. 51). They observe that "when we gather information to review or explore or improve, we describe this as 'formative" evaluation'. When the purpose is to make decisions about merit, promotion or tenure, for example, we call it 'summative evaluation'." Spencer & Schmelkin (2002) present the three main purposes for student ratings in pragmatic terms: a formative one of giving student feedback to faculty members about their teaching and areas in need of improvement and two summative purposes: providing administrators with evaluative data used in personnel decisions and enabling students to reach out to their peers and formalise their opinion. This dichotomy can pose a major challenge for student evaluations: they are expected to function both as a formative tool (improvemental aspect) *and* as a summative tool (compensational aspect) simultaneously.

The suitability of student evaluations for both purposes is considered contradictory among the scholars. For example, according to Henkel (1998, p. 292) evaluation is an integral part of the dynamic of higher education, its regulation being both summative and decision-oriented and formative and development-oriented. Similarly Marsh (1987, pp. 259–260) wrote earlier that student ratings are "important both as a process-description measure and as a product measure and this dual role, as a process description and as a product of the process is also inherent in their use as a diagnostic feedback, as input for tenure promotion decisions and as information for students to use in course selection". However, some years later Marsh along with Bailey (Marsh & Bailey 1993, pp. 12–13) acknowledged that the appropriate use for students' evaluations for teaching effectiveness, particularly for use in the personnel decisions, is more controversial than their use as formative feedback. Abrami (1989) criticizes the use of ratings for promotion and tenure decisions and Braskamp & Ory (1994) doubt the suitability of assessment for both promoting faculty development and meeting the collective institutional goals. Theall & Franklin (2001a, p. 53) take an even stronger stand against conjoining summative and formative evaluation, arguing that it is crucial to separate the two purposes conceptually and in practice. They suggest using formative data for classroom assessment and research, and warn against making program decisions without agreement on the suitability of the data and how such data should be used. Lubinescu et al. (2001) provide one example of the usage by stating that formative evaluations are "most useful in providing directions for programmatic improvements, which faculty and administrators most highly value".

Cannon (2001a) has outlined a contextually integrated system, where evaluative processes are not only aligned but are linked together vertically and horizontally. According to him, vertically integrated evaluation takes place when information from one level of evaluation

(e.g., the evaluation of student learning) is used appropriately and ethically at another level, for example, by a teacher in a portfolio submitted in a tenure application or by a course review team. Some current approaches to quality assurance foster vertical integration by looking at the way evaluation is carried out and the way information is used rather than the actual data. Horizontal integration occurs when the information derived from an evaluation is used in a planned, strategic way. Most student evaluation data are not horizontally integrated. (Cannon 2001a, pp. 95–96)

In practice, however, the data gathered from ratings have been used for both formative and summative purposes (Young et al. 1999; Centra 2000; Cannon 2001a), both to enhance the instructional processes and to make, for example, personnel decisions. As Abrami (2001, p. 59) notes, rating forms are often the major source and sometimes the only source of information available for assessing a faculty member's teaching performance. Nevertheless, as Young et al. (1999) remark, if evaluations collected for summative purposes are found to differ from evaluations collected for formative purposes, then student evaluations collected for one purpose should not be used for another purpose. Abrami (2001, p. 69) suggests that the reporting system for summative decisions should not include the results of individual-specific items which are best used for teaching improvement purposes, that is, for formative decisions about teaching. According to Abrami, the reporting system for summative decisions should include the results of individual global items or, preferably, the reporting of an average of several global items. Knapper's (2001) view is more uncompromising, stating that serving different, possibly conflicting, needs is an almost impossible task. According to him, information that is useful for one purpose (e.g., a student selecting a college or a department head assigning teaching loads) may be unhelpful or even harmful for other purposes, such as providing feedback to a newly recruited teacher.

The results of a recent study by Chen & Hoshower (2003) into students' expectations of student evaluations showed that students generally considered an improvement in teaching as the most desirable outcome of the teaching evaluation system. The second most desirable outcome, according to the authors, was using teaching evaluations to improve course content and format. Using teaching evaluations for tenure, promotion and salary rise decisions, on the other hand, were seen as less important as was making the results of evaluations available to students for course and instructor selection.

Johnson & Ryan (2000, p. 111) stress the importance of understanding the roles and expectations of instructors with regard to teaching and learning in the following contexts: establishing evaluation criteria, determining the appropriate use of evaluation results, developing evaluations that are of optimum use in faculty development and, aligning evaluation and development efforts with department and institutional goals. Theall & Franklin (2000, p. 103) also note that the skills a teacher needs for using ratings in the context of reflective practice are quite different from the skills needed by a department head

when reviewing ratings or those of a committee when negotiating their sometimes conflicting interpretations of ratings and students' written comments.

3.2 Prevailing Evaluation Methods

To provide an overview of the various ways of evaluating education, this chapter briefly introduces the most typical evaluation methods. The purpose here is to show the position and the significance of student ratings compared to other evaluation methods within higher education and the characteristics of each method. The idea, discussed later, that the results of student ratings should be combined with data collected from different sources using different methods is strongly advocated in several research articles (e.g., McKinney 1997; Becker 2000; Kreber & Brook 2001; Ory 2001; Saroyan & Amundsen 2001, Emery et al. 2003). The researchers point out that student feedback provides only one source of information for evaluating teaching and units and that a questionnaire is only one method of gathering information from students (Cummings & Ballantyne 1999; Brennan et al. 2003). On the other hand, despite there being numerous ways of evaluating education, each way incurs some degree of criticism among the researchers.

As mentioned earlier, student evaluations in U.S. are typically focused on assessing teachers' characteristics whereas in many other western countries the focus is more on the contents of the education and the curricula. Irrespective of the focus of the evaluation, the contemporary evaluation methods being used are more or less of the same. Berk (2005, p. 49), concentrating on teaching effectiveness[13], has presented twelve potential sources of evidence along with such salient characteristics as type of measure to gather the evidence, the person(s) responsible for providing the evidence, the person or committee who uses the evidence and the type of decision typically rendered on the basis of that data. Berk also supports the view according to which a unified conceptualization of teaching evidence is proposed to use multiple sources of evidence. Table 3 below presents Berk's outline of these twelve sources.

[13] In U.S. the emphasis on student evaluations is closely related *teaching effectiveness.* Even Marsh (Marsh & Bailey 1993) referring to his earlier widely cited article (Marsh 1987) uses the terms student evaluations and students' evaluations of teaching effectiveness (SETE) as synonyms. According to Marsh & Bailey (1993, p. 1) there are thousands of studies on SETE's only. To define how teaching effectiveness should be measured, national (U.S.) standards have been developed (Berk 2005, p. 48).

Table 3. Twelve potential sources of evidence of teaching effectiveness (Berk 2005, 49).

Source of Evidence	*Type of Measure(s)*	*Who Provides Evidence*	*Who Uses Evidence*	*Type of Decision**
Student Ratings	Rating Scale	Students	Instructors/Administrators	F/S/P
Peer Ratings	Rating Scale	Peers	Instructors	F/S
Self-Evaluation	Rating Scale	Instructors	Instructors/Administrators	F/S
Videos	Rating Scale	Instructors/Peers	Instructors/Peers	F/S
Student interviews	Questionnaires	Students	Instructors/Administrators	F/S
Alumni Ratings	Rating Scale	Graduates	Instructors/Administrators	F/S/P
Employer Ratings	Rating Scale	Graduates' Employers	Instructors/Administrators	P
Administrator Ratings	Rating Scale	Administrators	Administrators	S
Teaching Scholarship	Judgmental Review	Instructors	Administrators	S
Teaching Awards	Judgmental Review	Instructors	Faculty Committees/Adm.:s	S
Learning outcomes	Test, Projects, Simulations	Students	Instructors/Curriculum Committees	F/P
Teaching Portfolio	Most of the above	Instructors, Students, Peers	Promotions Committees	S

* *F=formative, S=summative, P=program*

To form an overview of how the gathered evidence is utilised, Table 4 presents the most frequently used sources of evaluation within higher education, linked to their most typical uses. Each of these is discussed later in more detail.

Table 4. Typical sources of evaluation and their purpose of use in higher education (e.g., Johnson & Ryan 2000; Cannon 2001a; Brennan et al. 2003)

Method	**Purpose of Use**
Student Ratings	Improvement of the course / teaching, to inform prospective students Improvement of student learning
Self evaluations	Promote faculty development
Portfolios	Capture the complexities of teaching
Benchmarking	Administrative procedures
Peer Evaluation	Assess instructor knowledge and the value of instruction
Alumni and Employer Surveys	Determine the utility of learning experiences to students' careers

Student ratings, as previously described, are by far the most commonly used and most thoroughly researched evaluation method in higher education. The most typical form of student ratings is a traditional end-of-the-term paper-and-pencil rating form consisting of several quantitative multiple choice items and a few qualitative open-ended questions (e.g., Marks 2000). During the last few years paper-based questionnaires have all more often been replaced by online rating forms (Ballantyne 2000; McGourty et al. 2002b; Hoffman 2003), discussed more thoroughly later in this study.

Self-assessment as defined by Harvey (2004) is the "process of critically reviewing the quality of one's own performance and provision". According to him, self-assessment may be undertaken on an individual basis or, in the context of external quality review, on a collective basis. Self-assessment is used interchangeably with *self-evaluation* and *self-study* in the context of higher education quality. They all involve a process of self-reflection by the institution or sub-institutional unit being reviewed and the preparation of a document reflecting that self-reflection. (Harvey 2004) The suitability of self-evaluations has also evoked criticism among researchers, for example, Marsh (2001) has argued that despite the apparent appeal of instructor self-evaluations as a criterion of effective teaching, it has limited application.

Portfolios have originally been adapted from art and architecture, in which professionals display samples of their work for clients or employer (Centra 2000; Cannon 2001a; Johnston 2004). In higher education, portfolios are typically applied for individual faculty but there are also references of the portfolio concept being applied at the level of academic department (Cannon 2001a). On an individual level the most typically mentioned types of portfolio are a course portfolio, a professional portfolio and a teaching portfolio. Cannon (2001a) has described the differences between student ratings and portfolios as performance indicators, stating that student ratings tend toward an atomistic approach to evaluation whereas the portfolio stands as a way of bringing together in one holistic document an integrated range of evidence about teaching. While there are advocates of portfolios (e.g., Seldin 1993) there are also recognised problems with the use of portfolios, such as the lack of guidelines for their use and interpretations of their use (Theall & Franklin 2000).

Benchmarking is typically used in management and accounting contexts. According to Doerfel & Ruben (2002) benchmarking is a process, and as a process, it is one of comparison. They state that in the context of higher education, the comparisons can be made, for example, at the level of an institution, department, program, workgroup or specific process. One may also compare across academic, support and administrative units. The analysis of examining on any level can focus on the inputs, internal processes or outputs and these comparisons can be either with other higher education organisations or with other sectors. (Doerfel & Ruben 2002, pp. 6–7) According to Cannon (2001a) benchmarking has the potential to fit well in the higher education context, especially because it relies on a

methodology that requires hard data. However, Cannon also notes that benchmarking has been applied mainly to administrative procedures in higher education and in those areas of teaching activity that have a strong, external quality-oriented focus, such as continuing education and business education. Doerfel & Ruben (2002) also remark that universities have to ensure that their benchmarking partners share similar procedures, structures and missions.

Harvey (2004), describes *peer review* as the process of evaluating the provision, work process, or output of an individual or collective operating in the same milieu as the reviewer(s). Perhaps the most common form of peer review, also called as peer evaluation, is classroom observation (Becker & Watts 1999; Goldstein 1993), since it provides direct evaluations of the act of teaching (Berk (2005). There are both proponents and opponents of peer reviews. Braskamp (2000), for example, argues that peer evaluation needs to be an essential element in any evaluation of the faculty. According to him, peer evaluation does not violate academic freedom, but instead provides one of its best defences. Marsh (1987), on the other hand, strongly criticises the reliability of colleague and administrator ratings based on classroom visitation. He claims that peer ratings do not appear to correlate substantially with student ratings or with any other indicator of effective teaching. Similarly, Goldstein (1993, p. 10) has criticised peer reviews for the often haphazard way they are carried out, with colleagues occasionally sitting in on classes taught by the teacher being evaluated. In sum, Berk (2005, p. 51), for example, considers peer ratings of teaching performance and materials as the most complementary source of evidence to student ratings, but according to Berk, it should not be used solely for personnel, i.e., summative decisions.

Alumni and employer surveys can be used to align program improvement with institutional assessment (McGuire & Casey 1999) and accountability (Borden 2005). Although widely used, they are problematic in the sense that the information gathered from alumni and employers comes at a late stage when considering improvement of single courses. On the other hand Berk (2005) notes that they can provide new information on the quality of teaching, usefulness of course requirements, attainment of program outcomes, effectiveness of admission procedures, preparation for graduate work, preparation for real work and a variety of other issues not measured on the standard student ratings scale. These latter issues can elicit valuable information across courses or for the program as a whole (ibid.).

In addition to above, there are several other methods for gathering and interpreting information collected from students, such as structured group discussion, nominal group technique, focus groups and student interviews (see, e.g., Tiberius 2001; Brennan & Williams 2004; Berk 2005). Many of them are considered useful and advantageous, though the facts that limit their usability are that many of them involve a small group of students (around 6–12 students), requiring much work from personnel and are thus difficult to apply across an entire university. Brennan & Williams (2004) also mention the problems of ensuring that groups are representative of the student body, the use of good and independent facilitators as

well as the time taken to obtain, analyse and interpret the results. They also note problems of assuring the anonymity and confidentiality of the participating students.

The most common measures used for collecting data are rating scales (Berk 2005). In addition, some scholars (e.g., Brennan et al. 2003) observe that student feedback can also be obtained in many ways other than through the administration of formal questionnaires, for example, meetings of staff-student committees, student representation on institutional bodies and even casual comments made inside or outside the classroom. However, Brennan et al. point out two advantages that surveys using formal instruments have: they provide an opportunity to obtain feedback from the entire population of students and they document the experiences of the student population in a more or less systematic way.

Several authors (e.g., Theall & Franklin 1991; Seldin 1993; Scriven 1995; McKinney 1997; Ory, 2000; Johnson & Ryan 2000; Abrami 2001; Cannon 2001a; Kreber & Brook 2001; Saroyan & Amundsen 2001) emphasize the need of multiple and multidimensional sources of evaluation, such as student ratings, peer reviews, self-evaluations, portfolios, site visits by outside experts as well as alumni and employer surveys. Regardless of the multidimensionality of student ratings as such (i.e. their ability to measure several different aspects of teaching), it is widely recognised that student ratings represent only one source of information and that no single student rating item, nor set of related items will be useful for all purposes (c.f., Cashin 1995; Scriven 1995; Theall & Franklin 2001a). The purpose of the evaluation should thus affect to the way the ratings are constructed. In practice, however, when evaluating teaching and learning, student ratings are still the predominant method in most universities worldwide (e.g. Seldin 1993; Kwan 1999; Saroyan & Amundsen 2001). Since the results of student ratings play a major role in such important decisions, such as tenure decisions or changing contents of a course or a module, a good deal of pressure is put on their usability and reliability.

3.2.1 Characteristics of Student Evaluations

Traditional student ratings have usually been implemented to explore students' opinions of their education. By analysing such ratings it is possible to find out the strengths and weaknesses of the lectures, the programme or other specific areas in education (see, e.g., Betoret & Tomás 2003). Missing from the analysis, however, is information about the *students* besides their direct survey responses. Worthington (2002) has addressed this inadequacy remarking that "little attention has been paid to measuring what appears to be a relatively important source of ratings bias, that is, the perceptions and characteristics of students themselves".

There are several phases where evaluation information is needed. Even though there are many cultural and country specific differences in the use of student evaluations, they are typically used specifically for evaluating the education the students are given. As Candoli & Stufflebeam (2003, p. 421) describe "evaluation is needed to identify, understand, and set priorities on students' educational needs and it is also needed in the course of preparing and assessing the merits of program plans. Since most programs either do not follow plans exactly or may improve on plans, staff and oversight personnel need evaluations of the implementation process. After programs have been planned and delivered, they must be evaluated to determine how they impact on students, to see if they have served all students adequately, or if adjustment is needed to reach students for whom they were intended."

Scriven (1995) lists several strong arguments for using student ratings to evaluate teachers: students are in a unique position to rate their own increased knowledge and comprehension as well as changed motivation toward the subject taught, they can also observe and rate facts that are relevant to competent teaching, and they can also identify and rate whether the teacher in enthusiastic. Despite the well-known problems regarding students' evaluations of teaching, such as the absence of any definition of "effective teaching" (see, e.g., Sproule 2000) or the lack of suitable criteria by which teaching faculties are evaluated (Stake & Cisneros-Cohernour 2000), they are generally considered to be reasonably reliable and useful. For example, Kulik (2001) argues that teachers profit from the information that ratings provide, and if they are provided with instructional consultation, they profit even more from ratings' results.

Student evaluations are typically collected at the end of the term during the last lecture before the final exams. Broder & Dorfman (1994, p. 236) divide the student evaluation of the teaching process into three parts: first, the students are asked to rate the quality of the teacher and the course with an end-of-term survey or questionnaire. Second, these survey data are compiled and a set of summary statistics is produced and distributed to faculty and administrators. Third, these summaries are used for both personal and personnel decisions. (ibid.)

This kind of information is often used for both formative and summative purposes, even though in some research papers the utility of end of term is questioned by criticising that the feedback gathered this way will not benefit those students who originally gave the feedback and thus have already completed the course. Some universities or lecturers also collect midterm feedback which provides a way to find out what students are thinking before the end-of-semester evaluation (Lewis 2001b, p. 38). One of the most unusual experiments of student evaluations is a study reported by McKenzie et al. (1998) where students of architecture presented their impressions of the preceding year through drawings.

Typical student evaluation forms consist of both multiple choice and open-ended questions. Multiple choice questions are often presented in a Likert-form 4- or 5-point scale (Heikkilä 2001) and contain both global and specific items. Many teachers tend to value most the open ended questions, but sometimes the response rates to those are perceived to be relatively low. According to Lewis (2001a, p. 31), it is possible to provide prompts for students. These can be used increase the usefulness and frequency of student comments, for example questions like "What helped your learning the most in this class?", "What hindered your learning the most in this class?" or "What suggestions for changes do you have that would have improved your learning in this class?" (Lewis 2001a). He also observes that students tend to write more and provide constructive comments when the prompts are provided than when they are not. However, there is also criticism as to the usefulness of students' written comments, for example Marsh (1987, p. 280–281) has denied the utility of open-ended questions. He argues that "The useful information from comments that cannot be obtained from rating items is idiosyncratic information that cannot easily be classified into generalisable categories, that is so specific that its value would be lost if it was sorted into broader categories, or that cannot be easily interpreted without knowledge of the particular context".

Many scholars highlight the importance of ensuring that the results of the evaluation form measure what they are intended to measure. For example Ouimet et al. (2004, p. 247) note that "even for survey instruments where it has been shown that students generally attach similar meaning to both items and their response options, certain groups of students may interpret some questions differently or find the wording of items to be unclear". According to Tiberius (2001, p. 63) the design and administration of student evaluations is a highly developed art. According to him, principles of good design help designers write questionnaires that produce valid, reliable and useful information. Mehrens (1998), on the other hand, argues that the type of format is used probably makes far less difference than how it is used.

3.2.2 Student Evaluations on Assessing Teaching and Education

For a full understanding of the scope of student evaluations, it is of importance to appreciate their dualistic nature: on one hand, they are recognised as a key way of receiving feedback from given education; on the other hand, they are also sometimes condemned as being faulty and even misleading. The latter is largely due to the fact that the results of student evaluations may even be used in salary and tenure decisions, as described earlier in this study, which naturally makes concerns over their accuracy understandable.

One of the most contentious issues in concerning student evaluations is the students' ability to assess the instruction they are given. Braskamp & Ory (1994, p. 99) consider students to

be competent in assessing the following issues: student-instructor relationships, their views of the instructor's professional and ethical behaviour, their work load, what they have learned in the course, fairness of grading and the instructor's ability to communicate clearly. However, the quality of course content and the instructor's scholarship in the field are examples of issues where students' ability to assess is often questioned (Braskamp & Ory 1994; Mason et al. 1995). Kulik (2001), in contrast, claims that students are competent to rate their instruction by pointing out that studies show student ratings to agree well with other measures of teaching effectiveness, such as expert observations and alumni ratings. The problems regarding the students' ability to evaluate the education they are given and especially the lecturers are discussed in more detail in Chapter 3.3.3, *Ethical and Other Issues*.

Many researchers stress that in the construction of an evaluation instrument the content of the dimensions to be surveyed is of major importance. The students' ability to answer the questions as well as the appropriateness of the questions themselves largely affects the kind of results the evaluations provide (e.g., Serva & Fuller 2004). Menges (1991) also stresses the importance of wording the items correctly by noting that "questionnaire items drafted by faculty may be poorly worded unless chosen from pretested, validated lists". The most typically assessed dimensions in student evaluation forms include the teacher's communication skills, rapport with students, course organisation, student self-rated accomplishment, course difficulty, grading and examinations (Braskamp & Ory 1994).

Rando (2001) places questions into two categories: self-report and direct assessment questions. In self-report questions the students are asked to rate, rank, report or mention some aspect of their learning experience, often using a Likert-type scale. In direct assessment questions ask students to show what they know or how they know it. According to Rando, the purpose of direct assessment questions is to assess the quality of instruction and its impact on student learning at any point during the course, i.e. before a lecture, during a break in the lecture or after a lecture.

Oliver & Conole (1998) also stress the importance of choosing an appropriate methodology for evaluation. For example, the question "What factors influence how well students do?" suggests an exploratory study, where the evaluator has little or no idea about the factors that will influence learning. In contrast, the question "Which of the following factors influences how well students do?" suggests a comparative study, where the factors have already been found and what remains is to test them to demonstrate their influence. (Oliver & Conole 1998)

Knapper (2001, p. 6) emphasises the importance of careful preparation by noting that before undertaking an evaluation of teaching, the motives of evaluation should be carefully considered. Johnson & Ryan (2000, p. 115) mention setting priorities as the most important

issue. According to them, it is important that those who design and plan evaluations consider evaluation priorities and how these priorities may vary in different situations. For example, administrative decision making is more pertinent to some evaluations than others, only evaluations in a certain situation need to stand up to legal scrutiny, and not all evaluations need a strong faculty development component. (Johnson & Ryan 2000)

Developing and implementing phases are not the only important steps in evaluations. Seldin (1993) remarks that even a carefully developed student rating form may be invalidated by poor administration, for example, a sporadic rating schedule or instructions that bias responses. And a third aspect which should be taken into account when student evaluations are discussed is the personnel's ability to interpret the ratings results. What has been found especially problematic is the fact that many factors influence learning. For example, Theall & Franklin (2000) have noted that "understanding how ratings relate to effective teaching requires someone with a wider range of skills than either a statistician or even an expert teacher may possess". Franklin (2001, p. 87) enlarges the theme by arguing that if ratings are used properly, they can provide useful information about the quality of instruction students have received. According to her, "No matter how good the questionnaire was, if reviewers lack the needed skills and knowledge, the ratings are no more than a Rorschach test with numbers instead of inkblots. Readers will look at the report and imagine they see what they already believe."

However, there are also researchers who doubt whether it is at all possible to utilise the results of student ratings for differentiating those underlying factors. For example, McKinney (1997, p. 3) has stated that "research indicates that students' evaluations should, generally, not be used to draw conclusions about a course. Research indicates that most of the variance in ratings is accounted for by the instructor, not the course." The problematics of interpreting ratings results will be discussed later in this study in more detail.

According to Fenwick (2001), there are at least three major influences besides teaching which affect student success: institutional (e.g., atmosphere in a department, class size, faculty workload), dispositional (e.g., students' attitude to instruction or individual capability) and social (e.g., students' financial situation or prior knowledge and experience). These three elements are complemented by the findings of Tynjälä et al. (2005) which indicated that in addition to the students' own study strategies and orientations, the students' perceptions of their learning environment were also related in their study orientations, which in turn were related to their study success.

There are several examples of instruments that have been developed to measure distinct components of teaching effectiveness. Marsh (1987, p. 266) lists the following four examples: Frey's Endeavor Instrument, The Student Description of Teaching (SDT) questionnaire, Marsh's SEEQ instrument and The Michigan State SIRS instrument, all which

are based on a factor structure. In Australia the prevailing evaluation method is the Course Experience Questionnaire (CEQ), which is used to evaluate teaching effectiveness in both departmental and institutional level (see e.g., Knapper 2001).

In many cases the student evaluations have focused on individual lecturers and their characteristics, rather than on education at a more general level. Quoting Knapper (2001): "Effective learning is dependent on far more than an individual instructor on a course". Cannon (2001a, pp. 88–89) has criticised the current tendency to focus evaluations on single instructors by arguing that "although individuals make important contributions to the work of the team, and these contributions can be identified, and evaluated, it is nevertheless true that the experience of students attending a university over the three or four years of an undergraduate degree program is shaped by the sum of the contributions of several different team members. From this perspective, a valid evaluation of teaching should focus less on individual instructors and more on the ways in which these contributions come together in students' experience of their program as a whole." Lappalainen (1997, p. 13) concurs, noting that even though course questionnaires are important from the viewpoint of development of that particular course or a particular lecturer, much more important for the unity is to evaluate the whole curriculum, the parts the degree is built upon and how the parts fit together.

In spite of the widespread use of the student evaluation questionnaires, there is also much criticism concerning the usability of evaluations given by students. Some scholars question the whole notion of student evaluations whereas others criticise specific issues such as the fairness of student evaluations to faculty members (Stapleton & Murkison, 2001). Goldstein (1993) disapproves of the use of plain standardised (i.e. computer-scored) student evaluations, arguing that without accompanying open-ended evaluations they are almost worthless and can even be misleading. However, he acknowledges that "although standardized questionnaires by themselves do not provide meaningful information, their very standardization can protect faculty against arbitrary and selective use of non-standardised "information" that can lend itself to manipulation in unscrupulous hands".

3.3 Problems in Using Student Evaluations

It may at first appear paradoxical that although there is said to be more research on student ratings than on any other aspect of higher education (e.g., McKeachie 1996), researchers still find the ratings such a controversial issue (Nasser & Fresco 2002; 2006). The disagreement between researchers can clearly be seen, for example, in the provocative titles of the articles on student evaluations, such as "Thoughts of the Lesser Evil: student evaluations" (Carey 1993), "What Do They Know Anyway?" (Felder 1995), "Do Online Ratings of Instruction Make $ense?" (Bothell & Henderson 2003) or "Characteristics of Mindless Teaching

Evaluations and the Moderating Effects of Image Compatibility" (Dunegan & Hrivnak 2003). This discrepancy is probably attributable to the fact, that student ratings are not only a subject of interest among the researchers in the field; they are also an actively used instrument in evaluating teaching and learning in higher education (e.g., Seldin 1993).

The general criticism of the use of student evaluations includes claims such as:

- the use of students' evaluations of teaching effectiveness may actually weaken the education of faculty, if lecturers alter their teaching in the hope of achieving good personal evaluation results (e.g., Emery et. al. 2003)

- student evaluations are nothing but "popularity contests" that can be manipulated by the instructor's grading policy or other elements, e.g., classroom entertainment quotient (Felder 1995; Becker & Watts 1999; Emery et. al. 2003)

- student evaluations lack discriminant validity; "no matter, how reliable the measures, student evaluations are no more than perceptions and impressions" (Marks 2000)

Student ratings are also a topic that triggers continuously criticism not only among the researchers and practitioners in the field but also among researchers in other disciplines as well as the public media. Several journals, such as *Perspectives on Political Science* and *Journal of Economic Perspectives*, have published articles (see, e.g., Platt 1993; Weissberg 1993; Young 1993; Becker 2000) in which student evaluations are seen mostly in a negative light. And the number of articles concerning student evaluations that have been published in different newspapers in different countries is myriad.

Though an actively researched topic, there remain several problems and pitfalls associated with the use of student ratings results. For example, the problems identified forty years ago by Remmers (1966) still persist, such as the tests being out of line and thus not providing valid measurement in the areas in which they are used, or the effectiveness of an individual teacher being judged on the basis of test scores made by the teacher's class ignoring other factors affecting the results. The following chapters consider the problems associated with the use of student ratings most often presented in the research literature of higher education, namely validity and reliability. Other issues particularly relevant to student evaluations are also discussed.

3.3.1 Validity

Validity, as a term, is usually defined as follows: "A test is valid if it measures what it purports to measure" (Scriven 1991, p. 372). Validity of an evaluation was defined in the 1960s by Remmers et al. (1966, p. 119) as "the degree to which it measures what it is intended to measure" and that definition still seems to be accepted by the researchers community (Benett 1993; Cashin 1995; Saroyan & Amundsen 2001; Deasy 2004). Validity is often discussed in connection with another term, reliability. According to Popham (1981) test reliability is a necessary, but insufficient, condition for test validity. As Popham notes, in order for a test to be valid, it must be reliable, and unreliable tests cannot possibly be valid. Thus, he concludes that merely because a test is reliable, this does not guarantee its validity.

The validity (whether the questionnaire measures teaching effectiveness directly or indirectly) and reliability (how faithfully it measures teaching effectiveness across the many students who responded and the many classes that were rated) of student ratings is the most contentious subject in the field of higher education (Franklin 2001). The ongoing debate concerning validity issues has caused heated discussion among researchers for decades. For example Popham (1981) reports that already by the 1950s the number of different ways of thinking about validity was getting out of control, with "all sorts of exotic validity types and validity terms finding their way into the measurement literature". Popham mentions the terms "intrinsic", "extrinsic", "divergent", "convergent", "face" and "two-faced" validity as examples of the polymorphous field of validity issues being discussed at the time.

The view of Braskamp & Ory (1994, p. 92) is that "validity refers to the integrity and appropriateness of the conclusions or generalizations drawn from the evidence collected". They consider the validity of evidence to be dependent on the purpose of the assessment or evaluation. Thus, according to them, validity is a relative matter and even more difficult to determine than reliability. They add that "Validity does not refer to the instrument or method of collecting data; rather it refers to the inferences and generalizations based on the evidence. An instrument can be valid for one use but not for another. For example, student ratings of teaching may be highly valid for assessment of a faculty member's ability to make a clear presentation but not valid for determination of the faculty member's knowledge of the subject matter or status in the discipline."

One of the most quoted publications on the validity aspects of student evaluations in higher education was published in 1987 by Herbert Marsh. In the findings of his extensive monograph on student evaluations of university teaching, Marsh (1987) states that "student ratings are clearly multidimensional, quite reliable, reasonably valid, relatively uncontaminated by many variables often seen as sources of potential bias, and are seen to be useful by students, faculty, and administrators". Marsh's findings intensified the debate among scholars which had started decades earlier and which remains unsolved (for details,

see, e.g., d'Apollonia & Abrami 1997; Greenwald & Gillmore 1997b; McKeachie 1996; Theall et al. 2000).

The previously recognised fact that student ratings are often, especially in the U.S., regarded as a synonym for faculty evaluations, is also a reflection of the way validity issues are discussed among researchers and practitioners. When the main focus of student evaluations has been on teachers and teaching effectiveness, the discussion of validity issues has and still remains similarly centred on teachers and teaching issues. For example Cashin (1995), Kulik (2001) and Deasy (2004) interpret the validity of student ratings as the degree to which they reflect teaching effectiveness, yet the authors note that there is no agreed definition of "effective teaching". As Kulik (2001) succinctly puts it, "without a perfect criterion, it is impossible to reduce the validity of student ratings to a single number". Cashin (1995) suggests that the best solution is to try various approaches, collecting data that either support or contest the conclusion that student ratings reflect effective teaching. Young et al. (1999, 181) emphasize the utility of student evaluations by stating that validity assesses the degree to which student evaluations of teaching performance in the classroom setting reflect actual teaching performance as exhibited by a faculty member.

Young et al. (1999) and Deasy (2004) suggest that in order to establish the validity of ratings for classroom performance of instructors provided by students, ratings must be related to other criteria purported to measure teaching performance. Cannon (2001a) on the other hand, questions the practice of focusing the evaluation on individual lecturers by stating that "valid teaching should focus less on individual instructors and more on the ways in which these contributions come together in students' experience of their program as a whole". Scriven (1995) observes that the validity of student ratings forms is also dependent on how and when the evaluations are administered. According to Scriven, to be valid, the student ratings results must be obtained from properly administered tests, stringently controlled data collection and thorough analysis of test results.

Scriven (1995) has identified the following nine potential sources of validity for student ratings of instruction:

1. The positive and statistically significant correlation of student ratings with learning gains.

2. The unique position and qualifications of the students in rating their own increased knowledge and comprehension.

3. The unique position of the students in rating changed motivation (a) toward the subject taught; perhaps also (b) toward a career associated with that subject; and perhaps also (c) with respect to a changed general attitude toward further learning in the subject area, or more generally.

4. The unique position of the students in rating observable matters of fact relevant to competent teaching, such as the punctuality of the instructor and the legibility of writing on the board.

5. The unique position of the students in identifying the regular presence of teaching style indicators. Is the teacher enthusiastic; does he or she ask many questions, encourage questions from students, etc.?

6. Students are in a good position to judge – although it is not quite a matter of simple observation – such matters as whether tests covered all the material of the course.

7. Students as consumers are likely to be able to report quite reliably to their peers on such matters of interest to them as the cost of the texts, the extend to which their attendance is taken and weighted, and whether a great deal of homework is required – considerations that have little or no known bearing on the quality of instruction.

8. Student ratings represent participation in a process often represented as "democratic decision-making".

9. The "best available alternative" line of argument.

Greenwald (1997, p. 1185) has analysed the prevailing views of several researchers and divided validity concerns into four issues: (1) conceptual structure: are ratings conceptually unidimensional or multidimensional; (2) convergent validity: how well are ratings measures correlated with other indicators of effective teaching; (3) discriminant validity: are ratings influenced by variables unrelated to effective teaching, and (4) consequential validity: are ratings results used in a fashion that is beneficial to the educational system. In the findings of a large study conducted at the University of Washington Greenwald & Gillmore (1997a) report an evidence for convergent validity of student ratings but a deficiency in discriminant validity which in their report was corrected to some extent by adding theory-based statistical adjustments to original ratings data.[14]

Kulik (2001) reflects the tendency of several researchers to emphasize the construct validation approach to student ratings. According to him, this approach "requires researchers to show that ratings correlate to a satisfactory degree with other admittedly partial and imperfect measures of effectiveness". Marsh (1987, p. 302) argues that since there is no single indicator for effective teaching, the validity of student evaluations must be demonstrated through a construct validation approach. According to Marsh (1987, p. 305) the construct validity of student evaluations requires that they are related to variables that are indicative of effective teaching, but relatively uncorrelated with variables that are not (i.e.

[14] These findings aroused resistance among other scholars who criticised e.g., the use of such adjustments. The criticism and the authors' comments on them are added on the postscript of the original article (Greenwald & Gillmore 1997a).

potential biases). Marsh considers student evaluations to be related to a number of varied criteria including the ratings of former students, student achievement in multisection validity studies, faculty self evaluations of their own teaching effectiveness and possibly the observation of trained observers. According to Saroyan & Amundsen (2001, p. 342) overall findings from research on construct validity (which they define as the degree to which student ratings accurately measure the construct of effectiveness) seem to be consistent across different studies.

There are several views on the various aspects of validity of the student ratings (e.g., Greenwald & Gillmore 1997a; Marks 2000; Marsh 2001) and numerous articles reviewing the vast amount of research for and against the validity of student ratings (e.g., Greenwald 1997; Kulik 2001; Felton et al. 2004). Probably the most frequently cited issues relating to the validity of student ratings concern the *potential biases of ratings* and especially the *grading leniency*. Worthington (2002, p. 49) has described the concern of bias as follows: "There is the possibility that background characteristics (or factors that have nothing to do with the instructor's behaviour or effective teaching) could bias student ratings. If this is the case, students' evaluations as a valid indicator of teaching effectiveness, whether for formative (quality improvement) or summative (quality assurance) purposes, could be called into question."

However, as Marsh (1987, p. 310) points out, the existence of a significant correlation between students' evaluations and some background characteristic should not be interpreted as support for a bias hypothesis. Marsh adds that even if a background characteristic is causally related to students' evaluations, there is insufficient evidence to support a bias hypothesis. According to him, it is not enough to show that a variable is correlated with student ratings and that a causal interpretation is warranted, but it must also be shown that the variable is not correlated with effective teaching. There are also study methods such as multisection validity study (e.g., Abrami 2001) which try to minimise contextual influences.

Oliver & Sautter (2005) divide sources of bias into three categories: (a) personal attributes of the student such as gender or expected course grade, (b) situational characteristics of the learning environment such as the size of the class and (c) genetic traits of the instructor, such as gender or attractiveness. In spite of the active research of recent decades, especially into the potential correlations between grades – actual or expected – and ratings results, there are still contradictory results concerning all the aforementioned elements.

Grading leniency is suspected to exist if the known or expected course grades and the evaluations the students give on their teachers are dependent on each other. Marsh (1987, p. 303), although arguing for the validity and reliability of student ratings, has noted that "too little attention has been given to a grading leniency or grading satisfaction effect in multisection validity research even though such an effect has been frequently posited as a

bias to students' evaluations of teaching effectiveness." Greenwald & Gillmore (1997a) present five theories, each explaining the grades-ratings correlation:

1. Teaching effectiveness influences both grades and ratings. That is, strong instructors teach courses in which students both learn much and give appropriately high ratings to the course and to the instructor.

2. Students' general academic motivation influences both grades and ratings. Students with strong academic motivation should do better in their course work and should more fully appreciate the efforts of the instructor (which may in turn increase the quality of the instruction) compared with unmotivated students.

3. Students' course specific motivation influences both grades and ratings. According to this theory, students' motivation can vary from course to course rather than being a fixed characteristic of the student.[15]

4. Students infer course quality and own ability from received grades. This theory leans on the attribution theory, which, applied in the academic context, suggests one to expect high grades to be self-attributed to intelligence or diligence and low grades to poor instruction.

5. Students give high ratings in appreciation for lenient grading. This would mean that the instructor praises the student by giving him/her a high grade and the student in turn, provides the instructor high ratings. This theory is also called leniency or grade-satisfaction theory.

The first three theories explain the grades-ratings correlation by assuming that a third variable influences both grades and ratings, whereas the last two theories assume grades to have a causal influence on ratings. What they emphasize, however, is that to observe that ratings measures are sensitive to grading leniency does not mean that the ratings fail to measure what they are intended to measure. They may just be measuring more than they are intended to measure. (Greenwald & Gillmore 1997a)

The first three theories also relate to the students' motivation and learning styles. In addition to Marton's (1982) classic division of students' learning styles into deep and surface learners and several theories relating to the way the students' learning approaches develop in the course of their studies (see, e.g., Biggs 1979), the recent findings of Tynjälä et al. (2004) also suggest that a connection exists between individual students' learning orientations and the approaches on how the university supports and evaluates learning. Thus, not only do

[15] Students' motivation may also vary even within a single course. This can clearly be seen, for example, in the findings of Salminen & Mirola (2001) concerning pedagogical experiments in one particular course.

students' personal attributes affect their study success, but also their compatibility with the university's guidance and evaluation methods (ibid.).

Worthington (2002, pp. 49–50) on the other hand, identifies four broad sets of factors of the biases in existing student evaluations:

1. the characteristics associated with the administration of student evaluation

2. the second group of background factors positioned to influence student evaluations of teaching that is concerned with the characteristics of the course itself

3. the third group of background factors related to the characteristics of the instructor, for which a large number of factors have been proposed and duly tested

4. the final group of background factors concerned with the characteristics of students themselves and the biases introduced into students' evaluations of teaching, such as the role of prior subject interest, the gender of students, the role of expected grade or student perceptions and the impact of student age.

Sojka et al. (2002) in comparing the attitudes of faculty and students, report that the members of faculty, in particular, were likely to assume that they can get higher ratings by grading more leniently, whereas the students were less likely to agree with this. The existence of grading leniency within higher education institutes has been under active research for decades, yielding contradictory results. There are several studies showing positive correlations between student ratings of teaching and students' grades (Greenwald & Gillmore 1997a; Millea & Grimes 2002), and some researchers (e.g., Felton et al. 2004) argue convincingly that instructors offering easy courses tend to be rated more highly. However, there are also studies (e.g., Gramlich & Greenlee 1993; Husbands 1998; Centra 2003) that show only minor, if any, correlations between course grades and high ratings results. For example, Gramlich & Greenlee found only minor positive correlation between the student evaluations of teaching and student grades.

Centra (2003) conducted out an extensive study of approximately 55 000 classes, concluding that expected grades[16] generally did not affect student evaluations and some courses were even rated lower when the expected grades were high. Centra suggests the results are due to

[16] As Centra (2003) points out, final grades in a course are typically not known to students at the time they complete a student evaluation form and therefore studies of a possible grading-leniency bias on ratings can appropriately use expected grades. This, however, applies only to traditional paper-based evaluations. As this study seeks to demonstrate, the shift towards online-based evaluations makes it possible to link separate evaluation results anonymously to corresponding background data, such as final grades and the date of completing the course. In the case of Centra's study, the unit of analysis was the class, not the individual student. Thus, the expected grades and other variables were all class average (mean) values.

the particular wording of the questionnaire: instead of asking students to rate the teacher or a course, generally on an excellent to poor scale, Centra used a questionnaire that asked students to rate the quality of the instruction as it contributed to their learning in the course. Similarly, Mason et al. (1995) discovered that students with lower grade point averages gave higher evaluations for course quality than those with better grade point averages. Franklin (2001) referring to her own and Michael Theall's earlier study observes that "academic ability of students as measured by grade point average has shown little relationship to the ratings they give; students who do poorly are just as appreciative of good teaching as students who do well". According to Mason et al. this would suggest that better students were tougher in rewarding course quality but more lenient in their ratings of lecturers.

Braskamp & Ory (1994, p. 100) suggest an explanation for the correlation between high grades and favourable ratings results by pointing out that high grades from a course may just indicate that the students learned a lot because of effective teaching. Similarly Kulik (2001, p. 20) is critical of research reports which suggest that high student ratings reflect lenient grading, arguing that good teaching may stimulate students to perform well in a course, and thus receive high grades, and may also lead students to give the course high ratings. According to Franklin (2001) giving students unearned good grades and thus reducing the effort needed to earn grades have been demonstrated to have virtually no payoff for raising ratings in the long run, provided ratings are anonymous. On the other hand Chambers & Schmitt (2002) present results that support the grading leniency theory. However, they note that more research is needed into the validity of students' performance ratings of their instructors.

The effect of class size, course level, the time of day that the course being evaluated is held, academic discipline, the subject matter of the course, gender of the evaluator as well as the gender of the teacher being evaluated have received intensive investigation for several decades yet continue to yield contradictory results (Feldman 1978; Mason et al. 1995; Adams et al. 1996; Husbands 1998; Kerridge & Mathews 1998; Kwan 1999; Franklin 2001; Griffin 2001; Worthington 2002). For example, Becker & Watts (1999, p. 344) have argued that correlation studies have not adjusted for the sample selection associated with student course withdrawals, or with absenteeism on the day evaluations are administered. Thus, according to them, if course persistence and class attendance are systematically related to student evaluation of teaching (SET) scores, least-squares estimators of the relationship between SET scores and measures of learning will be biased. Algozzine et al. (2004) examined the conflicting research results on the construct validity of student ratings in terms of the relationships between course, student and instructor characteristics and student ratings of teaching. Wilhelm (2004) has presented a summary of research factors that influence student evaluations of teaching which, for the most part, also illustrates the contradictory nature of research results. Similarly Ory (2001), Kulik (2001), Worthington (2002) and Nevgi & Lindblom-Ylänne (2003) also allude to a profusion of research findings presenting factors

potentially having an effect on the ratings results or correlating with other measures of teaching effectiveness.

The following research results exemplify this plethora of findings: Hiironniemi & Tuunainen (1995) note that students who had studied longer were more critical than students who were at the beginning of their studies. Similarly, the findings of the study of Chen & Hoshower (2003) showed that freshmen had a significantly higher regard for professors and student-generated evaluation systems than did senior students. Griffin (2001) has showed that students who had heard positive information regarding the instructor's reputation rated both the instructor and a course higher that those students who heard negative prior information about the instructor. Franklin (2001, pp. 96–97) observes that students tend to give slightly higher ratings to their majors or electives than to courses taken to fulfil a requirement and also that new or revised courses frequently get lower-than-expected ratings the first time out. However, Seldin (1993) argues that in general, factors that may initially seem to bias student ratings actually have scant or no effect. For example, Seldin (1993) estimates that the influence of significant relationships between extraneous variables and student ratings accounts for only 12 to 14 per cent of the variance between positive and poor ratings. Similarly, Marsh (1987, p. 308) argues that between 5 percent and 25 percent of the variance in student ratings can be attributed to the nature of the student ratings items, the background characteristics or the academic discipline and the institution where the study was conducted.

According to Marsh (1987, p. 309) the finding that a set of background characteristics are correlated with students' evaluations of teaching effectiveness should *not* be interpreted to mean that the ratings are biased, though this conclusion is often inferred by researchers. Marsh adds that "perhaps more than any other area of student-evaluation research, the search for potential sources of bias is extensive, confused, contradictory, misinterpreted, and methodologically flawed".

There is still an active interest in identifying new factors that have an influence on student evaluations. For example, Greimel-Fuhrmann & Geyer (2003) have analysed the effect of interest and the students' liking for their teacher on ratings. Ogier (2005) has examined the rating differences between native English-speaking lecturers and English as a second language lecturers, and Afonso et al. (2005) have compared the results of anonymous and open evaluations, showing that faculty receive significantly lower scores on the anonymous evaluations than on open evaluations where the evaluator's identity is known to the faculty member. In general, the importance of anonymity of responding students is recognised in several studies (e.g., Cummings & Ballantyne 1999; Ballantyne 2000; Cummings et al. 2001; Spencer & Schmelkin 2002; Brennan & Williams 2004). The findings of Spencer & Schmelkin (2002) showed that those students who were concerned about repercussions were the same who were more reticent about making evaluations. On the other hand, Platt (1993) has criticised the anonymity of student evaluations by arguing that students are not personally

responsible for erroneous answers and there are thus no personal consequences for a negligent, false, or even malicious misrepresentation.

Several researchers have presented examples of the possible problems with interpreting ratings results. For example Seldin (1993) cautions against misinterpreting small differences in ratings results, such as regarding a professor receiving a rating of 3.72 as a significantly better teacher than a colleague with a rating of 3.70. Likewise McKeachie (1996) argues against comparisons by claiming that "comparing teachers with averages such as 4.3 and 4.1 is like comparing apples to oranges. We can tell a good apple or a good orange, but judging whether a good apple is better than a good orange is a more difficult task." In a similar vein, Adams (1997, p. 10) notes that "One of the problems with this type of student evaluation of faculty is that few administrators are trained to interpret these numbers. It is not uncommon for administrators to eye-ball the scores and assume that scores below the mean are bad and those above it are good", or as Weissberg (1993) remarks: "Disputes over intellectual content can be replaced by low-key discussions of whether, for example, a 3.9 on a 5.0 scale is 'poor' or 'low average'." Theall & Franklin (1991, p. 84) also remark that an overall rating of teaching performance, although summary in nature, is not necessarily the "average" of more specific ratings; on the contrary, ratings on specific teaching skill areas may be very different from overall ratings. It is also important to observe whether we are measuring the key variables of teaching effectiveness or whether some other variables are becoming more important just because they are measurable (Shevlin et al. 2000, p. 398). Sproule (2000) derides the value of any evaluation data by arguing that the likelihood of gaining meaningful and valid inferences from raw student evaluation of teaching data is nil.[17]

Marsh (1987, pp. 309–310) has listed some methodological problems in the search for potential biases to students' evaluations:

1) Using correlation to argue for causation – the implication that some variable biases student ratings argues that causation has been demonstrated, whereas correlation only implies that a concomitant relation exists.[18]

2) Neglect of the distinction between practical and statistical significance – all conclusions should be based upon some index of effect size, as well as on tests of statistical significance.

3) Failure to consider the multivariate nature of both student ratings and a set of potential biases.

[17] In his article Sproule (2000) supposes that the data has been gathered by an instrument which does not provide for the collection of any background data on the student respondent, such as age, gender, major or grade point average.

[18] The problem of causality has also been brought out in more recent articles, for example by Paswan & Young (2002).

4) Selection of an inappropriate unit of analysis. Since nearly all applications of students' evaluations are based upon class-average responses, this is nearly always the appropriate unit of analysis. The size and even direction of correlations based on class-average responses may be different from correlations obtained when the analysis is performed on responses by individual students. Hence, effects based on individual students as the unit of analysis must also be demonstrated to operate at the class-average level.

5) Failure to examine the replicability of findings in a similar setting and their generalisability to different settings – this is particularly a problem in studies based on small sample sizes or on classes from a single academic department at a single institution.

6) The lack of an explicit definition of bias against which to evaluate effects – if a variable actually affects teaching effectiveness, and this effect is accurately reflected in student ratings, then the influence is not a bias.

7) Questions of the appropriateness of experimental manipulations – studies that attempt to simulate hypothesized biases with operationally defined experimental manipulations must demonstrate that the size and nature of the manipulation and the observed effects are representative of those that occur in natural settings, i.e., they must examine threats to the external validity of the findings.

Cranton (2001, p. 15) argues that the fact that student ratings are generally reliable and valid is an outcome of at least three factors: (1) the people who create forms generally tend to agree on what should be included; (2) generally students agree on what good teaching is within a special context; (3) individual differences among student responses are usually statistically removed. According to Cranton we should continue to use student ratings of instruction but also recognise that they are subjective and interpretive and that they represent only one perspective – a collective perspective where individual voices are lost.

The shift from traditional paper-and-pencil questionnaires towards online ratings has raised concerns of new potential biases such as lowered response rates and demographic differences among respondents and nonrespondents. These will be discussed in greater detail in Chapter 5. The empirical part of this study introduces the possibility of discovering other biases not previously reported in the research literature. These are investigated by utilising the link of an online rating system to a university-held database.

3.3.2 Reliability

Consistency is a term often used when defining reliability in higher education literature (Remmers et al. 1966; Marsh 1987; Benett 1993; Braskamp & Ory 1994; Cashin 1995; Deasy 2004). Remmers et al. (1966, p. 125) define reliability as "the consistency with which

a test yields the same results in measuring whatever it does measure". Similarly, according to Braskamp & Ory (1994, p. 91) reliability refers to the consistency and dependability of the information obtained in the measurement process: is an observation made at one point in time likely to be similar to an observation made at another? Braskamp & Ory (1994) point out of the relativity of reliability by stating that no perfectly reliable piece of information exists; increasing reliability is a matter of reducing errors in the measurement process. Remmers et al. (1966, p. 125) also note that "what a test measures may not be what it is being used to measure, i.e., a test may be invalid. But if it yields consistent results, it is reliable".

According to Popham (1981, pp. 128–129) the four approaches to reliability are stability, equivalence, equivalence and stability, and internal consistency. Another term frequently used when considering the reliability of student ratings is "interrater agreement". Marsh (1987, p. 275) claims that the reliability of student ratings is commonly determined from the results of item analyses (in other words, correlations among responses to different items designed to measure the same component of effective teaching) and from studies of interrater agreement (i.e. agreement among ratings by different students in the same class).

According to Cashin (1995), in the educational measurement literature, reliability covers consistency, stability and generalisability of items. For student rating systems, Cashin suggests that reliability refers most often to consistency or interrater agreement. According to Cashin, reliability is dependent on the number of raters: the more raters, the more reliable. Stability, he claims, is the product of agreement between raters over time, whereas generalisability is the product of how confident we can be that our data accurately reflect the instructor's general teaching effectiveness, not merely how effective a teacher has been in a particular course in a particular term.

The fact that any measurement always contains some error is well known and has been discussed in the research literature since at least the 1960s (Remmers et al. 1966, p. 8). When considering student ratings in general, Spencer & Schmelkin (2002, p. 398) state that "regarding the students, implicit in all of the literature is the assumption that they answer these anonymous instruments honestly and willingly." According to their research, Spencer & Scmelkin (2002, p. 406) claim that students are not reluctant to do ratings or fear any particular fear of bias; however, they say that students are unsure as to whether their opinions matter and for what purpose the ratings are put, and this explains why students pay little attention to the ratings. This lack of feedback for students as to how their opinions are taken into account is recognised as a major problem in many universities. Even the finest and most systematic student evaluation systems are practically useless if students are not motivated to give their opinions, which may be the case it they feel their opinions do not have an effect on anything (Kuittinen 2004, p. 17).

One important issue, which should also be mentioned in the connection with reliability, concerns the interpretation of the results of student ratings. Even if the results are perfectly reliable, they can be misunderstood by those who interpret the results. For example, Theall et al. (2000) state that informed faculty and administrator opinions are essential for good evaluation practice because "even if the raters are perfectly valid and the data completely reliable, improper or incorrect interpretation and use of data renders the entire process invalid and thus the questions of reliability and validity apply to the users of the data as much or more than to its providers."

Aleamoni (1999, p. 160) also claims that the disadvantages of gathering student ratings result primarily from how they are misinterpreted and misused. According to Adams (1997, p. 10), one of the problems with student evaluations is that few administrators are trained to interpret the ratings results. The most common misuse is to report raw numerical results and written comments on the assumption that the user is qualified to interpret such results validly (Aleamoni 1999, p. 160). For example, reporting an average by itself does not tell the reader the extent to which the responses are dispersed along the scale (Brennan & Williams 2004). As Cranton (2001, p. 11) states, student ratings of instruction are in fact subjective perceptions based on students' knowledge about good teaching: "We remain reassured by the charts, frequencies, means, and standard deviations produced by this assessment technique. We have the illusion of objectivity". It is also possible that there are even errors in the ratings results. Thus it would be extremely important that any numerical components used in the institutions meet at least minimum standards of statistical validity (e.g., Ruskai 1997).

According to Theall & Franklin (1991, p. 88) at minimum, analysis of evaluation results for teaching improvement should include descriptive information (distributions of responses by item), measures of central tendency (mean, median, standard deviation), and a direct estimate of error such as confidence intervals for means. Brennan et al. (2003, p. 35) name Cronbach's alpha as the most common measure of reliability. This estimates the internal consistency of an instrument by comparing the variance of the total scores with the variances of the scores on the constituent items (Brennan et al. 2003). Another frequently used measure is the interrater reliability of the raters to show that ratings by students are valid, reliable and consistent over time (e.g., Brennan et al. 2003). However the risks with statistical tools are that they can either be too easily misused if the principles in their usage are not properly understood or that all the relevant information possessed is not correctly taken into consideration when hypotheses are formulated (see, e.g., Dieks 1992; Toivonen 1999).

The methodologies typically used for analysing the student evaluations are factor analysis, Student's t-test, Chi Square Significance test, multivariate analysis of variance and multiple regressions (e.g., Marsh 1987; Job 2004, to name but a few). Factor analysis, as Marsh (1987, pp. 265–266) describes, provides a test of whether students are able to differentiate among different components of effective teaching and whether the empirical factors confirm the

facets that the instrument is designed to measure. There is however, a lack of consensus on factor analysis. For example Jackson et al. (1999, p. 582), have mentioned problems with factoring procedures and selecting the number and content of items, and Marsh (1987, p. 266) also notes that factor analysis cannot determine whether the obtained factors are important to the understanding of effective teaching; thus the selection of items is crucial for the validity of the evaluation. One main deficiency in using factor analyses in student evaluations relates to the fact that the collected background data has typically been based on class-averages instead of individual item specific information. What is problematic with this approach is that the class average response ignores the individual variability occurring within each class. Thus, the routine of using individual responses and class-average responses as the primary units of analysis raises problems for single level analyses. (Toland & de Ayala 2005, p. 273)

Abrami (2001, p. 73) suggests that the visual display of data can aid the interpretation of ratings results, especially for individuals lacking knowledge of statistics. A useful visual display, according to Abrami, should include the distribution of normative data, noting the norm group mean along with percentile, *z*-score, and raw score equivalents, which serve as informative points on the distribution. In addition to these normative data, the combined mean score for the faculty member and the confidence interval should be overlaid (ibid.). Theall & Franklin (2000, p. 103) also caution against the mere interpretation of numbers: understanding how ratings relate to effective teaching requires someone with a wider range of skills than either a statistician or even an expert teacher may possess. They note that the task of interpreting ratings also varies by purpose and the skill that a teacher needs to use ratings in the context of reflective practice are quite different from e.g., the performance appraisal skills that a department head needs to review ratings.

It is important not to rely too excessively on any statistical results and neglect the process by which the results are achieved. Becker (2000, p. 114) argues that departments often misuse the scores by comparing each instructor with numerical means or medians for all instructors of the course or of like courses, which endows the scores with far more precision than they actually do. One of the most problematic areas in the use of ratings, according to Franklin (2001, p. 93), is comparing raw mean scores among faculty (such as an average of 3.4 versus 3.5) or against some other kind of implicit or explicit normative standard (for example 3.0 as bad and 3.5 as acceptable). According to her, the difficulty with either approach is that the problem of error in ratings and its effect on ratings means is ignored. Scriven (1995) also warns of the common errors in data processing, report design and interpretation of the results, such as the use of simple averages without regard to the distribution, treating small differences as significant, just because they are statistically significant, or using factor analysis without logical/theoretical validation.

In addition to the above issues, it is also important to determine whether all students or only a subset of students are expected to respond to the ratings. Collecting responses from all students has traditionally been considered laborious and time-consuming. However, this is often necessary, especially if the purpose is to gather feedback at the course level where classes are often so small that the maximum number of answers is needed to form a representative overview of the responses. Limiting a survey to a subset of students may also incur a risk of sampling bias, which can arise if the relevant characteristics of the people who respond differ systematically from those who do not respond, in which case the results may be at variance with those that would have been found if responses had been obtained from the entire population (Dillman 2000; Brennan et al. 2003).

Finally, it is worth noting the warning of Ramsden (1991): one of the most common errors in the application of student evaluations is to forget that they do not constitute judgments in themselves. However valid and reliable, they can never be more than a guide to making decisions.

3.3.3 Ethical and Other Issues

Validity and reliability are important but not only criteria for student ratings. To be able to utilize the results of the ratings, the ratings have to be relevant i.e., they should have some bearing on the development of teaching and learning, be ethically implemented and be made public in an ethical way. In recent years the significance of ethical aspects has increased due to the increased role of evaluation for teaching in promotion and probation as well as the public availability of student evaluation results on universities web sites (Mc Caig 2002; McCormack et al. 2003; McCormack 2005). When discussing the ethical issues, there are two points of view particularly worth of note: the ethical aspects relating to how and what information is gathered from *students* and the ethical issues relating to the effect the evaluation results have on *lecturers*.

Zimitat & Crebert (2002, pp. 764–765), in their article on the ethics and validity of online surveys, point out that regardless of the context or nature of the research, the basic principles and standards for the conduct of research should remain the same. With regard to student evaluations, they stress that care should be taken to inform students about the survey, to ensure confidentiality and privacy and to disclose how collected data will be used and reported. Fowler (2002, pp. 147–151) lists four ethical issues which should be heeded in survey research in general: informing respondents, protecting respondents, benefits to respondents and ethical responsibilities of interviewers. The first three of these issues are valid for all student ratings, both online and paper-based. Creswell (2003, pp. 65–66)

describes the ethical issue in of two phases: data collection and data analysis. Ethical issues relating to data collection according to him include e.g.:

- developing an informed consent form for participant to sign before engaging in the research
- gaining the permission of individuals in authority to provide access to study participants at research sites
- anticipating the possibility of harmful information being disclosed during the data collection process.

In analysing and interpreting both quantitative and qualitative data, Creswell (2003, p. 66) suggests consideration of the following issues:

- how the study will protect the anonymity of individuals, roles and incidents in the project.
- storing the data after analysing for a reasonable period of time, and later discarding the data so that it will not fall into the hands of any outsiders
- the ownership of the data after it has been collected and analysed
- the need for researchers to provide an accurate account of the information in the interpretation of data.

Several problems may occur if ethical aspects are neglected. For example from the students' viewpoint a central issue is the fear of being identified from the responses (see e.g., Spencer & Schmelkin 2002), which may prevent them responding to some of the survey questions or even participating in the survey at all. The anonymity of students' responses should thus be guaranteed regardless of the way the responses are gathered. In paper-based questionnaires it is especially important to ensure that students will not be identified by their handwriting and in online surveys the student's personality must not be revealed through the login process. The anonymity of online ratings will be discussed in more detail in Chapter 4.2.3, *Other Advantages of Online Ratings.*

From the faculty point of view ethical issues often relate to the use of survey results. As Stake & Cisneros-Cohernour (2000, p. 52) remark, instructor evaluation can have several different purposes and data gathered for one purpose may unethically be used for another purpose, which may occur if, for example, data gathered for staff improvement is used for accountability. Similarly McKeachie & Kaplan (1996) emphasise that ratings on teacher characteristics can be helpful for teacher improvement, but "we fail ethically when we permit important personnel decisions to proceed on the basis of such potentially misleading data".

Vuorenmaa (2001, p. 153) and Linnakylä (2002, pp. 39–40) have examined the general motives and purposes of evaluation and present the following questions: to whom are the

evaluation results made public; who does the publicity of the results serve, when all the social effects to the participants of the evaluation are taken into account; do the evaluation results honestly bring our both strengths and weaknesses and what are the evaluands' rights to the information gathered?

Several researchers have mentioned evaluation misuse as a problem in the ethical use of student evaluations. According to Hofstetter & Alkin (2003, pp. 217–218) intentional non-use of competently made evaluations, is clearly an instance of misuse. In the same way, they deem the use of a poorly done evaluation by an informed user, who should be aware of deficiencies, as an attempt to deceive and thus misuse. However, they stress that one cannot categorise unintentional non-use of a well made evaluation to be misuse – it is simply "non use". Also according to Franklin (2001, pp. 88–91) the most common and potentially harmful types of error are misuse of data (over-reliance or conversely under-reliance on ratings data), bad data (misleading or uninformative data due to a badly constructed or incorrectly administered questionnaire or incorrectly processed, analysed or reported data), and misinterpretation of ratings e.g., errors of statistical understanding, unfounded generalisations about the characteristics of ratings and a general lack of applicable information about effective teaching practices and course design. Aleamoni (1999, p. 160) notes that the disadvantages primarily result from the way in which the ratings are misinterpreted and misused: the most common misuse being to report raw numerical results and written comments on the assumption that the user is qualified to interpret such results validly. Marsh (1987, p. 263) warns that "if a survey instrument contains ill-defined hodge-podge of different items and student ratings are summarized by an average of these items, then there is no basis for comparing these results with other findings".

Another of the most commonly cited problems with student ratings concerns the students ability to assess the education they are given. Many researchers and practitioners have questioned both the students' ability to evaluate the instruction and the instructors' pedagogical skills and contextual knowledge. For example Harvey & Green (1993) suspect that although students may be able to identify their short-term needs, they may not have enough knowledge and experience to know what they need in the long term, and thus they may not be in a position to evaluate whether their needs are met. Rautopuro & Väisänen (2000) also remark that the students' opinions about the content of their education may not be relevant if they lack the experience or ability to judge it. Adams (1997) has questioned whether students who are doing poorly in their courses are able to objectively judge their instructors. He is also sceptical that students, who are "almost universally considered as lacking in critical thinking skills", are able to critically evaluate their instructors. Becker (2000, p. 115), argues strongly the case that "students have little basis for judging an instructor's knowledge of the material, and students cannot know what goes into organizing a course if they have never taught it". According to Young (1993, p. 12) only from the later perspective of maturity and significant life experience have former students been able to

understand and appreciate the process that they have passed through. Seldin (1993) also comments that students should not even be expected to judge whether the materials used in a course are up to date or whether the instructor is familiar with the subject matter of the course. In Seldin's opinion, such judgments require a professional background and should be left to the professor's colleagues.

Sproule (2000) states that the students are too young and inexperienced to be sufficiently informed about the societal needs for educated persons and the skills demanded by employers. He also notes that the student evaluations lack "student responsibility". According to him, there are not personal consequences for a negligent, false, or even malicious representation in the process of students' evaluations of teaching. Franklin (2001) is of the same opinion on the students' limited ability to judge the course content. However, she comments that "compared with a visiting colleague untrained in systematic observation techniques who spends an hour and a half observing one class, students who spend an entire semester in a course are uniquely qualified to comment on the pace of the instruction, the classroom atmosphere, or whether difficult concepts are presented clearly".

Chen & Hoshower (2003) stress the importance of students' meaningful and active participation mentioning that students' input is the root and source of student evaluation data. According to them, the usefulness of student evaluation data is severely undermined if students do not willingly provide quality input. For their part, students' willingness to participate in evaluations is dependent on the visibility of the evaluation results. For example, Smith (1990) argues that how diligently students complete the evaluation may be a function of their knowledge of the evaluation's uses. He refers to his own survey findings which show that students do take evaluations seriously and believe in the importance and usefulness of evaluations, even though they were uncertain about whether professors take teaching evaluations seriously. Turner (2004, p. 12) expresses the same view, stating that since students seldom see changes that instructors may implement over time in response to student evaluations, they often feel that both administrators and faculty pay little attention to the feedback they provide. Smith (1990) also supports the idea of informing students of the uses of the evaluation results by stating that if students are informed that teaching evaluations are to be used for official purposes, they may be even *less* biased in their ratings. Similarly, Chen & Hoshower (2003) suggest listing prominently the uses of the teaching evaluation on the evaluation instrument. According to them, if these uses are consistent with the uses that students prefer (presuming that the students believe the evaluations are really used for such purposes), the students will assign a high value to the evaluation system. But, as Chen & Hoshower note, if students are kept ignorant of the use of student evaluations or if teaching evaluations are used for purposes that students do not value or if they see no visible results from their participatory efforts, they will cease to give meaningful input.

The increasing use of student evaluation of teaching surveys in all university courses may also prove to have its side effects. Barrie (2001, p. 10) warns of "questionnaire burnout" which may occur due to the escalating amount of questionnaires students in universities are supposed to complete. Bradley (1999, p. 387) also notes that any survey is only as representative as the subjects chosen to be interviewed. Braskamp (2000, p. 27), for his part, criticises current student ratings for focussing too much on teaching behaviour rather than student learning, arguing that if we focus only on the teacher or teaching, we will not move beyond our current practices. According to the author, many student-based evaluation procedures aimed at evaluating the teacher and teaching methods seem out of date and out of touch. Similarly, Serva & Fuller (2004, p. 21) argue that "although active learning and technology use have become acknowledged goals in many universities, many teaching evaluation instruments do not reflect changes in these areas and are rooted in the past".

There is also a risk that the use of student ratings may lead to unintended outcomes. In addition to the suspicions already mentioned that instructors may alter their teaching in order to receive high ratings or that students reward poor teaching by believing that they can give high ratings in return for high grades, Ory & Ryan (2001, p. 39) have listed the following examples:

- the campus rewards poor teaching by lowering faculty standards
- due to their convenience, the campus looks to student ratings as the only measure of teaching quality
- the content of the student rating may determine what is addressed in the classroom
- ratings are used to make discriminations between instructors that cannot be supported by the data
- due to the high stakes involved, instructors fail to follow proper administration procedures
- the rating process becomes a meaningless activity that is performed by students and instructors only because it is mandated.

Braskamp & Ory (1994, p. 145) are concerned about the motivation of faculty members observing that if normative comparisons are used to rank professors on any criterion, one half of the professors are always "below average". They continue to ask, whether this means that half of the faculty are incompetent or ineffective professors? According to them, the use of norms to interpret the effectiveness of faculty work tends to create a group of disappointed and discouraged faculty. Theall & Franklin (2000, p. 100) express concern about the increased use of student ratings data by external evaluators who lack the skills to interpret the underlying means of the numbers but whose decisions affect the entire higher education community. According to them, judging institutional effectiveness, for example by comparing average ratings would be a serious error.

Some authors have questioned the usability of any evaluation methods. Carey (1993, p. 20), who is critical of student ratings, although not totally denying their usefulness, argues that "student evaluations, even though they perform no useful function and probably militate against the goals of good teaching and the end of higher education, are the least harmful method for such evaluation". Carey (1993) adds that even though the good, bad and indifferent teachers usually come to be known as such over the years by word of mouth, evaluations of faculty cannot be based on these word-of-mouth reputations, because they are not measurable, they cannot be quantified, and they may be ill deserved. Braskamp & Ory (1994, p. 101) explain the faculty complaints about student ratings are typically prompted by "improper institutional uses, such as department rank-ordering faculty on the basis of student ratings and then using the ranks to help determine annual salaries. The complaint is not about the quality of the information but the misuse of it."

And even though most researchers support the use of student evaluations as a means of developing teaching and education, it is important to place the student evaluations to an appropriate scale. As previously quoted by Ramsden (1991) they can never be more than a guide to making decisions. McKeachie (1996) has reminded that whatever the source of data (student ratings, peer evaluation or gossip), some committee or administrator has to make an evaluative judgment, because students are not the evaluators, they simply provide data to the evaluators. According to Sproule (2000) a major conceptual problem with the student evaluations of teaching process is that opinion is misinterpreted as fact or knowledge. And finally, Theall & Franklin (2000, p. 13) also remind that even the best data can be misused, the best data collection, analysis and reporting system can be subverted and even the beast teaching can be ignored in favour of some other aspect of performance.

All in all, in spite of all the aforementioned deficiencies of student ratings and in spite of the shortcomings in the use of them (see also Parjanen 2003), they are still more widely used method for evaluating education than any other instrument (Cannon, 2001a; Saroyan & Amundsen 2001; Kember et al. 2002; Nasser & Fresko 2002). The use of student evaluations as well as the importance of evaluation results in decision making seems also to be still increasing. That is why also the possibility to analyse the validity of evaluation results becomes all more important.

4 ONLINE STUDENT EVALUATIONS

> *There are computers for planning class schedules, printing rolls, recording and issuing student grades and transcripts. There are computers for assisting, teaching, entertaining, tracking, retracking, and side-tracking students – there are computers everywhere doing almost anything.*
>
> – Pedersen 1983 –

As recently as the beginning of the 1990s, online surveys were a rarity in higher education as well as other areas. The first research reports on electronic survey methodologies in higher education were published in the late 1980s (Layne et al. 1999). The lack of computers and Internet connections both in and outside campus areas was one of the major factors delaying the widespread use of online surveys and therefore the research into electronic surveys in higher education. Even at the start of the new millennium Hmieleski & Champagne (2000) wrote that only 2 percent of the colleges Hmieleski surveyed in the USA reported using the Web as the predominant approach to their course evaluation. Since then the use of information technology has grown significantly and during the last few years there has also been an increase in research on online ratings. The first Internet tool used for survey research was email (Truell 2003). Now there are several methodologies for conducting online research (e.g., E-mail, bulletin boards, Web HTML, Web fixed-form interactive, Web customized interactive, downloadable surveys and Web-moderated or chat interviewing). Until a few years ago email surveys were the main way of conducting Internet surveying, but as the WWW has grown in popularity, the use of Hypertext Markup Language (HTML) forms of Web-based surveys have become the dominant method of gathering survey data. (MacElroy 1999; Solomon 2001)

This chapter introduces the present state of knowledge in the use, exploitation and research of online student evaluations and the characteristics of implemented online rating systems, the advantages and limitations of online student ratings as well as the major differences between online and paper-based surveys. Finally, this chapter discusses gaps in the current knowledge and the position of this research as well as the direction and suggested methods for further research.

4.1 Present State of Knowledge in the Use and Research of Online Student Evaluations

Shannon et al (2002) have described the development of electronic surveys starting from early disk-by-mail formats and simple e-mail surveys and evolving to current Web-based surveys. Possibly the first electronic survey in university surroundings was a survey on health and personal characteristics at Carnegie-Mellon University conducted in 1983 by Kiesler & Sproull (1986). The survey was a computer program which the respondents invoked at their own terminals and the purpose of researchers was to identify the differences in results caused by the medium in which the survey was administered. Since then modern technology has developed rapidly and many universities throughout the world have adopted new technologies for assessment and evaluation use. Hoffman (2003) reported in his study that ten percent of U.S. colleges were using a campus-wide Internet system as a primary means for collecting student-ratings data at their institutions, which represents an increase of around 8 percent over the year 2000. Furthermore, seventeen percent of the responding institutions Hoffman studied reported using the Internet in some capacity to collect student evaluation data for face-to-face courses and 56 percent of the respondents reported they were using the Internet for evaluating online courses.

In recent years there has been an enormous growth in the use of Web-based student evaluations (typically called as online ratings) in higher education, as well as in other surroundings. The increase in use is naturally related to the increased use of technology in education in general, and some researchers (e.g., Oliver & Sautter 2005) also regard the increase in use as attributable to the advent of online or distance learning technologies. In many universities the Web-based ratings were first taken into use within online courses, of which the evaluation had previously been carried out by mailing questionnaires to students and expecting them to complete and mail them back to the university (see, e.g., Bullock 2003). Traditional evaluation methods were, however, problematic to execute in online courses. For example, Theall & Franklin (2000, pp. 101–102) have stated that student ratings of distance education collected with traditional [paper-and-pencil based] instruments do not address the unique characteristics of the on-line teaching-learning situation and thus do not provide data specific enough to allow accurate understanding of the outcomes of instruction. Although the use of online ratings has become more common in all kinds of courses, the results of Hoffman's (2003) study showed that significantly more institutions were still using the Internet for student ratings of online courses than for face-to-face courses.

Sometimes the new online ratings have replaced traditional paper-based questionnaires throughout the university, sometimes they have been experiments directed at certain courses or other entities for research purposes. What is surprising in the reports of the use of Internet-based student evaluations, however, is that typically when universities have implemented online questionnaires, the other advantages of information technology have been neglected:

the online surveys contain often just the traditional paper-based items transferred to an electric form (e.g., OnSET 2005). Similarly, although the amount of research has grown during the last few years, there still seems to be a lack of research into and experience of online ratings with any more than just the basic elements from traditional survey forms. Recognising this, Cummings et al. (2001, p. 29) comment that although online surveys run in a number of universities, few research papers have been produced in this area.

Nowadays online ratings are used in thousands of universities worldwide. Several universities have also reported their intention to start using online evaluations in the near future.[19] In U.S.A. they are typically used in the same way as traditional paper-based evaluations, to evaluate courses and instructors whereas in Australia many universities are using not only online course evaluation surveys but also extensive online systems for graduate students (e.g., Cummings et al. 2001). In Europe there seems to be no coherent way in which universities use online ratings, instead the existing online systems have been executed to fulfil the needs of the individual institute.

In addition to the universities' own web-based student evaluations of faculty performance, there are web sites created and managed by students that are not approved by the university's administration, and there are even corporate websites featuring student evaluations (Mc Caig 2002). It is expected that the number of students accessing these teaching evaluations will increase in the near future, even though some universities in U.S.A. have attempted to restrict student access to the evaluation results (Wilhelm 2004, p. 17).

Research into online ratings, especially in the early years of their adoption in universities, has concentrated on giving descriptions of implementing online rating systems (e.g., Andrews & Feinberg 1999; Goodman & Campbell 1999; Scoles 2000) or on descriptions of implemented systems as such (e.g., McGourty et al. 2001; Ballantyne 2003; Llewellyn 2003; Tucker et al. 2003) or descriptions of experiences from implementing online systems (e.g., Recker & Greenwood 1995; Ha et al. 1998; Cummings & Ballantyne 1999; Cody 1999; Nulty 2001; Moss & Hendry 2002; Watt et al. 2002).

Most current research literature appears to concentrate on assessing the differences between online surveys and traditional paper-based questionnaires. One of the most popular research themes is the potential difference between response rates of online ratings and traditional paper-based questionnaires (Layne et al. 1999; Tomsic et al. 2000; Sax et al. 2001; Dommeyer et al. 2002a; Dommeyer et al. 2003; Sax et al. 2003; Dommeyer et al. 2004). Though the research results show generally lower response rates for online ratings than traditional methods, there is also evidence to the contrary. There are also studies investigating strategies for improving response rates (see, e.g., Cummings et al. 2001; Dommeyer et al.

[19] For example in a seminar "Opiskelija opetuksen laadunarvioinnissa" held by FINHEEC in 23rd January 2006 several universities reported their intention to start collecting feedback from students online in the near future.

2003; Porter & Whitcomb 2003; Conn & Norris 2004). The differences in response rates and the potential explanations for the differences will be more thoroughly discussed later in this chapter.

Other comparisons between online and paper surveys of interest to researchers have dealt with such topics as the costs of online ratings compared to traditional paper-and-pencil ratings (e.g., Bothell & Henderson 2003), the differences between student responses to online and paper-based surveys (Crawford et al. 2002; Carini et al. 2003; Hardy 2003; McGhee & Lowell 2003; Dommeyer et al. 2004; Gamliel & Davidovitz 2005) and possible biases and demographic differences between respondents (e.g., Tomsic et al. 2000; Sax et a. 2001; Ballantyne 2004). There are also studies investigating whether the design of a Web questionnaire influences response rates or the data quality (Couper et al. 2001; Forsman & Varedian 2002; Heerwegh & Loosveldt 2002). Some comparisons made in university settings have not compared student but rather teacher (Mertler 2003) or alumni responses (Muffo et al. 2003) to online and traditional paper-based surveys. There are also studies examining the extent to which universities have implemented online student evaluations (Sorenson & Reiner 2003) or have adopted the Internet for data collection and reporting of student evaluations of instruction (Hoffman 2003) or examining the roles of technology in teaching and learning (Driscoll 2001; Hoffman 2003).

Although the scientists have recognised both the possibility to (e.g., Kelly & Marsh 1999) and the need to link data systematically from students and others to available information (Welsh et al. 2001, p. 393), there are very few studies that address the potential for connecting survey data to other university-held databases.[20] There are some research articles describing the use of online technology in developing advanced computer-supported learning systems such as CECIL (Gardner et al. 2002) or developing a methodology for online feedback and assessment (Hanson et al. 2001), but integrated student online evaluation systems are a rarity. In an alumni study by Romano & Himmelmann (2002) examining the factors affecting whether a student decides to answer a survey via Internet or via paper the online response mode was connected to background variables but the actual answers collected online were not analysed. Welsh et al. (2001) developed a continuous quality measurement system that links student satisfaction data with other information available about the university. However, in their study the surveys were paper-based questionnaires mailed to respondents and the responses, which included the identification number of the student, were later scanned to computer. Theall & Franklin (2001b) describe the experiences of using TCE-Tools, a system for formative and summative evaluations which is capable of utilizing integration of data sets and supporting individual questionnaires and delivering Web-based online course evaluations. They make no mention, however, of integrating university databases to the results of the survey itself. Sax et al. (2003) executed a survey that

[20] As early as 1999, a small-scale experiment utilising the connections between online survey and the university's database had been conducted in TUT (see Pajarre 1997).

studied response rates and non-responsive bias by combining the survey results with the background information of the same students which they had previously gathered from the students in another questionnaire.

Possibly the only published research utilising the opportunities of combining online results with existing databases is Thorpe's (2002) investigation of non-response bias in which he studied three classes instructed by the same teacher. The purpose of his study was to examine whether significant differences exist in student responses on the basis of the delivery method (in-class paper versus web-based). The online part of the study was executed so that the students logged in by entering their university ID number (which was later connected to university's database) and answered questions relating to a particular course and teacher. The demographic differences among students were partly analysed by comparing paper-based responses to online responses and partly by comparing online responses to the particular course to four background characteristics gathered from the university database. The results of his study suggest that the evaluation responses did not significantly differ between online and paper-based responses. Since Thorpe's research is confined to a single course level, so far there still exists no extensive research combining the results of an online survey with the information about the students in university databases and utilising the results to their full potential.

4.2 Advantages of Online Student Evaluations

There are several proponents of Internet-based ratings system who consider online ratings to be "cutting-edge" practice compared to traditional paper-based systems (Sorenson & Reiner 2003) and claim, for example, that "with the current trends toward electronically mediated instruction, the delivery of course evaluations via a web-based application is mandatory" (Conn & Norris 2004). Regardless of the method of teaching, student ratings systems in universities are increasingly being transferred to Internet surroundings. The advantages most often mentioned are that the development of modern technology has lowered costs and enabled more timely feedback. Other frequently mentioned advantages are, for example, more thoughtful comments from students and fewer errors in data entry.

4.2.1 Lowered Costs

Numerous researchers report reduced costs as being one of the main advantages of online ratings over traditional paper-based surveys (e.g., Schmidt 1997; Dillman 2000; Cummings et al. 2001; Forsman & Varedian 2002; Moss & Hendry 2002; Thorpe 2002; Johnson 2003;

Muffo et al. 2003; Sax et al. 2003; Sorenson & Reiner 2003; Ballantyne 2004). Some researchers even claim the reduced costs to be the primary reason for putting student evaluations online (Bothell & Henderson 2003, p. 70). The reduced costs are not only actual direct monetary savings (e.g., reduced paper, mailing or data entry costs) but also savings in time otherwise lost from contact teaching to the administration of evaluations, savings in faculty time to administer evaluations or in the time taken for departmental assistants to manage the evaluation process (Dillman 2000; CETL 2004).

There are several research reports discussing the differences in costs between traditional and online ratings. Hmieleski and Champagne (2000) examined several surveys and conclude that the costs of online evaluations were consistently significantly lower than the costs of traditional evaluation methods. The degree of savings, on the other hand, varies considerably from modest to remarkable. For example, Forsman and Varedian (2002) found the difference in total costs between the Web and mail survey around 20 percent in favour of the Web alternative. Bothell & Henderson (2003, p. 76) calculated the annual costs of online ratings to be approximately half of the costs of traditional paper-based ratings. In a study by Crawford et al. (2002), the mail survey actually cost 222 % more than the web survey.

However, there are also opinions which contradict the above comparison results. For example, Theall (2000) questions the economic efficiency of online ratings, claiming that the figures provided by Hmieleski and Champagne (2000) present the best case for electronic data processing and the worst case for paper-based systems. And though most researchers agree on the reduced costs of online systems, these authors point out the less obvious additional costs of initiating, developing and financing such systems (see, e.g., Upcraft & Wortman 2000; Sax et al. 2001; Sorenson & Reiner 2003; Brennan & Williams 2004). According to Ha & Marsh (1998), it is estimated that an online system needs to be updated once every 3 to 5 years, the cost of which should also be taken into account. Initial costs can sometimes be lowered by purchasing and possibly modifying an existing online rating system (Ha & Marsh 1998; McGourty et al. 2002b) if it satisfies the university's needs. Sorenson & Reiner (2003, p. 9) also note that on many campuses the old paper-based systems have become outdated, making it necessary to invest in a new system, whether paper-based or online. Some researchers have also observed that the sheer ability to collect "thousands of responses at no more cost than collecting dozens of paper-based questionnaires has enticed a culture of web surveys to develop which tends to ignore the scientific underpinning of surveys" (Dillman et al. 1998).

4.2.2 Quicker Turnaround Time

One of the most often commonly cited advantages of moving student evaluations online is the quicker turnaround time (Layne et al. 1999; Dillman 2000; Upcraft & Wortman 2000; Cummings et al. 2001; Theall & Franklin 2001b; McGourty et al. 2002b; Moss & Hendry 2002; Hardy 2003; Johnson 2003; Sorenson & Reiner 2003; Brennan & Williams 2004, CETL 2004, Conn & Norris 2004). Traditional paper-based surveys are usually very time consuming by their very nature; it requires time to distribute, collect and manage paper-based surveys, to enter data from the questionnaires either manually or by using a scanning hardware and to process and analyse information (e.g., Ballantyne 2000; Ingham 2000; Conn & Norris 2004). In an online system the data entry is eliminated, the data acquired is free from common entry errors and the results are almost immediately available for the analysis and reporting (Layne et al. 1999; McGourty et al. 2002b; Bullock 2003; Muffo et al. 2003).

On the other hand, Theall (2000) is sceptical of the supposed efficiency of online ratings by pointing out that putting student ratings online will not improve the evaluation practice as such. According to him, moving paper-based evaluations online will only allow bad information to be misinterpreted and misused more rapidly by those who presently do so in paper-based systems. Upcraft & Wortman (2000) also note that the efficiency of an online system may be severely compromised if there are malfunctions in hardware, software or the server or human errors in programming or storing data should occur.

From the students' perspective, there are two different opinions concerning the efficiency of responding to an online survey. On one hand, online ratings are considered as a quick and easy way for students to respond to the survey (Cody 1999; Cummings & Ballantyne 1999; Ingham 2000; Upcraft & Wortman 2000; Dommeyer et al. 2002b). It is also agreed by many researchers that since responding to an online rating is independent of time and place (although the answering requires an online connected computer), students have more time to complete the ratings forms (Layne et al. 1999; McGourty et al. 2001; Sorenson & Reiner 2003). On the other hand, the lack of Internet-connected computers, complicated log-on processes and other time-consuming computer problems are sometimes recognised as a problem (Dommeyer et al. 2002b).

4.2.3 Other Advantages of Online Ratings

There are numerous other advantages of online ratings over paper-based versions presented in literature, in addition to reduced costs and more timely feedback. Such reports present the perspectives of students' and teachers' as well as faculty and administrators. For the students, the most often mentioned advantages relate to the increased anonymity in written responses compared to traditional handwritten comments (Cummings & Ballantyne 1999; Ballantyne

2000; Cummings et al. 2001; Hardy 2003) and to the increased freedom to select when to respond to the survey (Ingham 2000; Scoles 2000; Cummings et al. 2001; McGourty et al. 2001; Thorpe 2002; Sax et al. 2003). The commonly cited advantages for teachers include among others the savings in class time previously taken up with filling in the questionnaires at the end of the course allowing for more time to be spent on instruction (Kelly & Marsh 1999; Ballantyne 2000; Ha et al. 2000; Cummings et al. 2001; McGourty et al. 2001; Thorpe 2002; Hardy 2003; Sorenson & Reiner 2003; Turner 2004), longer and more thoughtful responses to open-ended questions (Hmieleski & Champagne 2000; Scoles 2000; McGourty et al. 2002b; Bullock 2003; Johnson 2003; Sorenson & Reiner 2003; Conn & Norris 2004; Turner 2004) and the possibility to modify the questionnaire or to add customised questions according to teacher's personal needs (Kelly & Marsh 1999; Ha et al. 2000; McGourty et al. 2002a; Sorenson & Reiner 2003).

From the administrative perspective, online ratings are seen as advantageous not only because of reduced cost in materials and reduced data entry, but also because of ease of administration (Kelly & Marsh 1999; Johnson 2003; Sorenson & Reiner 2003). Other benefits reported are fewer errors in data entry and more accurate data collection (Schmidt 1997; Ballantyne 2000; Cummings et al. 2001; Johnson 2003; Sorenson & Reiner 2003) and the opportunity to present data in a more user-friendly format than has previously been conceivable (Schmidt 1997; Johnson 2003). The possibility of accessing a large population simultaneously is also considered as an advantage for administrators (Schmidt 1997; Upcraft & Wortman 2000). Oliver & Sautter (2005) also emphasize the ease of downloading the ratings data into a statistical analysis framework for a more informed analysis of student responses which, in turn, enables closer scrutiny of the bias effects.

Even though there are several advantages of online ratings that are attributed to the development of modern technology, the potential for further developing online ratings by utilising the information technology are largely unreported in the literature. The most typical solution is that in making the conversion from paper-based questionnaires to an online survey, only the process – from paper to online – is changed and the other constituents are left unchanged (see, e.g., Llewellyn 2003).

Brennan et al. (2003, p. ii) make the constructive observation that in many institutions, more use could be made of feedback data. They advocate collecting less data and analysing and presenting it more thoroughly and imaginatively, which, according to them, would increase the commitment of faculty and students to the importance of the feedback process. What is dismissed by them and most other researchers is the possibility of utilizing the existing computer network and databases in the online ratings process. For example, several researchers have suggested that the length of a questionnaire affects the students' eagerness to respond to a survey: the longer the survey, the lower the response frequency. Linking the online rating to university's database makes it possible to omit several questions concerning

the students' background information from the online questionnaire and thereby reduce the time and inconvenience for students to respond to the survey. This possibility has not so far been utilised. Nonetheless, a reduction in the number of questions is only one minor factor in terms of the overall benefit to be gained by linking the online survey and student registers together.

4.3 Prevailing Views on Limitations of Online Ratings

The fact that the shift from paper-based questionnaires to online ratings has not occurred as fast as might be expected[21] may be the result of actual and anticipated limitations associated with online rating systems. The most commonly mentioned drawbacks with and reservations to the use of online ratings relate to potentially lowered response rates, demographic differences among respondents, possible biases and the reliability of online ratings.

4.3.1 Response Rates

Online student course evaluations have been criticized for their potentially low student response and the likelihood of non-response bias (Thorpe 2002, p. 1). However, there is no generally agreed level of what constitutes a reasonable response rate since it partly depends on what other information sources are available to test the accuracy of the survey data and on the absolute number of responses obtained. For example, a 50 per cent response rate from a class of 100 students will provide more useful information than a 50 percent rate from a very small class. (Brennan & Williams 2004)

The research results of response rates of online ratings in university settings are somewhat contradictory. Although Internet-based surveys are frequently associated with lowered response rates in academic studies, the research results show both increases and decreases in the response rates when compared to traditional paper-and-pencil questionnaires. For example Scoles (2000) reported response rates from online surveys as low as 10 % and several other researchers have presented research results from online surveys with response rates below 50 percent (e.g., Cody 1999; Cummings & Ballantyne 1999; Dommeyer et al. 2003; Muffo et al. 2003). In contrast, there are findings showing response rates as high as 75 percent or more (e.g., Ha et al. 2000; McGourty et al. 2002a; Sorenson & Reiner 2003).

[21] For example in USA according to Hoffman's (2003) study, only 17 % of institutions responding to his survey reported using the Internet to collect student evaluation data of face-to-face courses.

The raw figures for response rates do not in themselves reveal much about any single evaluation method. In order to judge whether the response rates gathered in a particular way are adequate, it is important to compare the rates with the rates gained by some other method. For example, a study by Watt et al. (2002) reports response rates of 32.6 % for an electronic survey compared to a rate of 33.3 % for a paper-based version, indicating no significant difference between evaluation methods. To date, research results have shown greater response rates sometimes for online and sometimes for paper-based surveys, varying from one study to another. For example, the studies of Cody (1999), Kelly & Marsh (1999), Cummings et al. (2001) and Dommeyer et al. (2003) showed lower response rates for online evaluations than for traditional paper-based evaluations in contrast to the findings of studies such as those of Ha et al. (2000), Crawford et al. (2002) and Thorpe (2002). In addition to these contradictory results there are also studies of Recker & Greenwood (1995), Forsman & Varedian (2002), Muffo et al. (2003), Liegle & McDonald (2004) and Oliver & Sautter (2005) showing no significant difference between online and traditional paper-based rating methods.

In many universities the major concern with online administration of student evaluations are response rates (see, e.g., CETL 2004; Conn & Norris 2004). There are a variety of limitations identified in the research literature to explain the low response rates of online student evaluations. The most common of these are multiple and frequently changing email addresses, slow or uncertain Internet connections, students' technological illiteracy, difficulty in accessing available computers in the campus area, insufficient information on the online survey for students, fear of a lack of anonymity in responses, lack of compulsion to complete the survey and student apathy (e.g., Bradley 1999; Rigsby & Smith 1999; Ballantyne 2000; Dommeyer et al. 2002b). Some of these concerns, it is claimed, will eventually disappear due to increased computer literacy and the development of Internet technology and networks (Tomsic et al. 2000; Dommeyer et al. 2002b).

Brennan et al. (2003, p. 29) have analysed the attitudes among researchers and practitioners to response rates, noting that some may argue that the purpose of feedback is simply to provide students with an opportunity to comment on their educational experience. Thus, students who do not respond present no problem because they have chosen not to contribute to this exercise. Nevertheless, according to them, most researchers assume that the purpose of feedback surveys is to investigate the experience of all the students being questioned. In such a case nonresponses do pose a problem since any conclusions have to be based on data contributed by a sample. (Brennan et al. 2003)

The researchers have considered several ways for increasing the response rates. The means vary considerably, from increasing information of the online survey or sending remainder emails to students (Layne et al. 1999; McGourty et al. 2002a; Conn & Norris 2004) to the use of incentives (Dommeyer et al. 2002b; Ballantyne 2003; Johnson 2003) or limiting students'

access to essential information until they have logged on to the online rating system (Johnson 2003). The most important reason for attempting to increase response rates is the fear that lower response rates are biased in some way and thus may compromise the usefulness of information gathered from such evaluations. However, the increase in response rates as such is not necessary always positive. For example, Hmieleski & Champagne (2000) argue that the value of an evaluation may be lost if return rates become the primary goal of evaluation. Their experience is that by making online evaluations mandatory, the response rates will rise near 100% but the number of useful open-ended comments will drop dramatically. Several scholars have studied the demographic differences of respondents and possible sources of bias as well as the effects of using incentives. Their findings are discussed in the following sections.

However, as Bosnjak & Tuten (2001) remark, while the potential bias that may result as a consequence of nonresponse is a well-covered topic in 'classic' survey modes (e.g., mail surveys), there is little explanation of nonresponse itself and relatively little is still known about response behaviours, especially in the case of Web-based surveys. The problematic issue of nonresponse is discussed more thoroughly, e.g., by Schnell (2002).

4.3.2 Demographic Differences among Respondents and Nonrespondents

Several researchers have attempted to explain the underlying factors affecting student's willingness to respond to an online survey. Their findings, however, are valid only for a limited time due to the rapid increase both in the use of Internet and in the network connections. For example, the findings of both Layne et al. (1999) and Tomsic et al. (2000) showed that younger students were more likely to respond to a Web survey than older students and that students enrolled in the area of technology or science were significantly more likely to respond to the online survey than students from other disciplines. These findings are likely to become outdated with the rapid growth in Internet use in many surroundings both within and outside university settings. As early as 1996, Kehoe & Pitkow observed that the early users of the Web were "primarily young, computer savvy users" and the "early adopters of technology" whereas with the expansion of the Web to include new user groups, the respondents to online surveys will increasingly come from more diverse segments of the population. The predicted saturation of Internet technology, possible side effects in the form of spam and other undesirable factors will all probably change users' attitudes towards the use of Internet surveys (e.g., Cho & LaRose 1999). It is also worth noting that the findings of Sheehan (2002) show that response rates to online surveys in general have already been declining in the past ten years.

The attempts to reveal demographic differences among respondents and nonrespondents are numerous and sometimes contradictory. The primary reason prompting the interest in revealing possible demographic differences among respondents and nonrespondents is the fear that low response rates to online ratings may bias the gathered responses (see, e.g., Sorenson & Reiner 2003). Even though it is claimed that low response rates alone do not necessarily suggest bias (Sax et al. 2001, p. 6) the concern of the possible biases is probably one reason preventing universities from moving their surveys online. Zimitat & Crebert (2002, pp. 766–768) like Sax et al. (2001, p. 411) discuss nonresponse error, which occurs when respondent's answers to the survey differ from those of non-respondents causing an outcome different from that which could have occurred had the non-respondents participated. According to Zimitat & Crebert (2002) non-response errors are not equivalent to response rate, but they are less likely to occur with a high response rate. Sax et al. (2003, p. 411) also note that nonresponse bias should not be confused with response bias since the latter typically indicates bias in the ways in which the questions themselves are answered, i.e., nonresponse bias is linked to *who* responds and response bias to *how* they respond.

Several studies investigate the effects of respondents' and nonrespondents' gender, age or grade point average. Many of the research studies have been executed by administering both an online and a paper-based survey and comparing the results of those surveys statistically. The background information is usually detected either by asking the students in the survey questionnaire or asking the students for their student number and later linking it to the student background data (e.g., Thorpe 2002). In addition to the findings that younger students and science students are more likely to answer online than other students, as mentioned above, there are research results suggesting that students with higher grade point averages are more likely to complete an electronic survey than students with low grade-point averages (Layne et al. 1999; Thorpe 2002; McGourty et al. 2002b). It has also been shown that female students were more likely to respond to an online survey than men (Sax et al. 2001; Thorpe 2002; McGourty et al. 2002b). The studies of McGourty et al. (2002b) show that juniors and seniors were more likely to respond than lower classmen. In contrast, the results of Forsman & Varedian (2002) and those of Ballantyne (2004) show that male students were more likely to respond online than women and that no differences exist between whether they were in the first or subsequent year of study. The results of research by Carini at al. (2003) support the latter findings in that they found no differences in mode effect by age of students or differences in respondents' gender. The studies of Ha & Marsh (1998), Cummings & Ballantyne (1999) and Muffo (2003) showed no significant difference between online and paper-based ratings. As Muffo (2003) puts it; whatever biases exist in the data are similar to those found in the paper-and-pencil surveys done previously.

There are assumptions that online evaluations will result in lower ratings than paper-based evaluations, especially if the response rate is low (CETL 2004). Layne et al. (1999) and Kelly & Marsh (1999) studied the potential differences in the way students respond to different

rating forms. They both found that students rated courses more favourably when they responded online than they did in paper-based evaluations. This conclusion, however, was later questioned by Hardy (2003), whose research results showed slightly lower scores (0.25 points) for online ratings than for paper-based evaluations. Johnson (2003, 55), reported response rates 0.1 points higher for online evaluations, observing that the online evaluations were less sensitive to response rate than were paper-based evaluations.

Liegle & McDonald (2004, p. 6) also tested, how "good"/"bad" students ranked instructors online in anticipation of their respective grade.[22] There was no difference between good students or bad students regarding the means of evaluation, i.e., for both online and paper, good students ranked faculty higher than bad students.

Kelly & Marsh (1999) searched for potential differences in responses according to response mode, noting that students who responded to open-ended questions online made more comments than those who responded in paper-based evaluations. They suggest that the tendency for students to rate courses more favourably online may partly be explained by the lower response rates for the online groups: if the students who answer online are more motivated than other students they may also be the ones with a more positive view of the subject of the evaluation. On the other hand, Kelly & Marsh noticed also that students in the randomly generated access code made significantly more comments than students who were required to use their ID to login. Students in ID mode appeared to be more inhibited about expressing both negative and positive feelings in writing. This supported Kelly's and Marsh's assumption that students' perception of anonymity is critical to the establishment of a valid and reliable system. Similarly, Oliver & Sautter (2005) argue that the guarantee of anonymity was likely a significant factor in the increase in response rates for online submission.

The tendency of producing longer and more thoughtful comments on the Web surveys has been noted in several research papers. For example Oliver & Sautter (2005) reported an average of 28,97 words of feedback provided per question in response to the open-ended questions in online administered ratings compared to an average of 7,83 words of feedback via traditionally administered paper-based ratings. Hmieleski & Champagne (2000) and Johnson (2003) assume the reason for longer and more thoughtful comments is because students can respond to online surveys at their leisure instead of during normally limited class time. According to Svinicki (2001) student feedback is too often solicited as an afterthought during the last few minutes of the class when students (and instructor) are already focussing on their next appointment rather than on the survey. Layne et al. (1999, p. 229) also found that students who completed the survey electronically were much more likely to provide comments about their course and instructor than were students in the paper-and-pencil group. A study by Ballantyne (2004), on the other hand, showed the percentage of respondents

[22] The dataset Liegle & McDonald (2004) employed did not allow for individual grade analysis. Instead they determined a multi-year average of what constituted a course consisting of primarily "good" students.

making comments was smaller in the online group. However the comments received online when measured in words per comment, characters per comment and lines per comment, showed these students' responses to be more prolific. According to Ballantyne, it would appear that the theory that more students make online comments does not hold but that the comments are more prolific does.

4.3.3 Other Possible Sources of Bias and Other Limitations

In consequence to the vast increase in Web surveys in general, researchers have identified a number of severe methodological problems, such as coverage, sampling, nonresponse, and estimation problems (see, e.g. Couper 2000; Forsman & Varedian 2002; Manfreda et al. 2002; Sheehan 2002). Potential problems due to the nonresponse bias were introduced in the previous chapter. The coverage bias[23] that is of most concern to researchers of Web surveys in general (Couper 2000; Solomon 2001) is usually of no concern in universities, where nearly all students have Internet access and are used to Internet registration for courses and submission of assignments, and where their email addresses are usually up to date (cf. Dillman & Bowker 2001; Crawford et al. 2002; Forsman & Varedian 2002; Zimitat & Crebert 2002).

According to Couper (2000, p. 475) relatively little attention has been paid to the problem of measurement error, which is the deviation of the answers of respondents from their true values on the measure. Measurement errors in self-administered surveys could be due to the respondent (lack of motivation, comprehension problems, deliberate distortion, etc.) or due to the instrument (poor wording or design, technical flaws, etc.) In order to minimise respondent error, the survey instrument must be easy to understand and complete, must be designed to keep respondents motivated to provide optimal answers, and must serve to reassure respondents on the confidentiality of their responses. (Couper 2000)

The effects of using incentives to increase response rates have aroused debate among researchers in the field. Several studies report increased response rates after taking incentives into use (Cummings et al. 2001; McGourty et al. 2002a; Dommeyer et al. 2004). Dommeyer et al. (2004) also show that online evaluations do not produce significantly different mean evaluation scores than traditional in-class evaluation, even when the incentives are used to increase response rates among students.

[23] According to Couper (2000, p. 467) coverage error is a function of the mismatch between the target population and the frame population. Sampling error arises from the fact that not all members of the frame population are measured. If the selection process were repeated, a slightly different set of sample persons would be obtained. (Couper 2000)

The need for online rating systems to provide information that is both valid and reliable is still crucial and is becoming ever more so with the growing use of evaluation results in promotion and tenure decisions. Although much has been researched and written about the validity and reliability of paper ratings on instruction, little has been researched or written in this regard about online ratings of instruction. (Ballantyne 2003, p. 108) For example the question presented by Moss & Hendry (2002, p. 588), about whether students who have very good or very bad course experiences are more likely to feel compelled to respond than those whose experiences have been unremarkable, remains unsolved. So far the research that has utilised both online ratings and background data has concentrated only on single courses. No research containing either larger entities or longer time periods has been conducted into online ratings with the use of background data.

4.4 Additional Differences between Online and Paper-Based Evaluations

In most universities the shift from paper-based evaluations to online ratings has been executed so that the questionnaires have remained unchanged, with only a change in the response medium (OnSET 2005). Despite the fact that the actual questionnaire has remained the same in most cases, there are many other differences in the process of gathering information from students online compared to the traditional paper-and-pencil method. These differences include, among others, changes in where and when students respond to the survey, changes in the visual appearance of the survey and changes in the way the results are published and distributed.

The change in response time and place has had two consequences. On the one hand, students are free to respond to surveys in their own time and place of choice, whereas the traditional paper-based questionnaires are usually collected during the last lecture before the exam. When filling in paper-based questionnaires students typically sit in classrooms under supervision in the company of their classmates. In contrast, when using an online system students are free to respond the survey anywhere, in any kind of groupings and probably without supervision (Ha et al. 2000; McGourty et al. 2002a; Sax et al. 2003). However, Moss & Hendry (2002, p. 585) point out that web-based surveys cannot be easily put aside to be completed in a different location at a user's inclination, and if a respondent wants to complete a survey at a later time, it is more difficult to retrieve the survey again.

With regard to the issue of time, an argument in favour of online ratings relates to the possible deceit on the part of the teachers. For example Griffin (2001) presents the possibility that to increase their scores, teachers can manipulate timing and procedures for student evaluations, for example, by driving "the unhappy out of the class". By placing the

questionnaire in the Web, it is impossible for teachers to control who responds and who does not, making the results received are more objective in this respect than those collected in the classroom. The decrease in response rates discussed in the previous chapters is largely because collecting traditional paper-based evaluations during the lecture times under supervision makes it relatively easy to achieve high return rates. The quality of ratings results achieved under such conditions, however, has been of the subject of great concern and much research for several decades. The validity and reliability of online result because of the suspicion that only certain types of student answer online has also interested several scholars in the field.

The importance of the visual appearance of the response medium is emphasized when the questionnaires are filled in online. Not only is there a risk that the layout designed for certain software and hardware will either look different or even fail to function at all (e.g., Couper et al. 2001; Dillman & Bowker 2001; Watt et al. 2002, Presser et al. 2004) but there are also the effects of different layouts or appearance of questionnaires on the students' responses in the focus of research. For example, in his dissertation Turner (2004) investigated the influence of the response formats, such as radio-buttons, drop-down menus or text boxes on the responses given in online student evaluations. Unlike other types of survey, Web page design skills and computer programming expertise play an important role in the design of online surveys (Gunn 2002). According to Presser et al. (2004, p. 122) online ratings require testing of aspects unique to that mode, such as respondents' monitor display properties, the presence of browser plug-ins, and features of the hosting platform that define the survey organisation's server.

Couper (2002) observes that good research takes time and effort. He notes that not only is the sheer number of [web] surveys increasing, but there is also increasing diversity in the quality of surveys being conducted. As Theall & Franklin (2000, p. 101) comment, relatively little expertise is required to collect ratings data; practically anyone can create a questionnaire or administer it on-line. Thus, as Sax et al. (2003) point out, closer scrutiny of the validity and usefulness of online data collection is critical as Web-based surveys proliferate. Similarly Shannon et al. (2002) also mention the potential problems arising from the fact that developing electronic surveys requires technological knowledge and skills with the result that the lead in developing online surveys has largely been taken by technology specialists. According to Shannon et al., before electronic surveys[24] are widely accepted and used on a regular basis, input must be gathered from survey methodology professionals. Dommeyer et al. (2002a, p. 456) also consider that since online procedures require little, if any, involvement of faculty, there is less scope for faculty to influence or alter the [ratings] results.

[24] The authors' article discusses all electronic surveys, not only those used in higher education surroundings.

As shortly mentioned in Chapter 4.3.2, at least two kinds of problem have been identified as arising from the vast growth in electronic communication. First, Ha et al. (2000) warn of "questionnaire burnout" which may accompany the increase in student surveys. Second, Shannon et al. (2002) and Sheehan (2002) highlight the concerns of several survey researchers that the email-based invitations to respond to Web-based surveys or mass mailings to published email lists might be perceived as junk mail or "spam". According to Sheehan, it seems reasonable to assume that the amount of unsolicited email correlates negatively with willingness to participate in email surveys. Sheehan (2001; 2002) also warns of the potentially lowered response rates. According to Sheehan (2001), while the number of studies that use e-mail to collect data has been increasing over the past fifteen years [in general], the average response rate to the surveys appears to be decreasing. According to the author, this early level of high response may reflect the novelty value of using e-mail to respond to surveys and thus, as time progresses, it seems likely that response rates to e-mail surveys will continue to decrease.

The other issue concerning online ratings is the anonymity of responding students. On one hand, many researchers claim that students write longer open-ended comments since they can not be identified by their handwriting (Hardy 2003), on the other hand, some researchers (e.g., Conn 2003) report that students are wary of revealing their identity through the log-in process which requires entering the student number or other form of identification. As Conn (2003) observes, if unique logons and passwords are required, student responses can be tracked, which undermines their anonymity. It is thus important to inform students how and when survey response data will be displayed and delivered as well as who will have access to the responses. The literature reports a number of advantages of using student identifiers. These include the fact that the response process of students can be traced by using cgi-scripts, java applets or log files to investigate how different students respond to online surveys (Bosnjak & Tuten 2001), and that by using exception reporting, survey administrators can target emails to students who have not completed all their requested course surveys (McGourty et al. 2002b). However, when the anonymity issue is considered by comparing online student ratings with plain e-mail surveys, where the respondent's e-mail address is generally included with his/her responses (Shannon et al. 2002; Truell 2003), there should be a greater guarantee of respondent anonymity.

According to Layne et al. (1999, p. 222) an additional weakness of paper-and-pencil faculty evaluations is the difficulty in ensuring the integrity of the administration procedures. According to them, with the increasing use of student ratings to make promotion and tenure decisions about faculty, the lack of survey administration procedures is particularly troubling: there is often no guarantee that the results have not been altered in some manner before officially becoming a part of a faculty member's file, or that survey results have not been reviewed before students' final grades have been determined.

The transition from paper-based questionnaires to online ratings also affects the publishing of the surveys in several ways. The rapid dissemination of results, mentioned earlier, is probably the most commonly cited advantage of online ratings. However, there are many other important differences in the way results are disseminated after the survey, e.g. the automated processing of results, more detailed user-friendly reports and rapid archiving and retrieval of data (Schmidt 1997; Kelly & Marsh 1999; Bullock 2003; Johnson 2003; CETL 2004).

4.5 Gaps in Current Research on Online Ratings and the Positioning of the Present Study

The previous chapters have presented an overview of the uses of student evaluations in higher education in general and the current state of knowledge concerning the uses and potential of online-based student evaluations. The following summary sets out the main conclusions:

- student evaluations are used in versatile ways worldwide, not only for improving courses and teaching but in many countries also for tenure and salary decisions. This has given rise to questions about the validity and reliability of evaluation results. As a result, hundreds of research reports have been published setting out to confirm (or sometimes reject) especially the validity of student evaluations.

- due to the lack of access to the respondents' actual individual background data, the validity of evaluations has usually been analysed statistically by manually linking the gathered responses with background information such as collective class *averages* and estimated *average* course grades etc.

- increasingly universities have replaced or are at least planning to replace traditional paper-based student evaluations with online-based systems. This also opens up new opportunities for analysing the evaluation results' validity. The emphasis of the research literature into online ratings, however, has mostly centred on investigating the differences between online and paper-and-pencil ratings systems and describing the experiences of implementing online evaluations or of the use of online evaluations. The possibilities of utilising the connections between online-based evaluation systems and universities' other data systems to analyse the validity of student evaluations have so far largely remained unutilised.

The vast amount of published articles for and against the validity of student evaluations clearly demonstrates the importance of the topic. The fact that new research findings are

continuously being published also shows that none of the existing research findings has been convincing enough to resolve the debate over the validity of student evaluations.

The purpose of this study is to make a contribution to earlier research in the following ways: an online student evaluation system is constructed and linked to the university's database. This makes it possible to analyse every question by connecting each respondents' item specific responses with their individual background data from databases (with their permission). Thus, for example, the analyses of potential biases can be based on individual responses and individual background data instead of class averages and respondents' given estimations which have formed the basis of background of prior research. To demonstrate the usefulness of the constructed integrated online evaluation system several analyses are made using the gathered responses and background data. The emphasis is on presenting the following novel findings:

- the concise matrix showing the students' responses about the courses' usefulness and difficulty compared with the actual course grades the same students were given. This has been one of the most widely researched issues in the field of student evaluations. However, in prior research the validity of the received results has been analysed utilising a substantial number of responses due to a lack of individual actual background information.

- the order and timing in which the respondents have completed their courses compared with the same students' responses concerning the same courses. This can shed light on whether students who undertake their courses according to the recommended schedule regard the courses differently from students who undertake the same courses at a later stage or in a different sequence.

A feature of this study is that all the responding students had been fully aware of all the final grades they received from all the courses which they were evaluating at the time they responded to the survey. Typical of most online ratings – as well as traditional paper-and-pencil evaluations – students are expected to evaluate a single course within a single survey with the evaluation period expiring before the actual grades are given.

In the present study the students had completed nearly all required courses for their M.Sc. degree and were evaluating all the courses together with their experiences over the entire period of their studies. Thus it was possible to link every single survey response item to the individual background data of the same respondent and analyse the results of the student evaluations more thoroughly than has previously been possible using traditional evaluation methods.

5 DESIGN AND IMPLEMENTATION OF THE EMPIRICAL STUDY

> *If you're doing an experiment, you should report everything that you think might make it invalid – not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked – to make it sure the other fellow can tell they have been eliminated.*
>
> – Richard P. Feynman –

As Brennan & Williams (2004) remark, "feedback data will be more useful if they contain or can be related to student profile data (for example, age, gender, mode of study, ethnic background)". This study was implemented to investigate the possibilities of conducting an online survey connected to university's database in order to examine the survey results combined with individual background characteristics of responding students.

In Finnish universities students have a well-established tradition of providing feedback on the courses they have undertaken. Typically the feedback has been gathered by individual lecturers from individual courses, but there have also been more systematic attempts to harmonise the evaluation methods, such as the national KOLA-program in 1990s (Lappalainen 1997, p. 27). During recent years more universities in Finland and elsewhere have implemented their own online evaluation systems or are at least intending to introduce one in the near future.[25]

Tampere University of Technology (TUT) conducts scientific research in technology and architecture, providing the highest education in these fields. Currently there are 12 600 students at TUT, 10 600 of whom are studying for an undergraduate degree. The number of students receiving their M.Sc. (Tech/Arch) degree per year has increased steadily, with 742 graduating in the year 2005. (TUT 2007) In Tampere University of Technology the systematic online-based gathering of feedback from students started as early as 1994 and since 2000 the interface of the feedback system has been web-based. In the final report of TUT's quality assurance system evaluated by FINHEEC, it was noted that TUT has a long tradition of investing in electrical systems which enhance teaching and studying. (Jokinen et al. 2007, pp. 30–31)

[25] Presentations given in a FINHEEC seminar "Opiskelija opetuksen laadunarvioinnissa" on 23rd January 2006.

The implementation phase of the integrated online evaluation system started with the university's permission at the beginning of 2002. The survey questions were modified from the former paper-based questionnaire survey used in TUT during the years 2000–2001 (Pajarre 2001; 2002) to meet the requirements of Internet surveys. The technical elements of the survey were designed together with TUT's system designer Jaakko Ruohtula, who also implemented the system. The survey was placed on the university's students' intranet, so that the site was accessible only to those students who were sent an email of the survey. As the target group of the survey were selected students who were supposed to be nearing graduation, because they had experience of almost the entire education process and because they were supposed to be able to recall their courses and programmes.[26]

The often mentioned problem of the validity of Internet-based surveys, which is due to the potential biases in the population having Internet access, is not found not be a problem when the surveys are conducted in university surroundings (Dillman & Bowker 2001; Solomon 2001; Zimitat & Crebert 2002). Nowadays, nearly all university students have e-mail accounts and are accustomed to use the Internet for enrolling on courses and gathering information on their studies. The registration system and the student course evaluations in Tampere University of Technology were transferred online in 1994, so it can be safely assumed that all the TUT students are computer literate and thus capable of responding to the survey.

The survey was carried out so that all the students received an automatically sent email of the survey when their total amount of credit units exceeded 130 cu.[27] By including all soon-to-be graduates in a survey group it was possible to avoid sample bias errors and to assess all such students' willingness to respond to the survey. When this threshold was reached, an automatic email was sent to inform the student about the survey containing a description of the survey and a URL-link to the survey Web page. The survey was implemented so that if the student responded to the survey more than once, only the last answer was recorded in the database. There were intentionally no reminders of the survey since the purpose was to study the response behaviour of students without the use of incentives or remainders.

The Internet-based survey consisted of two parts which were both accessed via the common survey Web site. The first part was a questionnaire on student's opinions of all the courses she/he has executed during her/his studies. The second part was a general questionnaire on the student's opinions of her/his education as a whole. The possibility to connect an online

[26] The idea of informing the students about the survey when they actually have graduated was rejected because the students' email accounts are closed very soon after graduation. Thus it wouldn't have been possible to notify the students about the survey nor would it have been possible for the graduated students to have access to the university's intranet to respond to the survey.

[27] Corresponding to approximately 217 ECTS. During the survey execution time the Masters of Science degree consisted of 180 credit units which included a Master's Thesis of 20 credit units. Thus a typical student receiving the email of the survey was in the final stage of his/her studies.

survey to other data systems was already utilised in constructing the course questionnaire part. The course questionnaire was particularly implemented so that when the student logged into the survey, the software connected to the university-held student database. The student was requested to evaluate *those and only those courses* that she or he had completed *in the same order* the student had completed the courses. Thus the latter part of the survey was common for all respondents whereas the first part of the survey was *unique to each student* requesting opinions only on those courses that the particular student had passed. As far as is known, this represents a novel, unprecedented way to construct an automatically generated, yet an individual, questionnaire for large student groups.

The topics and contents of the survey were modified from a paper-based questionnaire which was submitted to all soon-to-be graduate students at TUT over a year and-a-half during the years 2000 – 2001. While informative as such[28], it was paper-based and did not provide the possibility to analyse the respondents' backgrounds any further.

Because the survey questions were very similar to those presented in prior paper-based surveys (Pajarre 2001; 2002) the questionnaire was not specifically tested with any test group before taken into use. However, a couple of weeks after launching the survey all responses gathered so far[29] were examined for any possible errors or inadequacies in the survey. Since the results of the examination seemed satisfactory, no changes were made to the actual questionnaire. Some modifications were made to the source code of the user interface immediately after taking the survey into use, since some respondents' had reported problems with certain web browsers which had not been detected in tests with the most common browsers (Internet Explorer, Netscape). These modifications, however, did not alter the actual user interface or the survey questionnaire itself.

To inform the respondents of the purposes of the survey, the introduction page of the survey provided a short description about the survey and the intention to use of results of the survey for research purposes. The students were told that any background information on them would only be collected with their full consent and that their identity would not be revealed in any case. To ensure that no background information was gathered from students without their explicit approval, the default value for the question that requested the permission to gather background information was set negative.

The research literature contains descriptions of some of the administrative and organisational requirements for the Web-based student evaluations which have been recognised and

[28] The results of this survey are presented in two reports: "Tutkinnon sisältö ja läpäistävyys – raportti TTKK:sta vuonna 2000 valmistuneiden antamasta palautteesta" and "Raportti tutkinnon suorittamista hidastavista tekijöistä TTKK:ssa" (Pajarre 2001; Pajarre 2002).

[29] Unlike typical surveys where the respondents' all receive the invitation to respond to the survey at the same time, in this case each student received an automatic email invitation from TUT OPREK administrator to respond to the survey the day after her/his amount of credit units in TUT OPREK database exceeded 130 cu.

implemented already in some of the first online evaluations. For example Recker & Greenwood (1995) described system requirements where the "responses must be authenticated, confidential, and there must be no more than one response per user. Additionally the system must be reliable and robust." In order to prevent multiple responses from single students, the instrument in this survey was implemented so that if the student answered to the survey more than once, only the last response was stored into the database.

Unlike typical student ratings of teaching, which tend toward an atomistic approach to evaluation by asking for detailed information on the characteristics of the teacher and the course (Cannon 2001a, p. 90) this survey sought to gain a holistic overview of the whole M.Sc. degree. Thus the survey was designed to encompass not only questions about single courses but also questions about all courses as well as experiences gained from the studies and other issues related to studies overall. In the literature arguments have been presented against the use of questions posed long after the course has been taken since it is believed that the students cannot be expected to distinguish their experiences between different activities (e.g., Brennan et al. 2003). However, this study was designed to contain only two questions per course and thus the respondents did not need to remember any specific features of the courses, only their views on whether the courses were useful or not and the level of difficulty of the courses.

After the survey period was over, all survey data were downloaded together with background data from those students who had given their consent. In order to match survey data files with background data a software script in Perl was written, after which the data were imported into Excel for formatting and statistical analysis.

Typical student ratings presented in the research literature are student evaluations on teachers' or teaching effectiveness and contain dozens of items involving a 5-point or a 7-point Likert-type scale. Many of such ratings are based on widely-accepted multidimensional instruments such as SEEQ (Student Evaluation of Educational Quality), SIR or SIR II (Student Instructional Reports)[30] and are analysed by using factor analysis. In the present study, however, the focus of the survey is not on teachers' characteristics but on individual courses and the students' opinions of their education as a whole. The purpose of the present study is to show the usefulness of the constructed evaluation *system* in comparing the evaluation results with the actual individual background variables of the respondents to discover potential connections, such as those between students' grades and responses. Because of this the emphasis is less on statistical procedures but more on the results themselves.

It has been claimed that the multiple-choice questionnaires offer no latitude for the information that is gathered (Oliver & Conole 1998). To avoid this the survey instrument was

[30] These and other evaluation instruments are more thoroughly described e.g., by Marsh (1987, p. 266).

constructed so that it contained both open-ended questions concerning the students' opinions of their education and two course specific items involving a 3-point scale for each course the student had completed. The survey instrument thus consisted of both text-entry boxes and menu selection bars. The scale was confined to three alternatives due to the large number of courses each respondent was to assess.[31]

Some researchers have also questioned the appropriateness of using short forms for gathering summative information. For example Jackson et al. (1999) deny the value of shorter questionnaires arguing that "instructors would never consider evaluating students' performance with only four or five items" and stating that basing promotion or tenure decisions on four or five items is inadequate from a measurement point of view. It should be noted, however, that in the same article, Jackson et al. also refer to the opposite opinions which do support to the use of short measure for summative evaluation. In this study the survey results are not used for promotion or tenure decision purposes; in fact, there are no single questions concerning any attributes of the teacher.

According to Marsh (1987, p. 261) validating interpretations of student responses to an evaluation instrument involves an ongoing interplay between construct interpretations, instrument development, data collection and logic. He believes that each interpretation must be considered a tentative hypothesis to be challenged in different contexts and with different approaches. The present study concentrates on assessing the possibilities of utilising background information to analyse responses gathered from student evaluations. The focus, therefore, is on observing validity issues. However, the existence of potential grading leniency in the form of the students' attitudes to the personality of their teachers is not an issue of this study. Instead, this study analyses the role of background information in analysing students' responses concerning the courses and programmes they have completed, compared, for example, with the actual grades achieved in the courses or with the implementation order of the courses they have undertaken.

5.1 Contents of the Empirical Study

The implemented research study comprises two interlinked elements: a student ratings system gathering students' opinions from all their studies at a course level and a connection to the university's database containing background information from the students and the courses they have completed.

[31] Students who responded to the course survey part assessed each an average of 55 courses.

A widely known problem with the traditional student evaluations – both paper-based and online-based – is that the volume of questions in the survey can easily become unmanageable. If the questionnaire seems to be too long, the students' willingness to respond to the survey may decrease remarkably. Since the university's database already contained information about the students' age, gender, training program and the year they began their studies, these questions were intentionally left out from the questionnaire. Also any possible manual errors in giving responses were eliminated this way. A third, significant benefit gained from linking the actual survey to the university database (and not just the survey results) was that the questionnaire was unique to every respondent: when the students logged in to the survey web page, the software presented a list of specifically those courses with response alternatives which that particular student had completed.

The online-based questionnaire was implemented so that the following issues can be analysed:

- all the courses the student had undertaken and the detailed implementation order in which they had completed their courses (this was possible since the online survey was connected to the university database and implemented so that the student was asked to evaluate only those courses she/he had completed).
- the respondent's opinion of all the courses she/he had completed in terms of the courses' usefulness and difficulty (later referred to as perceived usefulness/difficulty)
- the respondent's estimate of his/her work experience during the time of study
- the respondent's opinion on which factors hindered his/her progress in his/her studies and a possible suggestion for a solution
- the respondent's opinion on which would be his/her selection for his/her main subject of study if he/she would have to choose it now
- the respondents satisfaction with his/her opportunity to study a satisfactory amount of foreign languages
- the respondent's opinion of whether certain factors (a given list) impeded him/her in his/her studies and a space for possible open-ended explanations
- a space for other possible opinions (open-ended space with maximum limit of 256 characters)
- the respondent's opinion of the overall degree he/she would award the university for all the teaching he/she has received throughout the studies
- approval/denial of permission for the use of background information from university's other databases.

The online survey was connected to Tampere University of Technology's student database OPREK which contains e.g., the following information:

- the student's identification number
- the student's training program
- the degree the studies lead to (i.e., M.Sc. in Technology or M.Sc. in Architecture)
- detailed information on the courses passed such as
    - the institute responsible of organising the course
    - exact name and number of the course
    - name of the lecturer who taught the particular course
    - execution date for the course (several dates if the student has taken an examination more than once)
    - grade from the course (all grades, if the student has taken an examination more than once)
    - credit units for the course
    - date of examination of the course (not necessarily the same as the date for completion of the course)
    - number of examinations taken on a course
    - number of examinations passed on the course.

The perceived difficulty and usefulness of the courses were selected as the focus of evaluation because they are among the most commonly presented questions in student evaluations (see e.g., Olivares 2001). As described in Chapter 3.3.1, *Validity,* the students' grade point averages have played an essential role in the ongoing debate over validity issues, especially the contradictory views regarding the existence of grading leniency in student evaluations (Marsh 1987; Greenwald & Gillmore 1997a; Chambers & Schmitt 2002; Millea & Grimes 2002; Sojka et al. 2002; Centra 2003; Felton et al. 2004 to name but a few). The online evaluation method constructed in the present study enables the analyses of students' actual *individual* grades instead of class averages (or students' individual *estimations* of their forthcoming grades). Therefore, the analysis in the present study seeks to highlight the various ways the validity of student evaluations can be observed in terms of individual data.

The selected sample of the population was all those students who were soon to graduate.[32] This was found appropriate for the following three reasons: They were assumed to have formed a general insight of their overall education since they have completed nearly all required courses, yet they were unlikely to have forgotten their experiences of the courses.

[32] A figure of 130 credit units was set as a threshold to determine when a student is approaching graduation.

Second, it was assumed that they had gained at least some work experience to enable them to compare their education to their needs in work life. And third, since they were still students, they had a valid student email address which they were assumed to use actively. This would make them accessible, unlike, for example, students[33] who had graduated but who would also have been able to assess their education and their work experience.

Examples of the survey's front page, the course part and the general part are presented in Appendices 1, 2 and 3.

[33] In Tampere University of Technology the student email addresses are closed within two weeks after graduation.

6 RESULTS OF THE EMPIRICAL STUDY

Statistics is the art of making numerical conjectures about puzzling questions.

– Freedman, Pisani & Purves –

This chapter introduces empirical examples of the diverse ways in which new kinds of evaluation information can be yielded by developing an online student evaluation connected with other university databases. The focus of interest is on demonstrating the novelty value of the type of yielded results. Thus the results are considered more from the viewpoint of how this kind of evaluation system enhances the assessment of student evaluations validity, and less on the validity of the presented exemplar results itself.

When analysing the results of the constructed integrated online evaluation system, several possible units can be analysed from the gathered responses and their background data: the characteristics of the respondents, the individual courses, the individual teachers or even the individual subject or department. The focus may be on assessing, for example, how different students respond to the survey, how different courses are regarded by the students, or how the students' success and progress in their studies correlate with their responses to the survey.

In the present study the focus is on the latter alternative, analysing the interdependence between students' responses and their success in their studies. Unlike most previous studies where the survey responses have been compared with *average* or expected *mean* course grades and other average background information, this study compares the individuals students' actual course grades and other information with the same students' actual survey responses.

Linking up the individual students' responses with their background information from university's database enables a number of possibilities, e.g., analysing in greater detail the interdependence between students opinions on a particular course and the actual grade from that same course. It also makes possible to observe the effect of the implementation order of courses on how students assess those courses (for example, do those students who have completed their obligatory basic courses in the first one or two years of their studies assess

the courses' usefulness and difficulties differently to those students who have undertaken those courses just prior to graduating).

To be able to analyse individual students' responses in terms of their individual background information, permission had first to be requested from students. This resulted in two groups of responses being formed: responses from students who approved the use of background data and responses from students who denied the use of background data. Figure 5 below shows the distribution of the two groups of respondents as a subset of all students in the university.

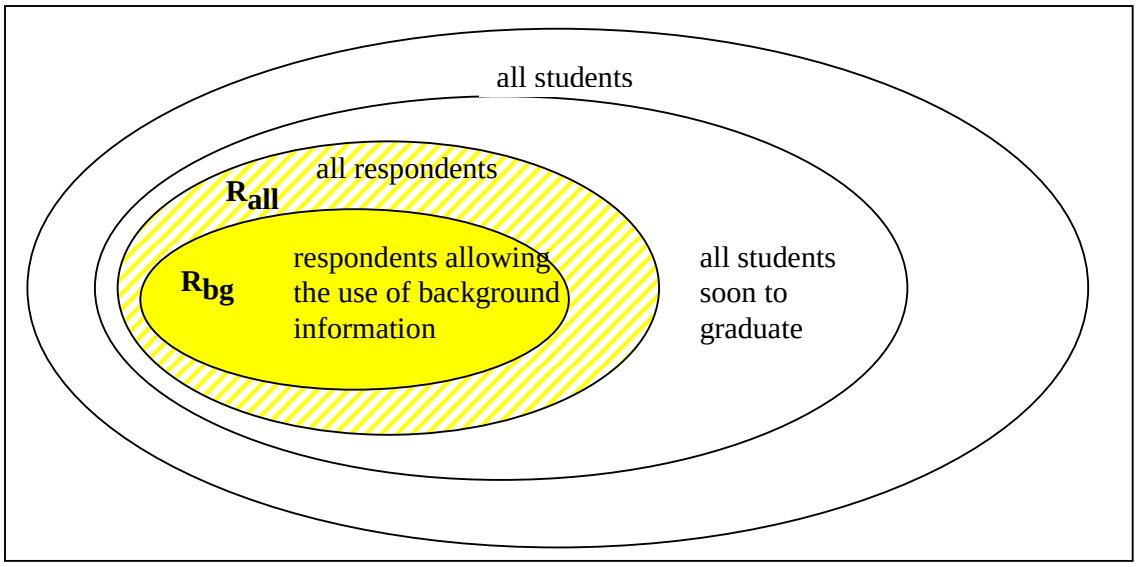


Figure 5. Respondents as a subset of the total group of students within a university.

In the following analysis the group of respondents who allowed the use of background data is referred as R_{bg} and the group of respondents who denied the use of background data is referred as R_{den}. All students responding to the survey, $R_{bg} + R_{den}$, are referred as R_{all}.

Typically the results of research studies of student evaluations have been presented numerically. There are however, arguments in favour of visually displaying evaluation results. According to Theall (2001) perhaps the greatest value of visual displays is that they can clearly show overlaps of the intervals and signal when a simple no-yes decision cannot be made on the basis of the data. In the following analysis the results are often presented both numerically and graphically to help provide overview of the observed issue.

6.1 Descriptive Analysis of the Data

6.1.1 Response Rates and Opinions on the Use of Background Data

The empirical survey part of this study is based on an online survey which was carried out starting in April 2002 and ending in August 2003. The total number of students visiting the survey web pages was 455. The exact total number of students who received the email notifying of the survey, was not recorded, but if the number of respondents is compared with the total number of students graduating from Tampere University of Technology[34] during the period of the survey, which was 986, an approximate response rate can be calculated at 46 per cent. However, the actual response rates in the case of many questions are substantially lower since several students left blank many sections of the survey pages.

Because of the constructed connection between the online survey and university's other databases typical basic questions such as the respondents' gender, year of study or study programme were not asked in the survey questionnaire. They were intentionally omitted from the questionnaire to shorten the questionnaire and eliminate possible human error in responding to such questions. Information on the respondents' gender was available in the university's database for those respondents who allowed the analysis of their background. Of these students (respondent group R_{bg}) 73 % were male and 27 % female. Of all students graduating at the same time 77 % were male and 23 % female. Thus female students were slightly more willing to allow the analysis of their background data than male students. Figure 6 shows how the responses distributed among respondents (R_{bg}) in different study programmes. Detailed figures for all respondents' (R_{bg}) gender and study programme are presented in Appendix 4.

[34] Retrieved from the university's official public statistics.

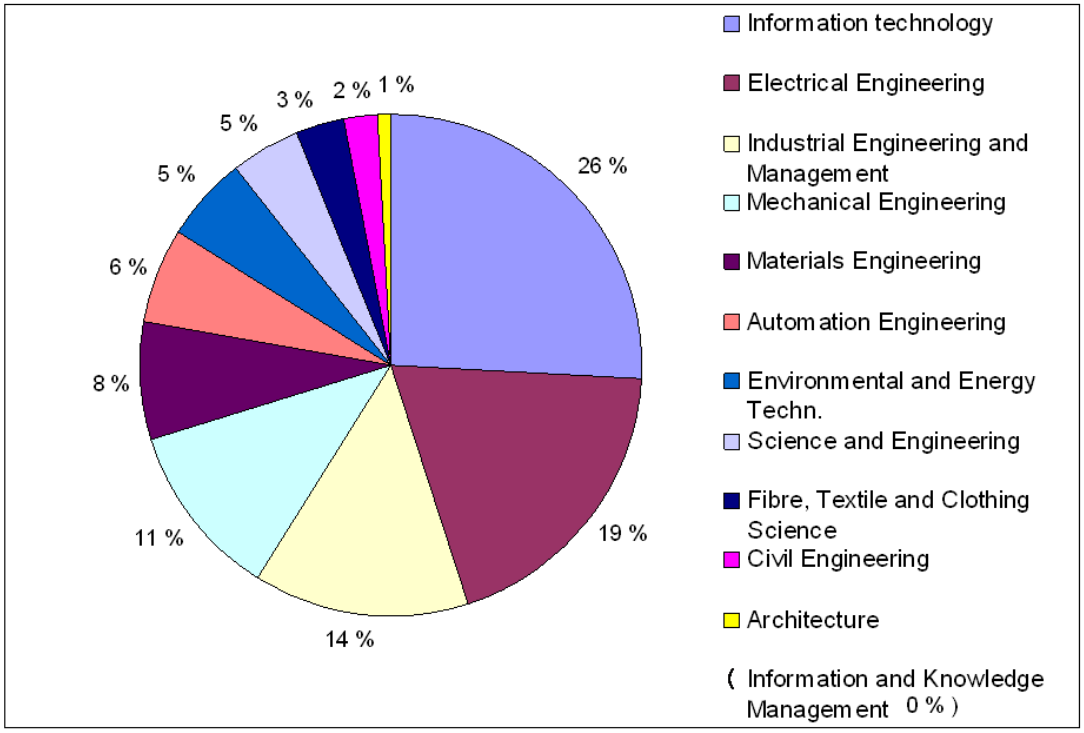


Figure 6. The distribution of respondents' (R_{bg}, n=131) study programmes.

As could be expected, there were more responses from students studying in those study programmes where the annual intake is higher than from study programmes where the annual intake is smaller. Table 5 presents the response rates among students (R_{bg}) in different study programmes. More detailed figures are presented in Appendix 4.

Table 5. Response rates among respondents ($R_{bg,}$ n= 131) in different study programmes.

Study Programme	**Response rate (compared to the number of students graduating during the survey execution time)**
Architecture	3 %
Automation Engineering	10 %
Civil Engineering	9 %
Electrical Engineering	13 %
Environmental and Energy Technology	15 %
Fibre, Textile and Clothing Science	18 %
Industrial Engineering and Management	15 %
Information and Knowledge Management	0 %
Information technology	10 %
Materials Engineering	15 %
Mechanical Engineering	14 %
Science and Engineering	32 %
Total	**13 %**

Several research articles have noted students' concern regarding their privacy and anonymity when responding to student ratings (Cummings & Ballantyne 1999; Kelly & Marsh 1999; Ballantyne 2000; Cummings et al. 2001; Hardy 2003; Oliver & Sautter 2005). According to such authors with traditional paper-based questionnaires it is possible that students' handwriting and especially handwritten comments can be identified. When student ratings are performed online the possibility of identifying students by their handwriting is eliminated, but some students may fear that their identity can be revealed through the log-in process (e.g., Conn 2003).

In this study responding to the survey required logging in to the university's intra pages with the students' username and password. This process was similar for all respondents regardless of whether or not they allowed the use of their background data. In the survey's informed consent the students were given explicit assurance of anonymity. However, a question of allowing the use of background data may have given the students the impression that their identity would be more easily revealed if they permitted its use than if they denied it.

In total 131 students answered positively to the question "Would you let your background data from your study history be combined to the survey?" Of the responding students, 297 respondents answered negatively (65 % of the total responses) and 27 respondents did not visit the question page (6 % of the respondents). The answer of the latter group was interpreted as being negative even though they actually did not answer the question. Thus the percentage of students participating to the survey and allowing the use of their background data compared with the total number of respondents was 29 %. The total response rate of those students who allowed the use of their background data (compared with all students graduating during the survey period) can thus be estimated to be 13 %. It should be noted that the default answer for the question was intentionally set at negative so that in order to answer positively, the students had to deliberately click the answer button in that question and select the "yes" alternative. This feature was programmed to prevent permission to use background data being given in error.

There was also a difference in response rates between the two parts of the survey, the student-specific course survey and a general questionnaire, which was similar for all respondents. Table 6 presents the response rates of these two parts of the survey:

Table 6. Number of respondents to the different sections of the survey.

	Number of Respondents (total)	**Number of Respondents to Course Survey Part**	**Number of Respondents to General Questionnaire Part**
R_{bg}	131	115	131
R_{den}	324	69	45
R_{all}	**455**	**184**	**156**

The relatively moderate percentage of students allowing the use of background data (R_{bg}) indicates that most respondents have been reluctant to risk their anonymity. For one reason or another, those students who denied the use of their background data (respondent group R_{den}) had also more often left one or both survey parts unanswered compared to those who allowed the use of their background data. However, the much higher overall response rate indicates that the students did not expect to be identified by logging in and visiting the survey pages[35], or that they were not concerned about their possible identification when they only responded

[35] Before the students were asked in the survey for permission to use their background data, they had already logged in to the survey by using their TUT intranet username.

to the survey, compared to the possibility that their background information was connected to the survey results.

On the other hand, the fact that nearly a third (29 %) of all the respondents did allow the use of their background information makes it possible to analyse, whether there are marked differences with the students who did allow the use of background data compared to other students. This is possible by analysing the data gathered from students who allowed the use of background information, R_{bg}, with all the survey responses (without the use of background data) and with public university statistics covering all students. These differences are analysed in greater detail in the following sections.

6.1.2 Differences between Respondents

Before analysing the results, it is important to discover whether the two groups of respondents, those who allowed the use of their background information, R_{bg}, and those who denied it, R_{den}, differ from each other significantly and, if so, in what way. This is done by analysing three issues:

1. differences between R_{bg} and R_{den} groups in their opinions on the difficulty of the courses,
2. differences between R_{bg} and R_{den} groups in their opinions on the usefulness of the courses, and
3. differences between R_{bg} and R_{all} in their overall grade point averages.

1) Opinions on the difficulty of the courses:

The total number of estimations concerning the difficulty of the completed courses made by students who *allowed* the use of their background data was 6337. The average number of courses assessed by one respondent (in respondent group R_{bg}) was 55 courses. Opinions concerning the difficulty of the courses were distributed as follows: course was “easy”: 2339 responses, course “rather difficult”: 3117 responses and course “very difficult”: 921 responses. Thus more than a third, 37 % of the courses were estimated as “easy”, 49 % were estimated as “rather difficult” and 14 % as “very difficult” (details in Appendix 4).

The total number of estimations concerning the difficulty of the completed courses made by students who *denied* the use of their background data was 3698. Their opinions on the difficulty of the courses were distributed as follows: course was "easy": 1411 responses, course "rather difficult": 1769 responses and course "very difficult": 518 responses. As percentages, these figures show that similarly more than a third, 38 % of the courses were estimated as "easy", 48 % were estimated as "rather difficult" and 14 % as "very difficult".

To discover whether the two respondent groups differ from each other significantly, a Chi-square test was performed. The results showed that no statistically significant differences existed between the responses of respondent groups R_{bg} and R_{den} concerning the courses' difficulty (details in Appendix 4).

2) Opinions on the usefulness of the courses:

The total number of estimations concerning the usefulness of the courses completed as assessed by students who allowed the use of their background data was 6380. The opinions concerning the usefulness of the courses were distributed as followings: course "not useful": 1088 responses, course "quite useful": 3231 responses and course "very useful": 2061 responses. Thus slightly more than a half, 51 %, of the courses were estimated as "quite useful", 32 % were estimated as "very useful" and 17 % were estimated to have been "not useful".

The total number of estimations concerning the usefulness of the completed courses as assessed by students who denied the use of their background data was 3747. Their opinions on the usefulness of the courses were distributed as follows: course "not useful": 800 responses, course "quite useful": 1790 responses and course "very useful": 1157 responses. In percentages these figures show that nearly a half, 48 %, of the courses were estimates as "quite useful", 31 % of estimated as "very useful" and 21 % were regarded as "not useful".

Again to analyse the differences between respondent groups R_{bg} and R_{den} a Chi-square test was performed. This time the results showed that at 5 percentage significance level, there was a statistical difference between the responses of the two respondent groups concerning the courses' usefulness (detailed results in Appendix 4). Those respondents who had denied the access to their background data had been somewhat more critical than those respondents who had admitted the access to their background data. This difference has to be remembered when the responses are analysed in the following chapters.

3) Differences in grade point averages:

To analyse if those respondents who allowed the use of their background data performed significantly better or worse in their studies than "average" students, i.e., whether success in their studies differs of the average student's success in his/her studies, the mean of students' grade point averages (referred here as "mean weighted average") was calculated from those students who allowed the use of their background data. These results were compared with a mean of grade point averages of all students graduating during the period the survey was active.[36]

The mean weighted average[37] of respondent group R_{bg} (n= 121) was 3.3. The mean weighted average of all students graduating at the time of the survey (n= 986) was 3.2.

6.1.3 Risk of Making the Analysis too Narrow

It is important to recognise that when a single course is selected as a target of analysis, opinions between different respondent groups or the distribution of opinions may vary significantly. To illustrate this, Figure 7 presents respondents' opinions concerning some basic courses lectured in TUT. The first four of the selected courses are obligatory for all M.Sc. (Tech.) students in TUT and the fifth course is obligatory for students in Automation Engineering study programme.[38]

[36] Retrieved from the university's official public statistical figures by "Opiskelijapalvelut".
[37] The grade for the M.Sc. thesis was excluded from the calculations since it is not included in the final average in the M.Sc. degree either.
[38] The courses were selected because the number of responses to these courses was higher than for many other courses. This may be because in addition to being obligatory courses, the names or identification numbers of these courses had changed little over the years in contrast to many other courses in TUT.

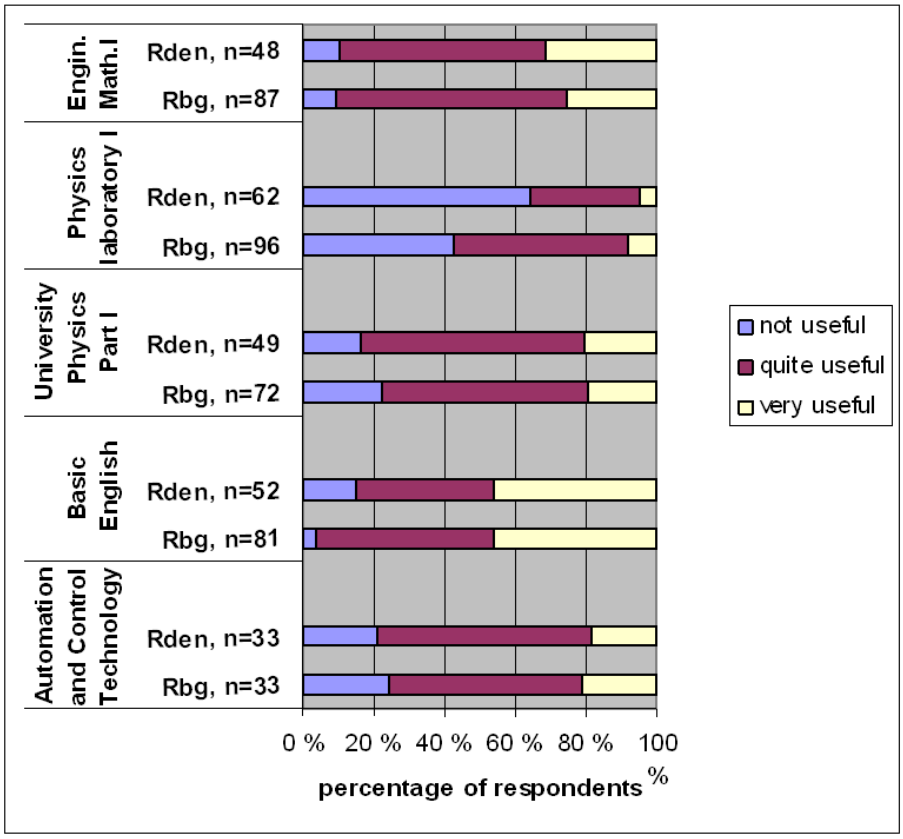


Figure 7. Respondent's (R_{bg} and R_{den}) opinions on the usefulness of some basic courses.

The results in Figure 7 above demonstrate how interpreting results with such limited data is difficult and may even be misleading. For example, the course Physics Laboratory I was more often considered as "not useful" by those respondents who had denied access to their background data compared to those respondents who allowed it. On the University Physics Part I course the responses were somewhat the reserve; those respondents who had denied access to their background data were slightly more often satisfied with the course's usefulness than those respondents who had refused access. To better understand the results gathered from student evaluations, information from students' background is essential. The following chapters present examples of survey results where the background data is utilised. By integrating individual background data with original responses even scarcer data can be analysed better and understood. This reduces the risks of misinterpretation which, at worst, can lead to incorrect actions on courses or lecturers.

6.2 Utilising Background Information in the Search for Potential Biases

Several variables are expected to influence student ratings; yet the following four characteristics are claimed to have shown consistent effects on student ratings: prior subject interest, expected grade/actual grade, reason for taking the course and workload difficulty (Olivares 2001). Grading leniency in particular is perceived to explain even 20 % of the variance in ratings (Greenwald & Gillmore 1997b).

Common to the numerous research findings analysing the existence of potential biases in student ratings, though based on ratings data from thousands of students, is that they have been founded on analysing student ratings results with average student course grades, average estimates of workload etc. For example, the findings of Chambers & Schmitt (2002) which showed that students who expected higher evaluations from their instructors rated them more highly in return, were based on class-averages. Centra (2003), on the other hand, showed in his extensive research that students with higher expected grades had given somewhat lower evaluations, the very opposite of a grading leniency expectation.[39] Due to the absence of any direct connection from individual student ratings results to a particular student's actual course grades and other information, the previous observations were all based on expected and/or average values instead of individual factual ones. One of the very few findings where background data has been included into the survey analysis is presented by Kerridge & Mathews (1998) who linked a paper-based survey manually to a university-held database to analyse four parameters: individual grade results, admission entry criteria, gender, and academic year/level.

This chapter presents findings in which student ratings results are connected and analysed together with individual student background data. The issues which are examined, some of them widely argued in the existing research and yet many of them not researched earlier, are as follows:

- Perceived usefulness of the course vs. individual students' actual course grades and perceived difficulty of the course vs. individual students' actual course grades
- Analysis of the interdependence between the individual students' order in which they have completed the courses, the same students' estimations of the difficulty and usefulness of the courses and the actual grades awarded
- Analysis of the interdependence between the courses' perceived usefulness, perceived difficulty and individual students' success in their studies

[39] Centra's (2003) study was based on students' own personal estimation of the grade they were expecting to receive which were then statistically analysed.

- Respondents' satisfaction with their major subject vs. the same respondents' overall success with their studies and the duration of their studies
- Analysis of the individual respondents' gender, in terms of the responses and the same respondents' opinions on factors hindering their studies
- Analysis of the individual students' work experience, grade point average and the duration of their studies
- Other analyses, e.g., the respondents' satisfaction with the opportunity to study foreign languages compared with their overall grades from their language courses.

The results were gathered by combining online survey results with students' individual background data from university held database with students' permission. In the following sections are first presented the results from the survey's course-specific questions and then the results from the general part of the survey.

6.2.1 Opinions on the Perceived Usefulness and Difficulty of the Course versus Students' Actual Course Grades

One of the most widely researched potential biases in student ratings is the connection between students' success in their studies and students' opinions of the courses (e.g., Gramlich & Greenlee 1993; Mason et al. 1995; Husbands 1998; Chambers & Schmitt 2002; Centra 2003). A frequent observation is that students' receiving (or expecting) good grades from the course also evaluate the same course favourably. In typical research studies this topic has been observed so that students' responses have been compared with expected mean grade point averages of the whole class. In the present study, the students' opinions on the course (usefulness and difficulty) were compared with the *same* student's *actual* grade from the *same* course. This eliminates or at least reduces the need for complex statistical operations such as factor analyses, which are extremely common in the field. Such analyses are necessary when the absence of actual individual data is replaced with average expected values.

In the survey's course questionnaire part, the students were asked to assess the usefulness and difficulty of every course they had undertaken.[40] Of the students responding to this part of the survey, 115 had agreed the use of their background data and 69 had denied it. The total

[40] The survey was implemented so that when a student logged into the survey page (i.e. the student was identified) the software connected the university held student database and loaded visible for the student to evaluate those and only courses that she or he had executed.

number of courses[41] the respondents had given their opinion on was 1390. In the following, the responses of those students who allowed the use of their background data are compared with their actual course grades from the same courses. The possible differences between the responses of those students who allowed the use of their background data and those who denied it were compared in Chapter 6.1.2 and are taken into account when analysing the results.

Usefulness of the course vs. actual grade from the course

The connection between the online survey and other university data systems made it possible to compare how those students who allowed the use of their background data had estimated the courses with the actual grades they had been awarded in the same courses.

Table 7 below presents the deviation between the students' (R_{bg}) opinion on the usefulness of the course and the actual grade the same students received for that particular course.[42]

Table 7. Connection between perceived usefulness of a course and the actual grades given to the same students (R_{bg}) from the same courses presented as percentages (total amount of responses 6901, total amount of courses estimated 1390).

Course grade	very useful, %	quite useful, %	not useful, %	not estimated, %
Approved (no grade)	3.8	5.0	2.6	1.3
5	6.6	8.4	1.8	1.4
4	7.8	10.3	2.9	1.6
3	6.3	11.2	3.4	1.7
2	3.0	7.4	2.7	1.5
1	1.5	4.4	2.4	0.9
Σ	29.0	46.7	15.8	8.4

The same grade-estimation connections can be presented graphically as in Figure 8:

[41] The actual number of different courses is somewhat smaller since in some cases the identification number of the same (or equivalent) course has been different in different years.

[42] The table includes all courses the respondents have assessed except the M.Sc. thesis. This was excluded from the data because even though it is recorded in the database with a course number it differs fundamentally from all other studies leading to the M.Sc. degree.

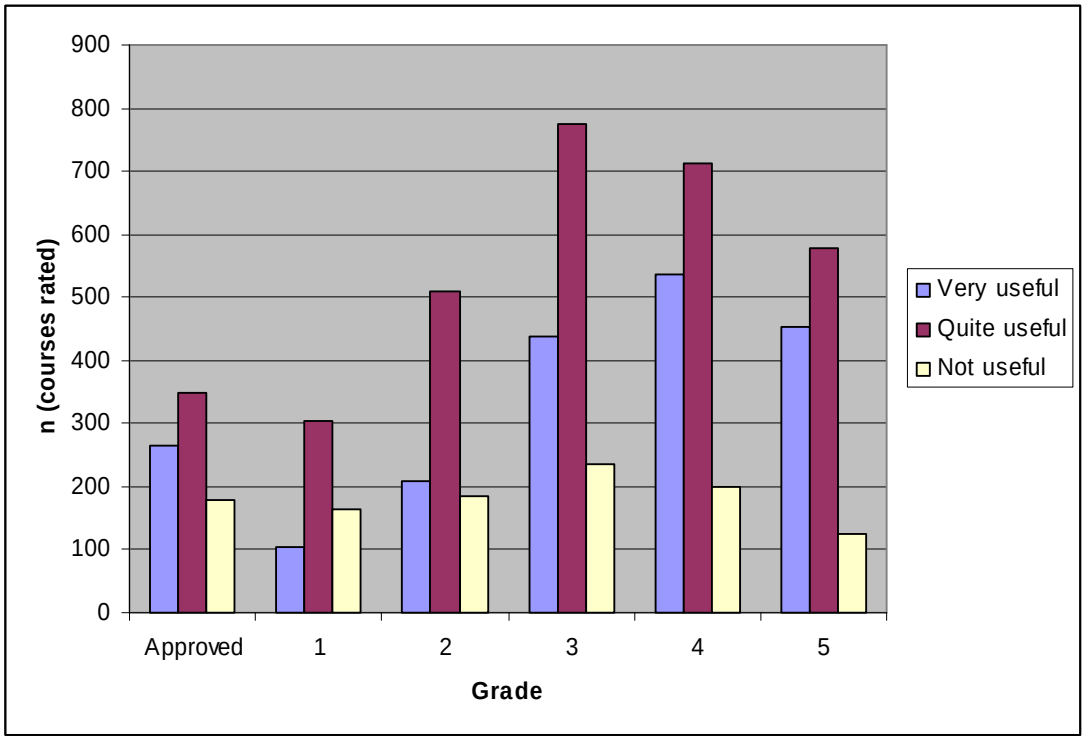


Figure 8. Distribution of the respondents' (R_{bg}) opinions on the usefulness of the course and the corresponding grade received by the same student for the same course, n(responses given)=6901 (M.Sc. theses excluded).The total number of courses undertaken = 1390.

According to Figure 8 even though the students had most often responded "quite useful" to those courses where their actual grade hade been three or four, there are no clear connections with the course grade and the perceived usefulness of the same course. In fact the reverse is the case, for example, there were more mentions of the course having been "not useful" among those respondents whose final grade had been three or four than among those students whose final grade had been lower. It is also worth noting that, regardless of the grades the students received, the estimation "quite useful" was dominant in all grade groups. Nearly a third of the estimated courses (32 %) were given the value "very useful", whereas half of them (50 %) were considered "quite useful" and only 18 % were considered as "not useful".

When analysing these results it has to be remembered that those students who allowed the use of their background data (n=131) and responded to these questions did regard the courses usefulness in a rather more positive light than those students who denied the use of their background data (n=324). However, since there were no major differences between the two respondent groups (see Chapter 6.1.2, *Differences between Respondents*) these figures can also be assumed to reflect to some degree the responses of "average" students. Thus the frequently presented assumption that students performing poorly on a course also give low evaluations of that same course is not supported by these results. For example, 71 per cent of

the courses where the respondents' final grade had been one were estimated either as "quite useful" or "very useful".

Another issue to be noted when analysing previous results is that the respondents evaluated all the courses they had executed at the same time. Thus some of the courses had been completed several years earlier and other courses only recently. Later in this study (Chapter 6.2.2, *Opinions on the Courses compared with the Accomplishment Order of the Courses*) there is discussion of certain issues relating to the implementation order of the courses. With this data, however, it is impossible to know whether the respondents' opinions of the courses have changed during their university career. It is possible that as they near graduation they have understood better the value of the courses in retrospect and rate a course more highly. Conversely with the passage of time and greater experience, they could also become more confident in making adverse judgements about their former courses.

Perceived difficulty of the course vs. actual grade from the course

In addition to the usefulness of the courses the respondents had also estimated the difficulty of all the courses completed at the time of responding to the survey. Table 8 presents the distribution of opinions given for the same course by those respondents consenting to the use of their background data connected to the actual final grades awarded to the same students for the same course.[43]

Table 8. The connection between perceived difficulty of the course and the actual grades given to the same students from the same courses (R_{bg}), presented as percentages (total amount of responses 6897 from 115 respondents).

Course grade	very difficult, %	quite difficult, %	easy, %	not estimated, %
Approv. (no grade)	0.7	3.3	7.4	1.3
5	1.3	8.0	7.5	1.5
4	2.2	10.5	8.3	1.6
3	2.8	11.2	6.9	1.7
2	3.1	7.2	2.8	1.4
1	3.0	4.4	0.9	0.9
Σ	13.1	44.6	33.8	8.4

[43] As with Table 6, presenting the respondents' (R_{bg}) opinions on the usefulness of the courses, Table 7 includes all courses the respondents have assessed, except the Master of Science theses.

These figures are presented graphically in Figure 9:

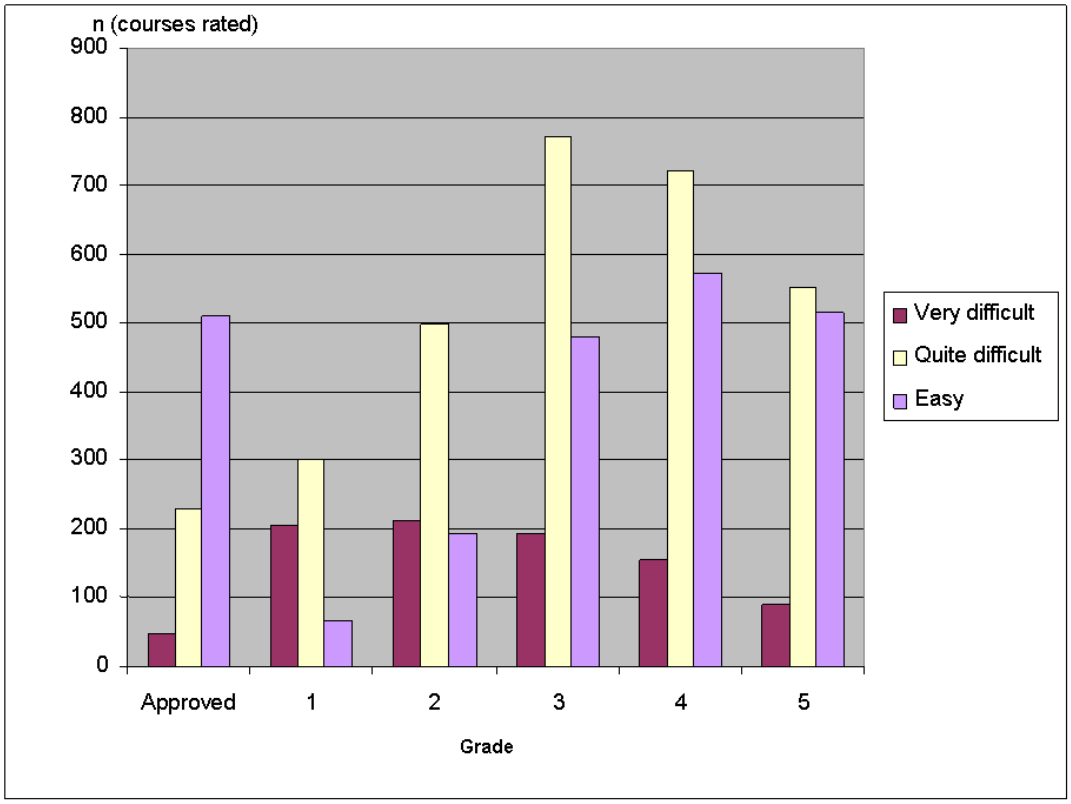


Figure 9. Distribution of opinions considering the difficulty of the courses and the grades the respondents (R_{bg}) received from the same courses (total amount of course evaluations given 6897).

The figures indicate that most of the courses, 48 % (n=3073) were perceived as "quite difficult". Additionally, 37 % (n=2335) of the estimated courses were perceived as "easy" and only 14 % (n=905) as "very difficult". Among those courses where the students did not receive a numerical grade, but only an "approved"/"failed" (or pass/fail) evaluation the most typical estimation was "easy" (65 %) and only 8 % of the courses were judged to be "very difficult".

Since the opinions of the analysed respondent group (R_{bg}) did not statistically differ significantly from the responses of all respondents (see Chapter 6.1.2, *Differences between Respondents*) it can be assumed that these results of the respondents' opinions on the difficulty of the courses can be generalised across all respondents.

The previous analyses have focused on evaluated courses as a whole. In the following sections 6.2.2 and 6.2.3, examples are presented of analyses of separate courses which have been analysed in conjunction with the individual background data.

6.2.2 Opinions on the Courses Compared with the Accomplishment Order of the Courses

The access to respondents (R_{bg}) background data made it possible to analyse whether the order in which the students carry out their courses affected the way they estimated the courses usefulness and difficulty and also their performance in the courses. In Tampere University of Technology the courses are designed so that accomplishing them in the recommended order is beneficial in terms of both learning and scheduling. In practice, however, the students can choose quite freely the order in which they undertake the courses and sit their exams. This freedom can, however, be problematic, not only in TUT, but in many universities, since it often leads to a protraction in studying time (Kärkkäinen 2005). Often the students leave some obligatory courses unfinished until near graduation and in some cases the graduation is delayed or even suspended due to the unfinished courses (e.g., Lehtimäki 2006). For example, certain compulsory courses in mathematics, physics and chemistry, which all students studying in M.Sc. (Tech.) program[44] are recommended to carry out during the first two study years, are in many cases among the last courses to be completed (Pajarre 2001).

To analyse whether the accomplishment order has an effect on the students opinions on the usefulness and difficulty of the courses, the students (R_{bg}) opinions on an obligatory course, Physics Laboratory I course[45], were analysed together with their individual course grades and dates of completing the course. Both types of data were retrieved from university-held database enabled by connection to the survey. This course was selected as an example because previous research (Pajarre 2001) has shown it to be one of the most difficult courses in the curriculum. Since the physics course is prescribed for either freshmen or sophomores, the results were divided into two categories: course passed as recommended (i.e., during the first two study years) or course executed later as recommended.

[44] Except those students who already held B.Sc. degree in engineering. After the transition to the two tier degree study programmes at TUT (Autumn 2005) these courses also form part of the B.Sc. degree programme.
[45] The corresponding course Physics Laboratory 1 was also included in these analyses.

Results of the course Physics Laboratory I:

Altogether 96 students (83 % of all respondents in respondent group R_{bg}) had assessed the usefulness and difficulty of the obligatory Physics Laboratory course. Less than half, 41 respondents (43 %) had passed the course in the recommended schedule, either during the first or second study year, and respectively 55 respondents (57 %) had passed the course later than officially recommended.

The distribution of the perceived usefulness of the course and the corresponding average course grades (calculated from individual actual grades[46]) are presented in the Figure 10:

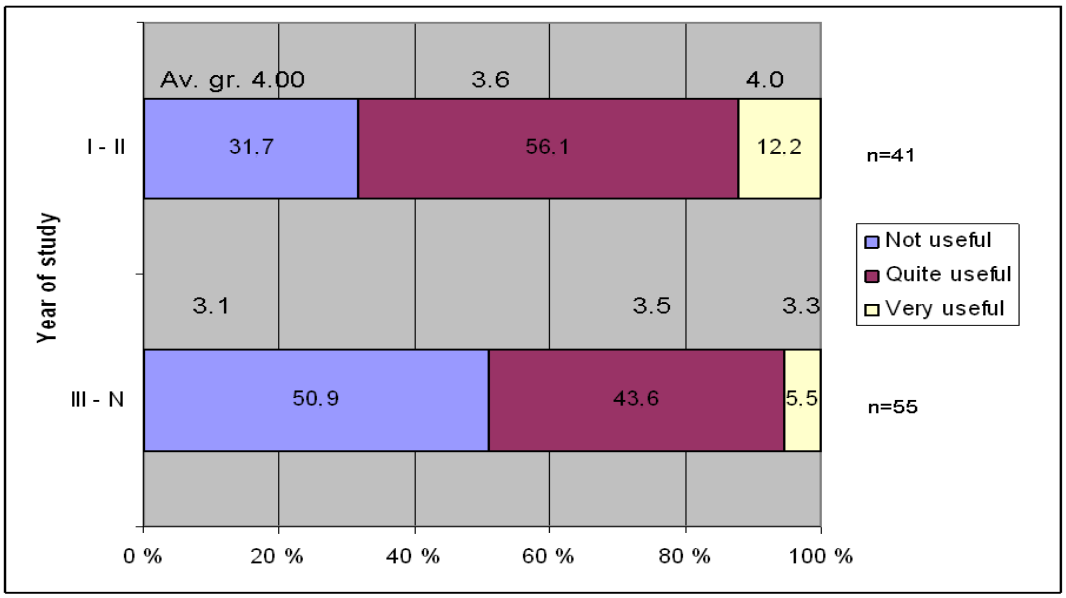


Figure 10. Distribution of respondents' (R_{bg}) opinions on the usefulness of the course, Physics Laboratory I, and the averages of the same students' final grades from that course and the effect of the time of executing the course on the responses.[47]

The results presented in Figure 10 suggest that those students who had passed the course in the recommended schedule were more often satisfied with the course's usefulness than those who had passed the course later than recommended. Less than a third (31.7 %) of the students who passed the course according to the recommended schedule considered the course as "not useful" whereas about half (50.9 %) of the students who passed the course later than recommended gave the course the same estimation. The grades the students received from the

[46] Studies at TUT are evaluated on a scale excellent (5), very good (4), good (3), very satisfactory (2), satisfactory (1) or failed (0). In some cases a scale passed/failed can also be used.
[47] Some respondents have been given a grade "passed" instead of a numeric grade. Their averages have thus not had an effected on the average grades. This should be noted when analysing the utilisability of the results.

course were also consistently higher among those respondents who had passed the course as part of the recommended schedule.

What is also interesting is that among the students who passed the course in the recommended schedule, those respondents who estimated the course as either "not useful" or "very useful" also gained the highest average course grades, i.e., the most successful students valued the course either most or least. Among those students, however, who had passed the course later than recommended, the students who estimated the course moderately as "quite useful" were those who had been most successful on the course (n=24, average course grade 3.5). Overall, the results indicate that those students who passed the course as scheduled, were more satisfied with the courses usefulness and succeeded better on the course than those students who passed the course later than recommended. It can also be seen that among those respondents who passed the course as scheduled, the more successful students on the course did not consider the course any more useful than the less successful students on the course.

Similarly, the respondents' opinions on the difficulty of the course can be analysed by dividing the respondents into two groups: those who passed the course according to the recommended schedule and those who passed the course later than recommended. The results are presented in Figure 11:

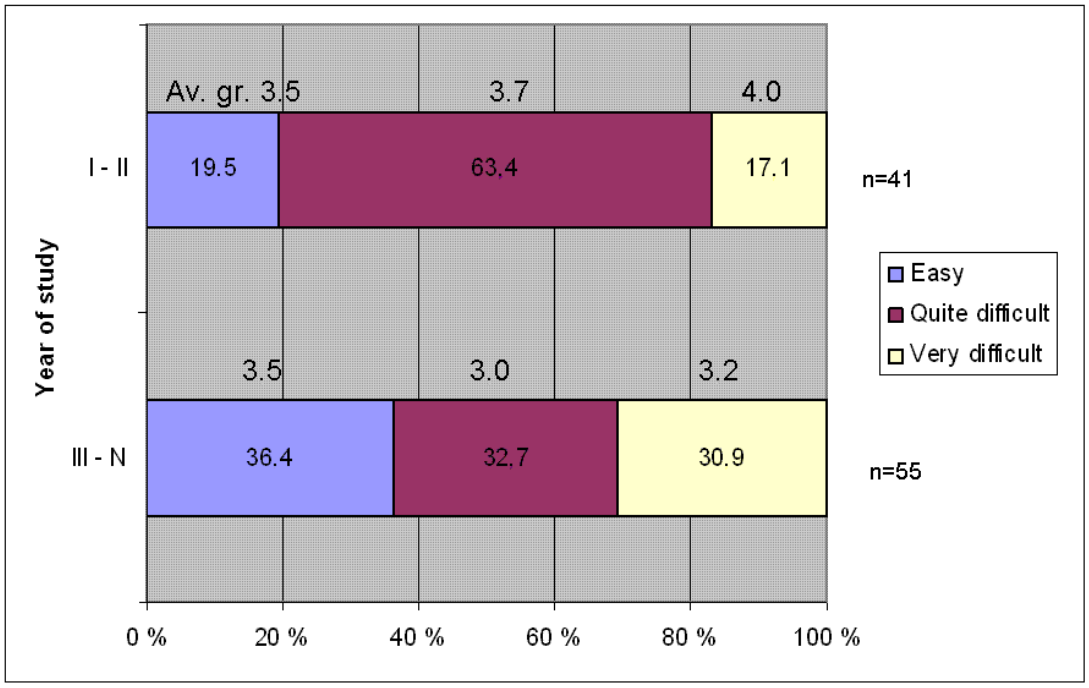


Figure 11. Distribution of respondents' (R_{bg}) opinions on the difficulty of the course, Physics Laboratory I, and the averages of the same students' individual final grades from the same courses. The upper section represents the students who passed the course according to the recommended schedule and the lower section those students who passed the course later than recommended.[48]

[48] Some respondents have been given a grade "passed" instead of a numeral grade. Their averages thus have no effect on the average grades. This should be noted when analysing the utilisability of the results.

The results suggest that among the respondents who passed the course following the recommended schedule, those who estimated the course most difficult were the same who succeeded *best* in the course and those who estimated the course as "easy" were the same who received *lowest* grades from the course. Interestingly, among those students who passed the course later than recommended, the connection between grade and difficulty was quite the reverse. Figure 11 shows that most of the students who passed the course in the recommended schedule estimated the course as "quite difficult" whereas among those respondents who passed the course later than recommended, opinions on the difficulty of the course distributed more evenly.

These results give rise to new questions regarding potential biases: Do those respondents who passed the course as scheduled differ from "average" students? Do they, for example, perform better in all their studies than "average" students? Do they study in different study programmes or is their sex ratio different from the whole respondent group? Since the analysis had been directed only at responses gathered from students who had consented to the use of their background data, a further analysis was also possible. Of those students (R_{bg}) who passed the course, Physics Laboratory I, as scheduled, 86 students' gender, study program and course grades were retrievable from the university held database. Analysis of their background showed that 26 % were female and 74 % male compared to all the respondents (R_{bg}) of whom 27 % were female and 73 % were male. The mean weighted average of those respondents (R_{bg}) who passed this particular course as scheduled was 3.2. This was very near of the mean weighted average of the whole respondent group R_{bg} (3.3) and the same as the mean weighted average for all students graduating during the survey time (3.2). The study programmes of respondents[49] who passed the course as scheduled are presented in Table 9.

[49] M. Sc. (Tech.) students in group R_{bg}. In addition, there was one architecture student who had completed the course and responded to the corresponding question.

Table 9. Percentages of respondents (in group R_{bg}) passing the course, Physics Laboratory I, in the recommended schedule compared to all respondents (R_{bg}).

Study Programme	**Percentages of respondents who passed the course in the recommended schedule, N=86**
Civil Engineering	100 % (*f*=3)
Automation Engineering	88 % (*f*=7)
Industrial Engineering and Management	72 % (*f*=13)
Materials Engineering	70 % (*f*=7)
Science and Engineering	67 % (*f*=4)
Information Technology	65 % (*f*=22)
Electrical Engineering	64 % (*f*=16)
Mechanical Engineering	60 % (*f*=9)
Environmental and Energy Technology	57 % (*f*=4)
Fibre, Textile and Clothing Science	0 % (*f*=0)

The above figures should be treated with caution, since the number of respondents in many study programmes is very small and even one respondent's execution time of the course can alter the percentage significantly. However, in the case of those study programmes where the number of students is larger, such as Electrical Engineering, Information Technology, and Mechanical Engineering, the previous figures can be informative to a certain extent. The more detailed statistics are presented in Appendix 4. The effect of the implementation order of a course is also discussed in Chapter 6.2.3, below along with the analysis of a course Fundamental University Physics Part I (72021).

6.2.3 Respondents' Opinions on the Courses compared with the Respondents' Success in their Studies

To observe in greater detail the potential connections between respondents' (R_{bg}) evaluations and final grades the following analyses were made:

1) Difficulty and usefulness of a course vs. the average course grades of the respondents:

One course was selected to discover if a connection exists between the respondents' (R_{bg}) opinions on the courses' usefulness and difficulty and the respondents' actual grade from the same course. The course selected was Fundamental University Physics Part I (course number 72021) because it was obligatory for all students in engineering programmes[50] and there were thus more student responses on this course in the survey data than from most other courses. From the respondent group R_{bg}, there were 81 students who had executed that particular course and 72 respondents who had also given estimations of the courses difficulty and usefulness. Table 10 presents the distributions of the respondents (R_{bg}) estimations of the course's usefulness and difficulty and the average final course grades of the same students from this course.

Table 10. Distribution of respondents' (R_{bg}) opinions of the course's difficulty and usefulness and respective average course grades (n= 72, N=131).

Perceived usefulness	Perceived difficulty
Not useful: n=16 average course grade 3.0	Easy: n=21 average course grade 4.1
Quite useful: n=42 average course grade 3.2	Quite difficult: n=46 average course grade 3.1
Very useful: n=14 average course grade 4.0	Very difficult: n=5 average course grade 2.0

To analyse the same data further, the respondents' estimations were classified according to the final grades of the respondents and the corresponding estimations of the course's usefulness and difficulty given by the same students. The distribution of the respondents' opinions grouped according to their final grades from the course Fundamental University Physics Part I (72021) are presented in Figure 12 below.

[50] The names and numbers of courses often change making it relatively difficult to find a course having the same name and course number over several years. Thus students beginning their studies in different years undertake the same obligatory studies but often with different course names and numbers. This course was lectured under the same course number for quite a long time, yet there were also respondents who had undertaken a corresponding course instead of this exact one.

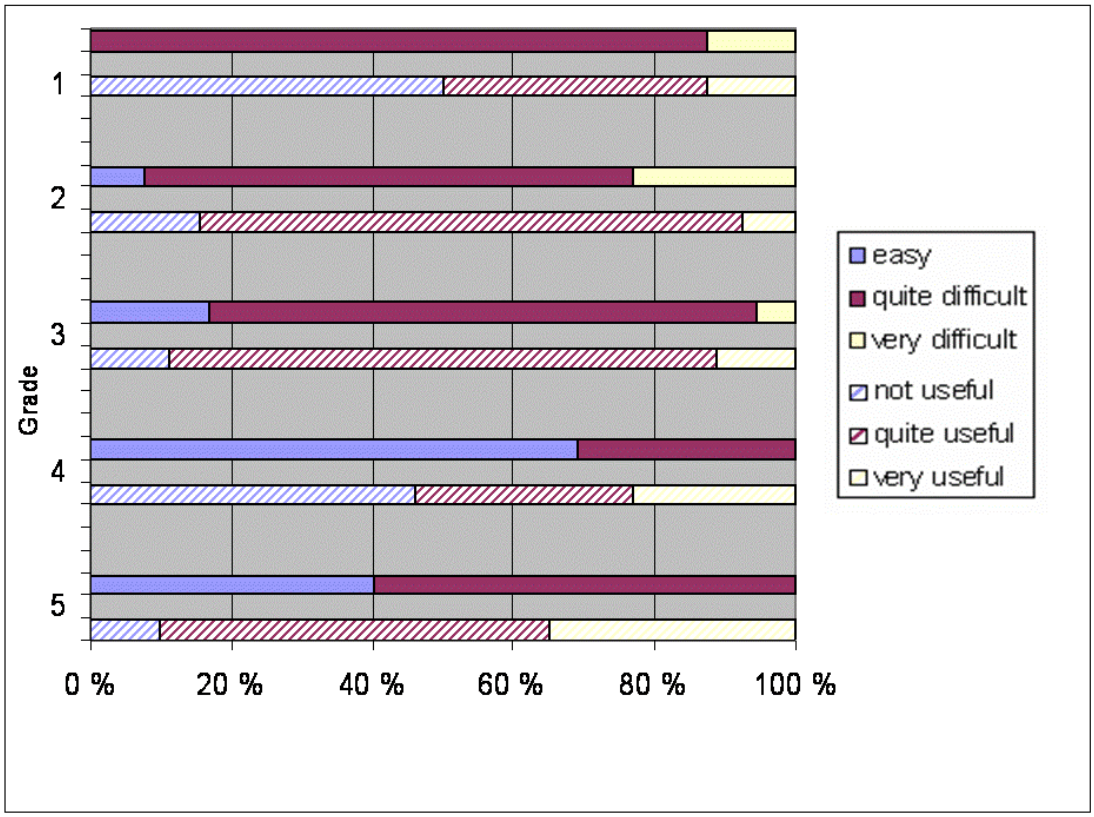


Figure 12. Distribution of the respondents' (Rbg, n=72, N=131) opinions on the course Fundamental University Physics Part I (72021) classed by their final grades from this course.

The results in Figure 12 show the distributions of the respondents' estimations of this particular course. It can be seen that the respondents' opinions on the difficulty of the course change as the grades rise, whereas the responses to the usefulness of the course are much less linked to the grade obtained.

When such results are analysed, it is important to note that the responses in one course are often fragmented so that the number of responses in each category can be very low. For example, in this course the total amount of responses was 72, however, when divided into several subcategories the number of responses in each category is so small that these results should be interpreted with some caution. However, these results demonstrate that performing this kind of analysis is possible. With a longer time period or more responses the results would also be statistically more reliable.

The differences in opinions raised the question as to whether those students who had passed this course within the recommended schedule evaluate the course's usefulness and difficulty

in a different way to those students who had postponed carrying out the course.[51] To answer this, the respondents' (R_{bg}) background data was analysed to determine the time of execution of the course. It was surprising to discover that most of the respondents (93 %, n=75) had passed the course according to the recommended schedule and only 7 per cent (6 respondents) had passed the course later than recommended. The average final grade from the course, Fundamental University Physics Part I, was 3.4 for those respondents who had passed the course as scheduled and 3.3 for those respondents who had passed the course later than recommended. There were, therefore, no notable differences between the respondents' study success on the course, regardless of when they undertook it.[52]

2) Perceived usefulness of the course vs. the mean weighted average of the respondents:

To find whether the more successful students evaluate the courses differently from, those students who performed "poorly", several analyses were made. First, the average grades from all the courses evaluated "not useful", "quite useful" and "very useful" were calculated i.e., for each mention the respondents (R_{bg}) had given to a course the corresponding grade for that same student and the same course was retrieved from the university's database. The results are presented in Table 11[53]:

Table 11. Average course grades from differently assessed courses (concerning courses' usefulness). The total number of given responses = 6379 given by respondent group R_{bg} (n=109, N=131).

Perceived usefulness	**Average final grade from the course**	**Total number of estimations of courses (N=6379)** [54]
Not useful	2.94	1088
Quite useful	3.26	3230
Very useful	3.60	2061

[51] As noted earlier, one of the recognised - or at least expected - problems in many universities is that the students tend to defer taking obligatory courses which is believed to delay graduation.
[52] It should be noted, however, that the survey and background data were gathered so that only the last execution date of the course was analysed. If the respondent had taken the course (or examination) more than once (which occurs if a student attempts to raise his/her grade from the course) only the last date was analysed.
[53] These data cover all executed courses, including the M.Sc. theses of those students who had it executed by the time of responding to the survey.
[54] The figure is the total sum of all respondents (Rbg) responses of the courses they had performed. The actual total amount of *different* courses evaluated in this survey was 1390.

Analysis these results can lead to at least two different interpretations:

1. the better the grade received from the course the higher the respondents' opinion of the same course's usefulness

2. only 17 % of the courses (n=1088) were estimated as "not useful", i.e., most of the courses were considered either "quite useful" or "very useful".

The previous results demonstrate how employing plain averages to assess the validity of responses can raise more questions than answers. Even though the previous results were obtained by gathering individual grades from a database, since they were used only as averages they revealed little profound information.

To analyse *the same responses* in greater depth, the responses (of the respondent group R_{bg}) were divided into three groups according to what had been the most typical estimation by each respondent when responding to the survey. For each group, the mean average of the respondents' weighted averages (from all courses completed) were then calculated to see whether those students who had most often been very satisfied with the courses' usefulness had on average gained better grades than those students who were more critical with the courses' usefulness. The following results were obtained:

- Of those respondents (n=7) who most often judged the courses in the survey as "not useful", the mean weighted average from all courses was 3.57.

- Of those respondents (n=78) who most often judged the usefulness of the courses as "quite useful", the mean weighted average from all courses was 3.27.

- Of those respondents (n=24) who most often judged the courses as "very useful", the mean weighted average from all courses was 3.34.

These results show, surprisingly that the highest mean weighted average was for those students who had judged most of the courses they had completed as not useful at all. The lowest mean weighted average was for the student group in which the most typical estimation of the courses' usefulness had been "quite useful".

Comparing these two result sets makes it easy to understand why the results of traditional and conventional student ratings are often viewed with suspicion and their validity questioned. All the above results were calculated from the same respondents and the same background information. However, it is possible to interpret these results in different ways: either the successful students also gave high value of the course's usefulness (the first interpretation) or conversely, the most successful students are those who value the courses' usefulness least

(the second interpretation). This discrepancy, however, does not imply the results were invalid, but rather underlines, how easily potential misinterpretations can arise if the results are not analysed thoroughly.

With traditional student evaluation systems, where there either is no background information available or where the background information is gathered from average university statistics and mathematically analysed, there is a high risk of misinterpretation in analysing the results. In the present analysis, the first results were gained by grouping the students' responses and corresponding final grades into three categories according to the courses' usefulness. The latter results, for their part, are gained according to the most typical estimation of each respondent and each respondent's grade point average. Together these two result sets show that there were *seven* students who judged the usefulness of the courses very critically while also succeeding excellently in all the courses they studied. On average, however, *most of* the students have been at least quite satisfied with the usefulness of the courses they have completed. Observations at this level of detail would not have been possible without the direct linkage between individual responses and individual background information.

3) Perceived difficulty of the course vs. the mean weighted average of the respondents:

Similarly, the potential interdependence between the students' course grade and the course's perceived difficulty can also be analysed. Following the same procedure as in the previous results, the average grades of all the courses evaluated as "easy", "quite difficult" and "very difficult" were calculated, i.e., for each mention the respondents (R_{bg}) had given to a course the corresponding grades from the same students and the same course were retrieved from the university's database. The results are shown below in Table 12.

Table 12. Average course grades of differently assessed courses (concerning courses' difficulty) in the respondent group R_{bg} (n= 111, N=131).

Perceived difficulty	**Average final grade for the course**	**Total number of estimated courses (N=6377)**
Easy	3.70	2339
Quite difficult	3.27	3117
Very difficult	2.68	921

Analysing the previous results can lead to the interpretation that there is a somewhat linear dependence between the students' final grade for a course and the same students' estimation of the courses' difficulty. The easier the courses have been estimated, the higher have been the students' final grades from the same courses and vice versa. Of all the course evaluations given, 14 % (n=921) were considered as "very difficult", and the remainder 86 % (n= 5456) as either "quite useful" or "very useful".

As with the analysis of the usefulness, the same responses were also divided into three groups according to the most typical estimation made by each student responding to the survey. The averages of each three respondent groups' grade point averages (for all courses undertaken) were calculated to see whether those students who had most often considered the courses "easy" had, on average, gained better grades than those students who had estimated the courses as "quite difficult" or "very difficult". The following are the results:

- There was only one respondent in the respondent group R_{bg} (n=1) who had given the estimation "very difficult" as the most typical alternative to describe the difficulty of the courses. The weighted average of all courses for that respondent was 2.72.

- Of those respondents (n=71) whose most often presented estimation of the courses was "quite difficult" the mean weighted average for all courses was 3.33.

- Of those respondents (n=39) whose most often presented estimation of the courses was "easy" the mean weighted average for all courses was 3.27.

According to the above results, on average the highest weighted *averages* were found for those students who have mainly considered the courses as "quite difficult", rather than "easy". This, again, contradicts the earlier observation that "the easier the course, the higher the final grades". However, as with previous results concerning the usefulness of the courses, these results also reflect the differences when results are analysed using different *categories*[55]. By linking these two result sets it can be noticed that a majority of the respondents had given the estimation "quite difficult" as the most typical alternative to describe the courses' difficulty. On average, students have received higher grades for the courses they have estimated as "easy" than for other courses. The mean weighted average of those students who considered most of the courses as "easy" was lower than of those students who considered most of the courses as "quite difficult".

These results are similar to Centra's (2003) findings on student results and the students' experience of course workload. According to Centra, the course grade the students expect to gain (the actual final grades were not accessible) and their estimations of the course's

[55] Respondents' grades from the course vs. respondents' grade point averages from all courses completed.

workload did not correlate. On the other hand, according to the author, the students had been most satisfied with the courses where the workload had been appropriate, not with those courses which they had considered either too easy or too difficult.

6.2.4 Respondents' Satisfaction with their Major Subject vs. the Same Respondents' Overall Success in their Studies and Duration of their Studies

As Tynjälä et al. (2004, pp. 10–11) remark, using students' grade points as the sole indicator to measure their study success is questionable because the learning results may vary according to the way in which the teacher has evaluated the students' learning results. To avoid this, the students' study progress was also analysed as an indicator (i.e., the cumulative amount of credit units as a function of time); the same indicator as that used in the study of Tynjälä et al. (2004).

Altogether 152 students (33 %, R_{all}) had responded to the question "What would be your choice, if you were to select your major subject now?" The given responses were distributed so that 77 % of the respondents (n=117) were satisfied with the major subject they had originally chosen, 18 % (n=27) were dissatisfied with their current major subject and 5 % of the responses (n=8) were unclear.

Of those respondents who had consented to analysis of their background data (R_{bg}) 116 students (86 %) responded to the question. Of these, 78 % (n=90) were satisfied with the major subject they had originally chosen, 16 % (n=18) were not satisfied and 7 % (n=8) of the responses were unclear.

The linkage between the survey and the respondents' (R_{bg}) background data made it possible to analyse whether there were differences in study success (as measured by course grades) or study progress (the duration of studies) between those respondents who had been satisfied with their major subject and those respondents who had been dissatisfied. Table 13 presents the distribution of the results:

Table 13. Distribution of respondents' in response group R_{bg} (116 responses) mean weighted averages and study years and the respondents' satisfaction with their major subject.

	Satisfied with current major subject	**Unsatisfied with current major subject**
Gender		
- male	84 % (n=62)	16 % (n=12)
- female	81 % (n=22)	19 % (n=5)
Mean weighted average	3.4 (n=84)	3.1 (n=17)
Duration of studies (years)	5.6 (n=90)	6.4 (n=18)

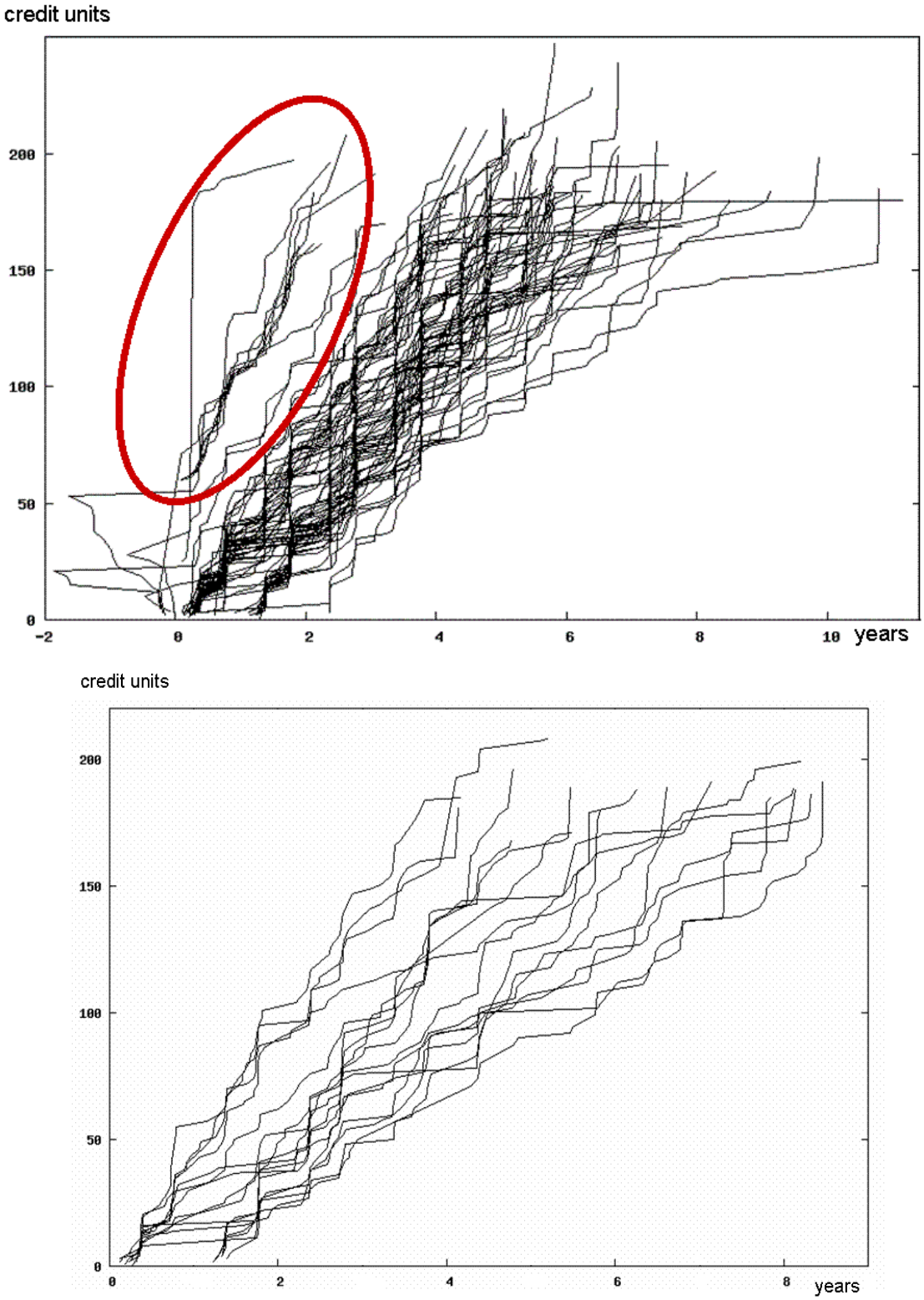
Table 13 shows that the mean weighted average of those respondents who were satisfied with their major subject was 3.4 and for those students who had been dissatisfied, the figure was 3.1. The average study time of those respondents' who were satisfied with their major subject had been 5.6 years and the corresponding average study time of those being dissatisfied with their major subject had been 6.4 years.[56] These results support the findings of Kärkkäinen (2005, p. 60) which show a dependence between students' satisfaction with their study programme and the duration of their studies.

More detailed analysis of the previous results suggests that the respondents' earlier education should also be taken into consideration. When the respondents were categorised into two groups by their previous education[57], i.e. those already possessing a B.Sc. level degree at the start of their studies in TUT and those with no previous technical education, the differences in results were markedly smaller.

Figures 13 and 14 below demonstrate the differences between respondents' (satisfied and unsatisfied with their current major subject) cumulative number of credit units as a function of time. When those students already holding the B.Sc. level degree are eliminated from the analysis (represented in Figure 13 by the lines which start with 60 credit units at the year 0) are eliminated, the differences between Figure 13 and Figure 14) are markedly reduced.

[56] Since these analyses were made during the year 2005, three years later than the actual survey, it was possible to calculate the total length of the respondents study time, not just the length of studies at the time of responding to the survey.

[57] It was possible to detect engineers from other respondents in group R_{bg}, since the admission criteria into TUT was visible in OPREK database. It could also have been possible to insert the question of earlier background into the survey page, but that would have made the questionnaire longer. Retrieval from the database also prevented erroneous information being collected on respondents' backgrounds.



Figures13 and 14. Distribution of respondents (R_{bg}, n=90) who were satisfied (upper figure) and respondents (R_{bg}, n=18) who were not satisfied (lower figure) with their major subject cumulative credit units as a function of time.

Figures 13 and 14 also demonstrate clearly that those respondents (R_{bg}) who already possessed a B.Sc. level degree were satisfied with their major subject. This finding supports the view of Tynjälä et al. (2004) according to which the student's existing knowledge also affects the student's learning process. In this case the results are probably because those respondents who had earlier studied engineering subjects elsewhere were much more likely to know what they wanted to study in TUT than, for example, those students who had started their studies at TUT directly after leaving upper secondary school.

6.2.5 Perceived Usefulness and Difficulty of the Same Course

To analyse how all the respondents perceived the usefulness of the *same* courses *and* their level difficulty, the responses were collected in the form shown in Table 14 below.

Table 14. Distribution of respondents' estimations of the same courses' usefulness and difficulty. Total number of students responding to the question = 184, total number of evaluated courses = 1390 and total number of evaluations of courses = 10057. N=455.

Respondent group		**very useful**	**quite useful**	**not useful**
Respondents consenting to analysis of their background data, (R_{bg}, n=115)	**very difficult**	4.8 %	6.8 %	2.8 %
	quite difficult	18.2 %	24.6 %	6.1 %
	easy	9.3 %	19.3 %	8.1 %
				Σ 100 %
Respondents who denied analysis of their background data, (R_{den}, n=69)	**very difficult**	4.3 %	5.9 %	3.8 %
	quite difficult	17. 0 %	23.8 %	7.0 %
	easy	9.3 %	18.2 %	10,6 %
				Σ 100 %

The same results are depicted graphically in Figure 15 and Figure 16. Figure 15 presents the distribution of estimations given by respondents allowing the use of their background data.

Figure 16 presents the distribution of the estimations of the same courses' usefulness and difficulty given by those respondents who denied access to their background data.

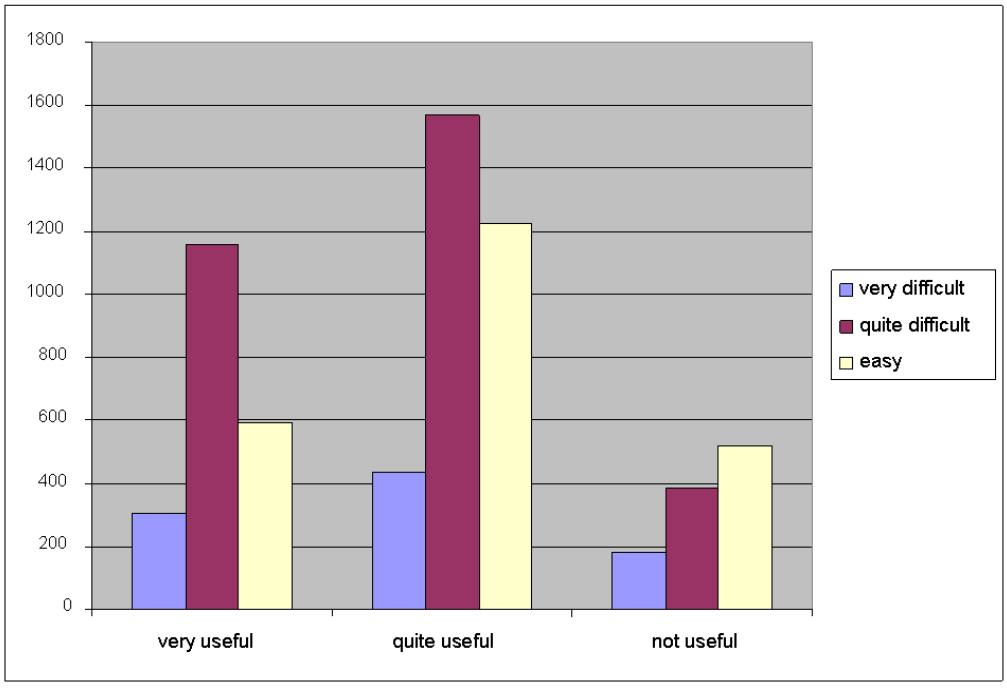


Figure 15. Perceived usefulness and difficulty of the same course (respondent group R_{bg}, n=115).

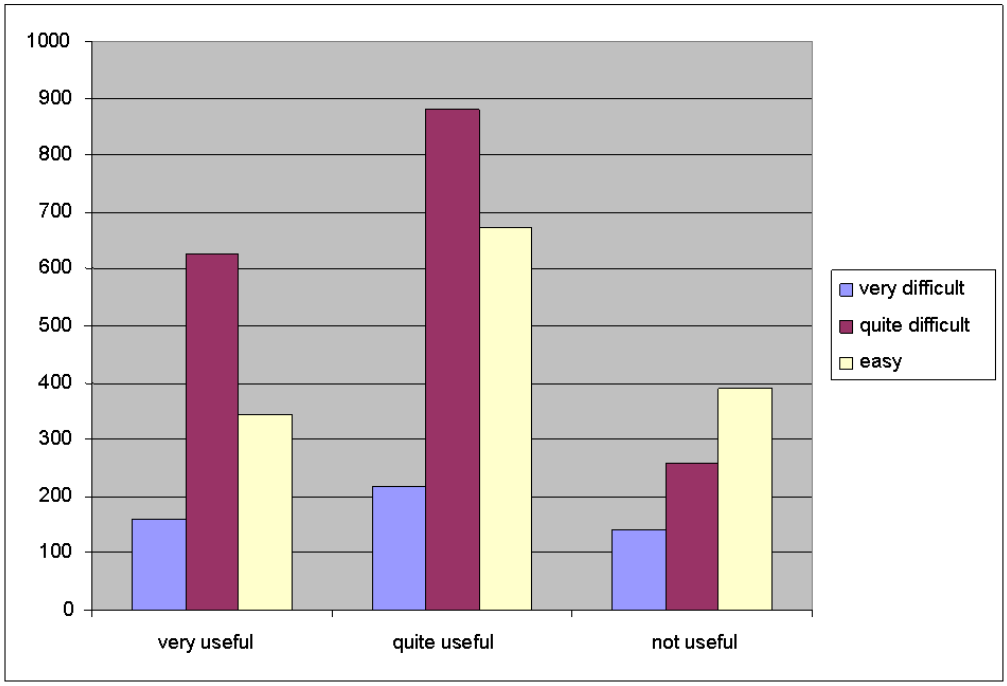


Figure 16. Usefulness and difficulty of the same course estimated by respondent group R_{den} (n=69).

From Figures 15 and 16 it can be seen that the respondents (in both groups, R_{bg} and R_{den}) did not consider only those courses useful which they had estimated as easiest.[58] There were also courses which were perceived as being both "very useful" and as "very difficult" and courses which were perceived as being both "easy" and as "not useful" as well as other combinations.

Altogether there were 465 responses concerning 253 courses which were perceived as being simultaneously "very useful" and "very difficult". The most frequently mentioned courses perceived as being both very useful and very difficult are presented in Appendix 4.

It should be noted, however, that the sheer number of responses to a certain course is insufficient to form an overview of which courses have been considered simultaneously very useful and very difficult (or any other combination). The number of students participating in a course varies significantly from course to a course. If only the number of responses is considered, just those courses having a large number of students will be noticed. A better approach would be to analyse the proportion of "very useful" and "very difficult" responses for each course. In the present study, however, the 465 "very useful" & "very difficult" responses were so evenly distributed among 253 courses that only the most voluminous courses were listed.

6.2.6 Perceived Usefulness, Difficulty and the Actual Grade Received for the Same Course

To further analyse the respondents' opinions on the usefulness and difficulty of the courses undertaken, the final course grades of the respondents (R_{bg}) were included in the analysis. To see what kind of final course grades the respondents had achieved in the courses for which they had estimated usefulness and difficulty, all the given course estimations (n=6306) which could be connected to corresponding final grades were analysed. Figure 17 shows the distribution of the respondents' final grades in each usefulness/difficulty category:

[58] A suspicion often voided is that students evaluate highly only those courses in which they expect to succeed well, and for students to succeed well the lecturers are tempted to make the courses easier (discussed in more detail in e.g., Parjanen 2003, pp. 15–16).

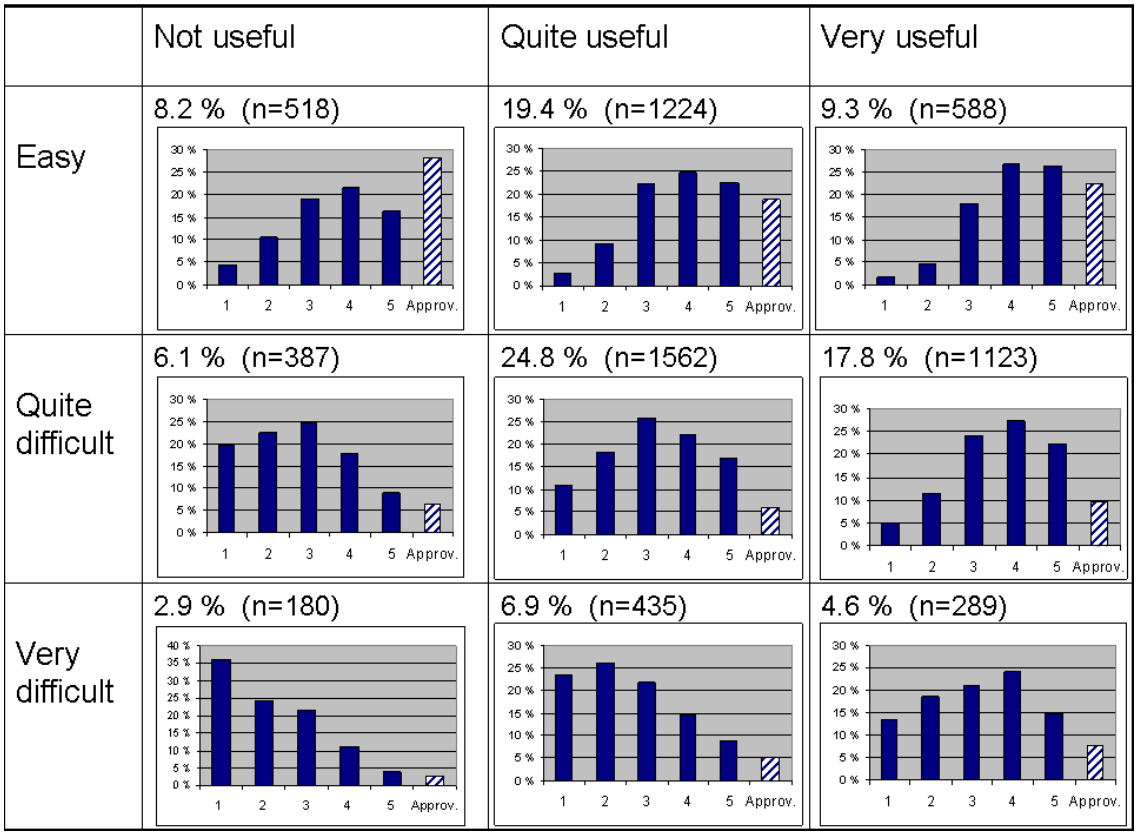


Figure 17.Distribution of the respondents' (R_{bg}) course specific estimation of the usefulness and difficulty of the course and the respective final grade from the same course (total number of course estimations was 6306).

From Figure 17 it can be seen that in each category the final grades distributed quite differently. For example, when those courses which the students had estimated to have been both "very difficult" and "not useful" (less than 3 percent of all responses), the most frequently gained final grade had been "1". For those courses which had been most frequently estimated as both "not useful" and "easy" the most typical final grade had not been numerical, but a simple "Approved".

The Figure 17 demonstrates a novel approach to assessing the validity of student evaluations. Earlier research work has been based on attempts to purposefully fit retrieved background information based on class averages or estimations connected to individual student responses. Analysing the validity of such data without the use of complex statistical methods has been impossible. As can be seen in Figure 17, linking online survey results directly to the corresponding individual background data makes it possible to analyse any survey responses at the individual level. For an individual teacher such simple deviations as demonstrated in Figure 17 may suffice to make conclusions on the validity of responses gathered for his/ her

own course. For more demanding analyses the same statistical algorithms can be used as earlier, but being based on actual data, not on averages or assumptions.

Exact figures for the distributions are presented in Appendix 4.

6.2.7 Amount of Work Experience and the Respondents' Average Course Grades and Duration of Studies

Altogether 174 students had responded to the question on their work experience during their studies. Of these 130 consented to access to their data and 34 denied it. There were seven responses which were unclear so that the final number of analysed responses was 167. Of these, 125 were from students who had allowed analysis of their background data.

The total estimated amount of work experience of all respondents (R_{all}) was an average of 23 months work experience supporting their M.Sc. studies and 12 months other work experience. The distribution of the respondents' work experience into work experience supporting studies and other studies is presented in Table 15.

Table 15. Distribution of work experience by the type of work.

Respondent group	**Work experience supporting M.Sc. studies > other work experience**	**Work experience supporting M.Sc. studies < other work experience**
R_{all}, n=167	71 %	29 %

The average work experience of those students not already holding the B.Sc. degree (n= 160, calculated from all responses, R_{all}) was 23 months work experience supporting M.Sc. studies and an average of 12 months other work experience.

Of those students who had allowed the access to their background data (R_{bg}), the cumulated work experience during their studies is shown in the distribution presented in Table 16:

Table 16. Distribution of work experience by type of work (respondent group R_{bg}).

Respondent group	**Work experience supporting M.Sc. studies > other work exp.**	**Work experience supporting M.Sc. studies < other work exp.**
R_{bg} (total)	76 %	24 %
R_{bg} (no previous engineering educ.)	73 %	27 %

The estimated average amount of work experience in respondent group R_{bg} was 24 months supporting M. Sc. studies and 10 months other work experience. The average duration of studies of those respondents who had more work experience supporting M.Sc. studies than other work experience was 5.9 years. Their mean weighted average from all executed courses (excluding the M. Sc. thesis) was 3.3. The average duration of studies of those respondents with more other work experience than work experience supporting their M.Sc. studies was 5.4 years. Their mean weighted average from all courses completed (excluding the M.Sc. thesis) was 3.2.

To ascertain if the duration of work time during studies affected the duration of the respondents' studies or success in their studies, the respondents ($R_{bg,}$ n=110) estimations of the length of their work experience were compared with the actual duration of their studies and their grade point averages (both retrieved from the university database). The results are shown in Figure 18 and Figure 19.

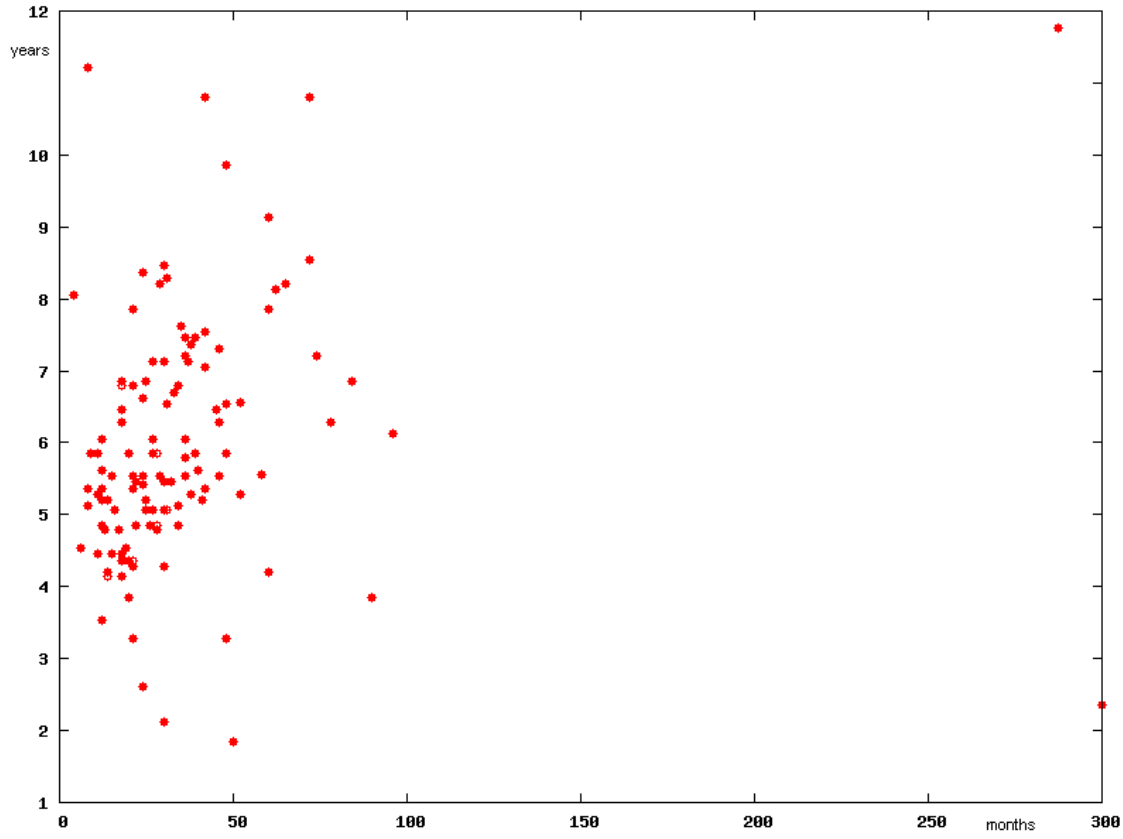


Figure 18. The deviation of respondents' ($R_{bg,}$ n=110) cumulative work experience (in months) during their studies and the duration of their studies (in years).

A commonly presented claim is that students' working during their studies will prolong the duration of their studies (discussed e.g., by Kärkkäinen 2005, pp. 14–15). The results in Figure 18 suggest that this connection is not straightforward. There are respondents with little work experience graduating both quickly and slowly as well as respondents with over 50 months of work experience also graduating both fast and slowly.

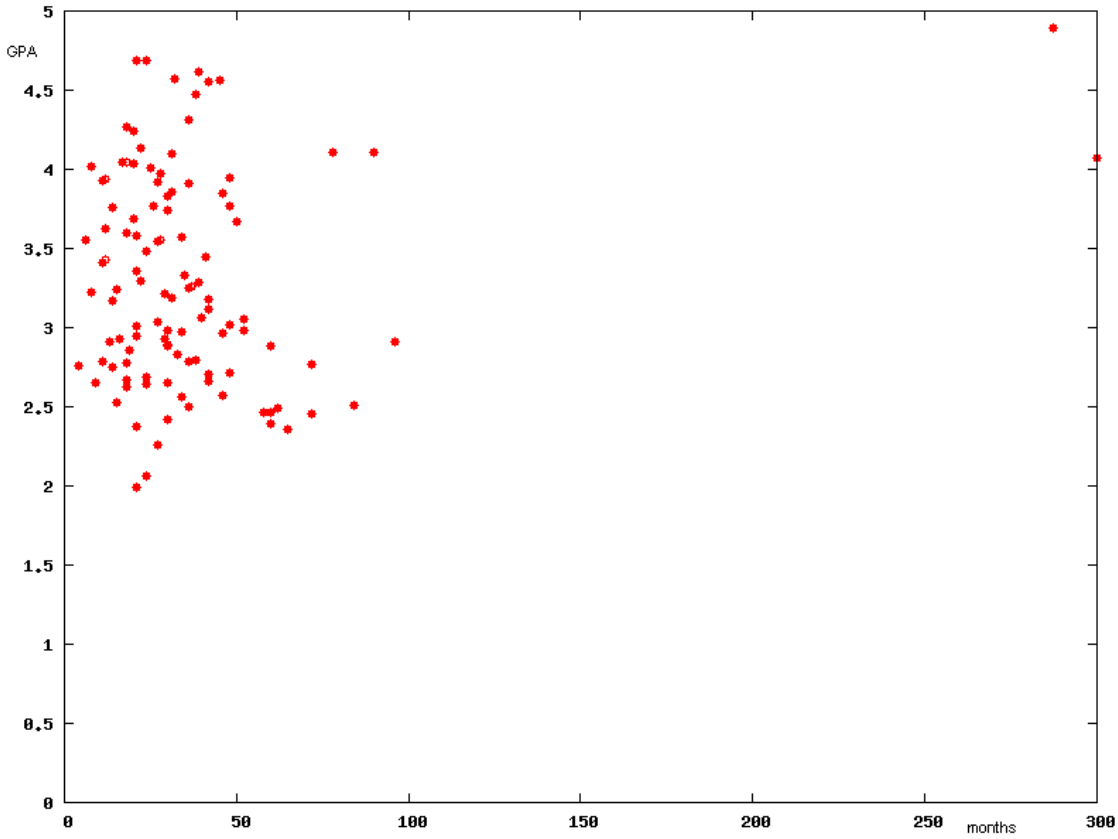


Figure 19. The deviation of respondents' ($R_{bg,}$ n=110) cumulative work experience (in months) during their studies and respondents' grade point average.

The results presented in Figure 19 suggest as in the case of duration of the respondents' studies that there are students with little work experience with both low and high grade point averages and also students with much work experience with both low and high grade point averages. To make reliable statistical analyses of whether there are correlations between these variables, more responses would be required. However Figures 18 and 19 demonstrate the fact that each student is unique, and that it may be risky to generalise their behaviour. This may be overlooked if only statistical figures are taken into account.

6.2.8 Analysis of Factors Impeding the Respondents' Study Progress

The second part of the online survey focused on charting potential factors which might impede or hinder students' progress in their studies. In the first question the respondents were asked to assess whether each of the following items had delayed their progress in their studies or not[59]:

- schedules (e.g., lectures held mostly only on Tuesdays, Wednesdays and Thursdays, or other overlapping issues)
- exams, delays in receiving results from exams, delays in receiving feedback from exercises
- lack of or insufficiency in student counselling
- accessibility of lecturers
- availability of other services
- gaining recognition for studies undertaken at other universities for the M.Sc. degree of Tampere University of Technology
- other issues

In contrast to the first part of the survey which focused on individual courses, the focus here is on the respondents' curricula as a whole. Altogether 176 students evaluated at least one item in the question. The distribution of students' (R_{all}) responses is presented in Table 17.

[59] The original idea was to gather results in a 3-point scale (did not hinder, hindered a little, hindered a lot); however the final data was received in a form hindered/did not hinder.

Table 17. Respondents' ($R_{all,}$ n=176, N=455) opinions of factors hindering progress in their studies at Tampere University of Technology (presented as frequencies).

Factor hindering student progress in their studies	**did hinder, frequency**	**did NOT hinder, frequency**	**not answered, frequency**
Schedules	128	48	279
Exams, delays in receiving results of exams, …	83	93	279
Lack of or insufficiency in student counselling	53	123	279
Reachability of lecturers	38	137	280
Availability of other services	19	156	280
Receiving approval for studies undertaken in other universities…	16	132	307
Other issues	36	95	324

Table 16 shows that schedules was the only item which the majority of respondents (who responded to the question) estimated as having hindered progress in their studies. None of the other given items were considered by the majority of respondents to have hindered their studies.

To examine the issue further, the responses of those students who had responded to this particular question were divided into two groups; responses of those students who allowed access to their background data and the responses of those students who denied it. The distribution of opinions is presented in Figure 18. The exact figures for the distribution of these results are presented in Appendix 4.

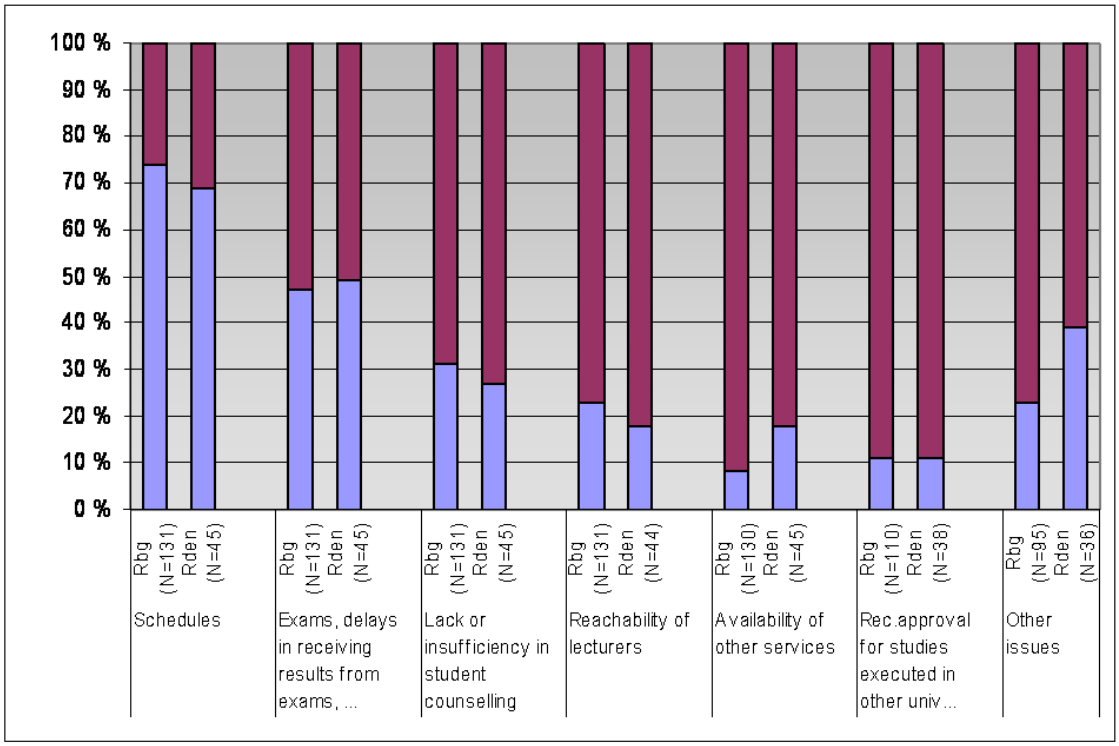


Figure 20. Distribution of responses from groups R_{bg} and R_{den} on potential factors hindering progress in their studies.

As Table 17 and Figure 20 show, schedules were the most frequently mentioned item which the respondents cited as having hindered their studies. Regardless of whether or not the respondents had allowed access to their background data schedules were considered to have hindered the respondents' studies in more than half of the responses. In contrast, none of the other items was considered in the majority of given responses to have been a factor hindering their studies. Those respondents who had consented to access of their background data were more likely to consider schedules, lack of student counselling and reachability of lecturers as having hindered their studies than respondents who denied such access. At the same time they reported fewer problems with exams and delays in receiving results from exams, with availability of other services or with other issues. In both respondent groups 11 % had experienced problems with gaining recognition for studies undertaken at other universities for their M.Sc. degree in TUT.

Factors hindering study progress and respondents' gender

To see whether there were differences between the opinions of male and female respondents as to the factors which had hindered progress in their studies, the results (in respondent group R_{bg}) were divided into two groups according to gender. Figure 21 presents the percentage of reports of each hindering factor for both female and male respondents. Exact figures of the analysis for these factors by respondents' gender are presented in Appendix 4.

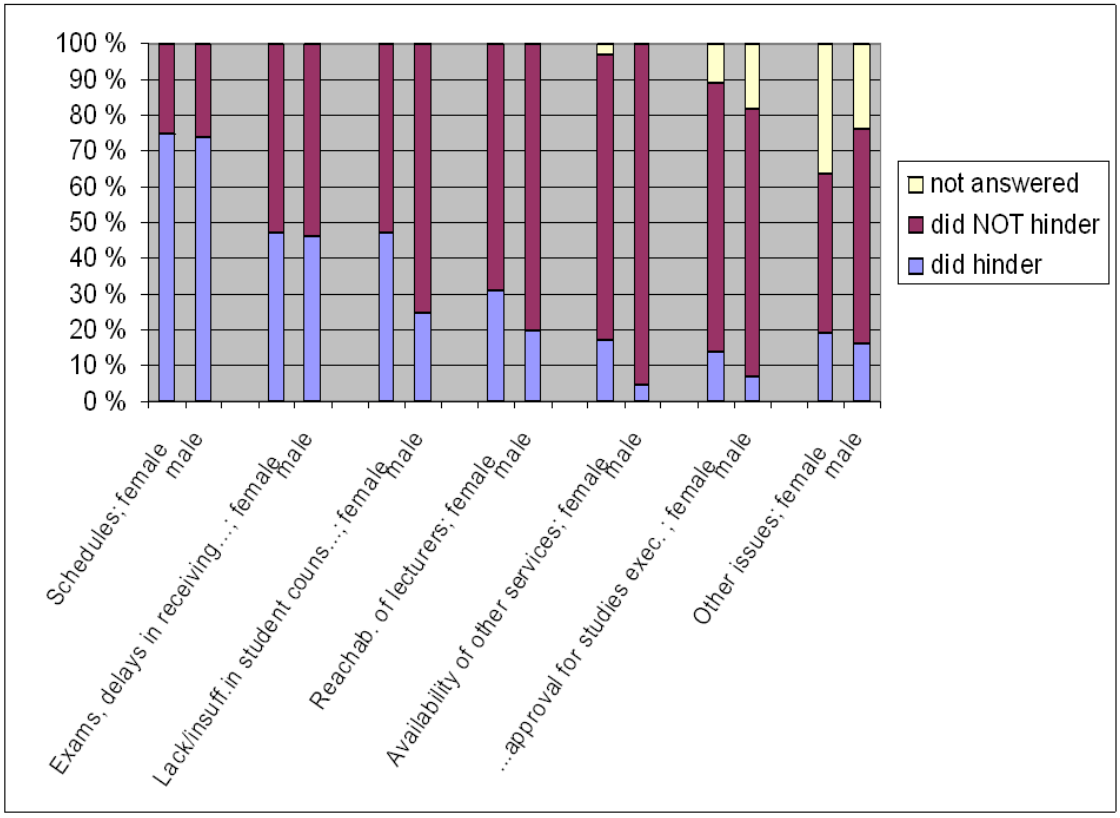


Figure 21. Factors hindering respondents' (R_{bg}) progress in their studies, analysed by the respondents' gender.

With the two most frequently mentioned hindering factors, schedules and exams and results of exams, there were practically no differences between female and male responses. Nonetheless, there were differences of more than 10 per cent between their responses regarding the sufficiency of student counselling, accessibility of lecturers and availability of other services, with female students making greater mention of these items.

Factors hindering the progress of studies and the respondents' mean weighted averages and duration of studies

The next step was to analyse in greater depth the background characteristics of the respondents compared with their opinions on the factors hindering their studies. Figure 20 presents the mean weighted averages[60] of those respondents who considered that the particular factor had hindered progress in their studies and those respondents who did not consider the same factor had hindered their study progress. All respondents belong to the respondent group R_{bg}.

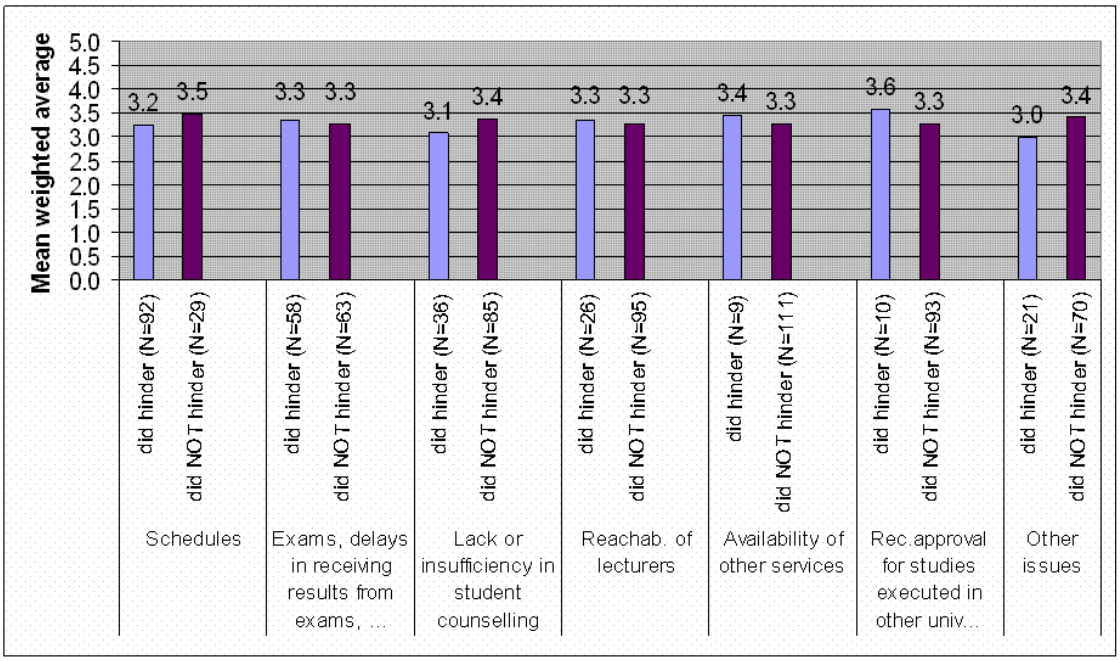


Figure 22. Respondents' (R_{bg}) opinions of whether certain factors had hindered their studies and the mean weighted average of the same respondents.

From Figure 22 it can be seen that those students who had experienced problems with schedules, with adequate student counselling or with "other issues" had succeeded less well in their studies than those students who did not consider such factors as hindrances. Exams and delays in receiving exam results as well as the accessibility of lecturers were factors where the mean weighted average was the same for both respondent groups.

The background analysis was also performed to ascertain how the respondents' experiences of factors hindering their studies correlated with the average duration of their studies. The results of this analysis are presented in Figure 23.

[60] Excluding M.Sc. theses.

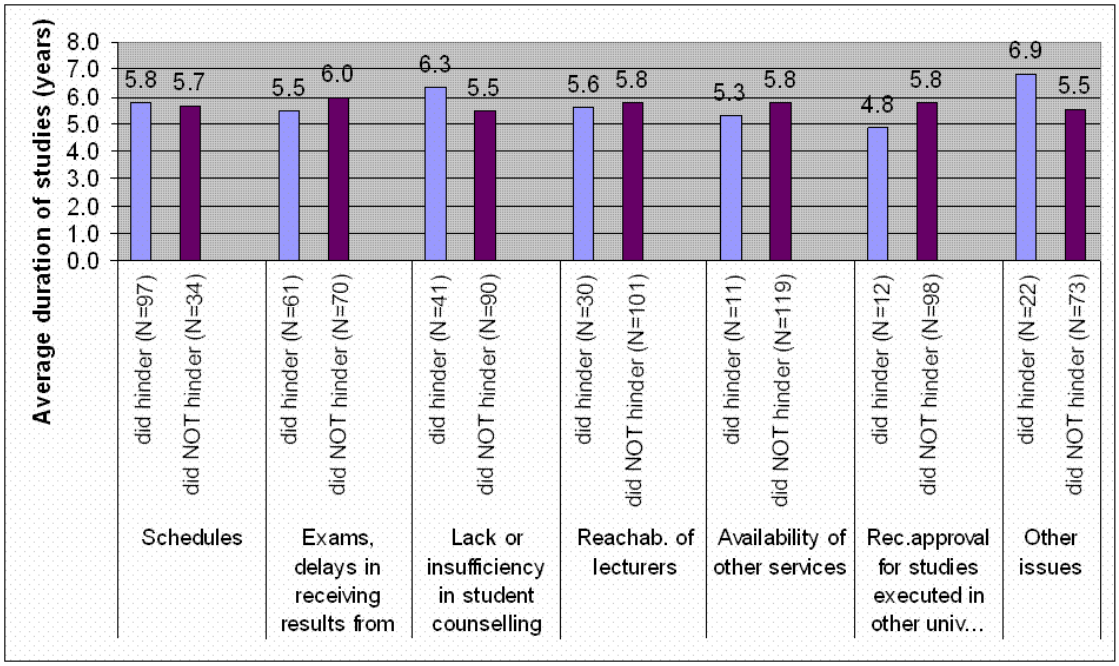


Figure 23. The respondents' (R_{bg}) opinions of whether certain factors had hindered their studies and the average duration of studies in years for the same respondents.

The results of Figure 21 show interestingly that those respondents who reported problems with receiving approval for studies undertaken at other universities had graduated on average faster than other respondents.[61] To analyse this in more detail, the number of credit units obtained at other universities and approved as a part of M.Sc. degree in TUT were analysed for each respondent (R_{bg}). Those respondents[62] who estimated to have encountered problems had an average of 14.3 credit units gained from other universities and approved in TUT. Of those respondents who had not encountered problems, 32 % (n=31) had had an average of 17.4 credit units studies gained elsewhere and approved in TUT. The remaining respondents who had not encountered problems, 68 % (n=67) had not included any courses taken outside TUT in their M.Sc. studies.

In addition, "exams, delays in receiving results from exams etc." as well as "availability of other services" were factors where those respondents who reported problems had graduated faster than those respondents who had not reported problems. The longest study times were found for those respondents who had encountered problems with "other issues".

[61] It should be noted that not all students undertake courses in other universities and thus do not face this problem. It is also possible (although not investigated in this study) that those respondents who had had problems receiving approval for their other studies had been more ambitious than "average" students and thus also graduated faster, in spite of the problems encountered.

[62] There were 4 respondents who had not responded to the course survey part, so the average could be calculated only for those 8 respondents whose credit units were available and the average result should thus be treated with caution.

Since some respondents had written in open-ended comments that going to work while doing their studies had hindered their academic progress, the average work experience of those respondents who had marked "other issues" as a hindering factor was calculated and compared with the average work experience of those respondents who had not reported having had problems with "other issues". The calculations showed that those respondents (n=21)[63] who had mentioned problems with "other issues" had an average of 50 months work experience (34 months' work experience supporting their studies and 16 months of other work experience). Those respondents (n=71) who had not reported problems with "other issues" had an average of 31 months' work experience, which was included 22 months' work experience supporting their studies and 9 months' other work experience. Thus it can be seen that those respondents who had reported problems with "other issues" had had considerably more work experience than those students who had not reported such problems.

The connection between the duration of the studies and the amount of work experience does not itself, however, explain the possible causality between the variables. It is possible that those respondents who were doing more work while studying also studied longer because of the time spent in their jobs. However, it is also possible that those respondents were just studying at a slower pace and therefore actually had more time to take jobs as well. In those cases, though, where the respondent had specifically written an open-ended comment that working while studying delayed their graduation, it can be considered as an explanatory factor.

6.2.9 Satisfaction with the Opportunity to Study Foreign Languages

In the general part of the survey the students were asked to estimate whether they had had the chance to study foreign languages adequately. There were 202 respondents (44 %) who had answered the question, of whom 130 (male 94 and female 36) had allowed access to their background data. Altogether 71 % (n=144, R_{all}) were satisfied with their possibility to study foreign languages. Of those students who had consented to access to their background data 80 % were satisfied with the possibility.

The background analysis (R_{bg}) showed that those respondents who had been satisfied with the amount of their language studies had an average of 8.6 credit units in foreign languages, of which the average grade was 3.3. Of those respondents who would have liked to study more foreign languages, the average amount of executed language courses was 8.0 and the average grade 3.3. This was exactly the same as with those who had been satisfied with the amount of their foreign language courses. These results seem to indicate that the respondents'

[63] Altogether 22 respondents reported "other issues" of whom 21 respondents had mentioned the amount of their work experience.

eagerness to study foreign languages is not directly dependent on the respondents' language skills, if measured by the final grades they had received. The fact that those students who would have liked to study more foreign languages had actually studied them *less* than other respondents was somewhat surprising. It is possible that their individual schedules have prevented them from studying more languages; however, making more profound conclusions would also require more specific questions concerning this issue.

6.2.10 Respondents' Readiness to Make Open-ended Comments

In the survey's general part the respondents were asked to give verbal feedback in two questions:

- to give examples of factors which hindered them progressing in their studies
- a general question "What else would you like to bring out about your experiences of studying at TUT?"

Altogether 80 % (n=105) of those respondents who allowed access to their background data had responded to one or both of the above questions. Of those respondents who had denied access to their background data, only 10 % (n=32) had responded to one or both of the open-ended questions.

Many of the responses for the latter, general question, contained both positive and negative comments or suggestions for improvement. Those 80 respondents who had allowed the access to their background data had given a total of 102 open-ended comments and those 23 respondents who had denied the access to their background data had given 32 comments. The distribution of comments is presented in Figure 24.

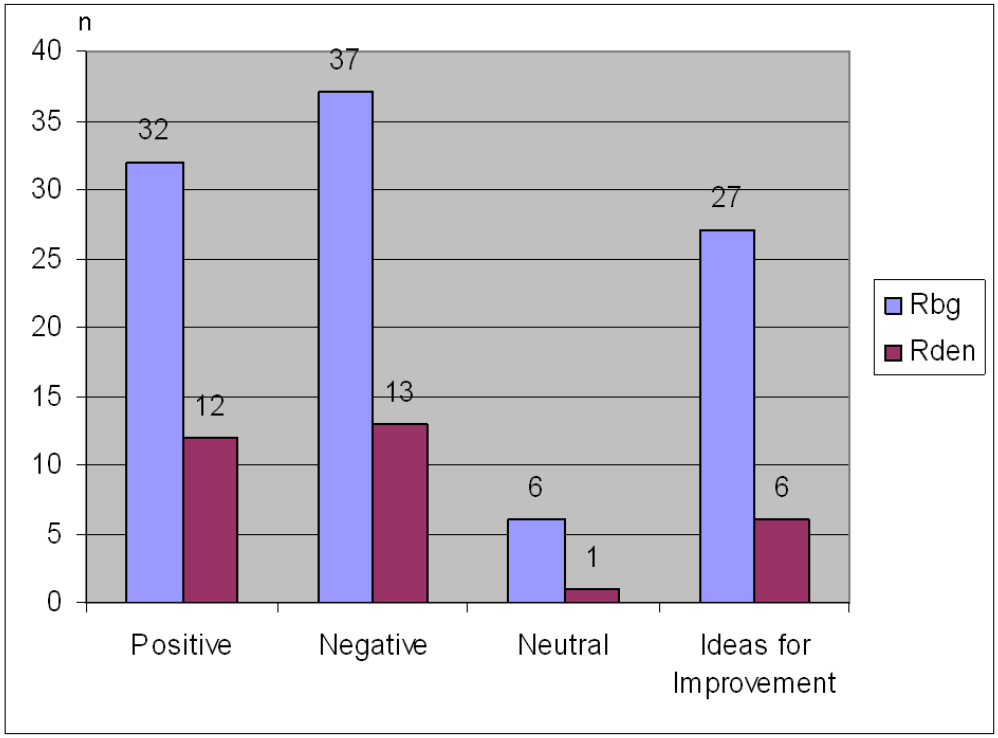


Figure 24. Distribution of the amount of given open-ended comments.

Both the positive and negative comments varied widely from small detailed responses to more general comments. The most typical positive comments dealt with issues such as the following:

- tutoring system
- good quality in teaching
- possibilities for flexible studying
- possibilities for distance learning
- the size of the university (“human size”)

The most typical negative comments were the following:

- working while studying hindered studies
- lectures too theoretical
- lack of motivation (either students’ or lecturers’ motivation)
- obligatory lectures or exercises
- overlapping lectures or exercises
- lack of resources in education / too many students participating in a course
- too many obligatory courses etc.

Ideas for improving education varied greatly. There were, for example, suggestions for more teamwork and seminars, seminars held in English, more practical exercises in courses, more connections with industrial companies and better pedagogical skills of lecturers, and even course-specific suggestions.

6.2.11 Generalisability of the Survey Results

The results presented in this study were collected from students who were given the possibility to decide whether or not they would give consent to allow access to their background data in databases held by TUT. Of the respondents, 29 % granted access and the remainder refused (the default answer was negative so that nobody would grant permission by mistake). In the previous results the specific respondent group (those who allowed the background usage: R_{bg}; those who denied it: R_{den}; and both groups together R_{all}) has been referred to each time the results have been presented.

To assess the generalisability of the findings, three characteristics of the respondents groups R_{bg} and R_{den} were studied in Chapter 6.1.2, *Differences between Respondents*. The results showed that the respondents in group R_{bg} had evaluated the difficulty of courses similarly to the respondents in the group R_{den}. However, in terms of the usefulness of courses respondents in the group R_{bg} had been somewhat more positively disposed than the respondents in group R_{den}. The mean weighted average of the respondent group R_{bg} was compared with the mean weighted average of all students graduating during the survey period and was found to be nearly the same. Thus is quite safe to assume that the respondents' opinions in group R_{bg} on the difficulty of the courses can be generalised to reflect the opinions of all respondents. In contrast, the opinions on the usefulness of the courses may seem somewhat more critical when all respondents' opinions are analysed. The average success level in their studies (measured by mean weighted course averages) of respondents in group R_{bg} can similarly be generalised to reflect the average overall success of all student graduating at the same time. Comparisons were not, however, made between the respondents' opinions on the usefulness and difficulty of the courses and all students' opinions of the same issues. This was because the main objective was to demonstrate the potential of the system used to connect background data to the original gathered responses and thus reveal the possibility to analyse what kind of students had responded to a particular question and how.

More important than the generalisability of any single research result presented in this study, is that this kind of survey, where students' responses are (with their permission) connected to

their individual background data, can be carried out in almost any university today. Thus it is possible for any institute to focus an investigation on those issues of particular interest to itself.

6.3 Limitations of the Results

There are several issues to be taken into account when the examples of the results of the present study are being considered. Among the most important of these are the objectivity of the study and the accuracy and representativeness of the results of the empirical part. In the following, both elements are discussed.

When estimating the objectivity of the study, the first question concerns the relationship between the researcher and the research object. Välimaa (2000, pp. 73–74) refers to the traditional scientific view, according to which it is not possible for researchers to be objective, because we cannot detach ourselves from our research objectives. With regard to evaluations, according to Välimaa (2000, pp. 73–74), a researcher cannot be objective if he/she is working at the same university as the objectives of the evaluation. Epistemologically, Guba & Lincoln (1989, p. 88) have described the essence of the constructivist paradigm in the following way: "[I]t is impossible to separate the inquirer from the inquired into. It is precisely their interaction that creates the data that will emerge from the inquiry." Thus, the dilemma can be seen in two ways. On the one hand in terms of Välimaa's remarks that the objectivity of those researchers who evaluate higher education is to be as objective as possible or, on the other hand, as the interaction between the researcher and research objectives as an essential element of the methodology.

This research has been conducted at the same university in which the researcher works. This may be construed as having limited the objectivity of the survey, even though the intention was to make it as objective as possible. However, the advantages of conducting the survey at the "home" university were that the university procedures and practices were already well known. This obviated the need to ask irrelevant questions from students and also made interpretation of the research results significantly easier when the background phenomena were already familiar. For example sometimes the differences in the survey results were due to the backgrounds of the students (high school vs. engineers) which may have been neglected were the environment not familiar to the researcher. It is still possible, or actually very likely, that the actual survey would have been designed differently if the researcher had not earlier conducted comparable paper-based surveys in the same university. It should also be noted that the actual survey was conducted only in Tampere University of Technology meaning that all respondents came from only one institution, which also may limit the applicability of the survey results to other academic environments. What is essential,

however, is that the actual survey was designed purposely to demonstrate the utility of an interlinked evaluation system. The gathered student responses, although valuable all as such, served mostly to demonstrate the usefulness of the connection between survey results and corresponding individual background information. Thus the fact that the researcher was familiar with the university was helpful when analysing the students' responses but this is not a major issue when considering the contribution of this study as a whole.

In designing a survey Dillman (2000, p. 11) has distinguished four sources of survey error and their consequences:

1. sampling error: The result of surveying only some, and not all, elements of the survey population,

2. coverage error: The result of not allowing all members of the survey population to have an equal or known nonzero chance of being sampled for participation of the survey,

3. measurement error: The result of poor question wording or questions being presented in such a way that inaccurate or non-interpretable answers are obtained and

4. nonresponse error: The result of people who respond to a survey being different from sampled individuals who did not respond, in a way relevant to the study.

The survey in this study focused on all students soon to graduate. All were all equally informed about the survey and thus given equal opportunity and encouragement to respond to the survey. Thus the first two errors are not likely to occur in this study. When analysing the responses there were some questions which had been understood in different ways by different students, such as the question about the amount of work experience. Classifying such answers slowed down interpreting the results to some extent, but it did not totally hinder understanding the responses, even though a few of them were ultimately rejected. In questions where a possibility for misunderstanding may exist, this has also been clearly presented when interpreting the results. In general the questions were well understood and thus could be analysed without ambiguity. In part, this may be due to the fact that the questions in the survey were intended to be similar to in many parts of the preceding paper-based survey which was in use at TUT during the years 2000 – 2001 and which had been considered useful. The fourth type of error, mentioned by Dillman (2000), was the nonresponse error, and studying it and its possible consequences form a part of this study.

It should be noted that the questions in the survey were not tested by pedagogical experts and so lack analysis of their psychometric properties as well as any extensive use in the field with subsequent reanalysis of their results as, for example Theall & Franklin (2000, p. 98) insist. Another issue which can be considered as a deficiency in the survey questionnaire is that the

level of difficulty and the workload were the only aspects to be rated for the courses the respondents had undertaken and the rating scale for both of them consisted of only three levels. The reason for designing the course questionnaire in this simple way that every respondent was expected to have executed an average of 50 – 60 courses by the time of responding to the survey. Generating more questions or more response alternatives would have made the survey too laborious and detailed for students to answer. In addition, the already mentioned fact that the emphasis in this study was on the connection between the survey and existing databases, not on the plain survey, also guided the whole study in a more comprehensive direction.

One of the most typically cited problems with online ratings concerns the response rates which, it is claimed, are often lower than with traditional paper-based evaluations. In this study the overall response rate was 43 %, however there were several responses where many sections of the survey pages were left blank. Of all respondents 29 % allowed the use of their background data. Statistical comparisons with the groups of respondents allowing the use of background information and respondents denying it showed that there were no significant differences in the respondents' opinions concerning the courses' difficulty. However, those respondents who allowed the use of their background data evaluated the courses' usefulness somewhat higher than those who denied it. The average of grade point averages of the respondents allowing the use of their background data was nearly the same as the average of grade point averages of all students graduating during survey execution time. Thus even though the response rate of this study was relatively low, it was possibile to analyse the background data of those respondents who allowed it and the information on how the respondents groups' evaluations differed from each other. It was therefore possible to analyse the results more thoroughly than could have been done using traditional evaluation methods (even with greater response rates) where the background information is not individually linked to the original survey results.

Another issue to be noted is that the results gained from the empirical survey concerning M.Sc. students opinions in this study cannot strictly be generalised to other environments. There already exist numerous research findings which have proved that students in different study programmes value different characteristics in education and also adopt different approaches to their own studies. For example Lizzio et al. (2002) found in a study of more than 2100 respondents from an Australian university that science students were the only group to show a positive association between their school achievement and deep approach to study at university. Commerce students, in contrast, were the only group to show a positive association between school achievement and surface processing (ibid.). As the authors note, the better the results of science and commerce students the more likely they were to report deep and surface approaches to study respectively. Humanities students, on the other hand showed no relationship between these factors.

In the literature it has also been recognised that if the respondents are registered in the same department where the course is offered, there is often a small but a significant "home ground" advantage (Husbands 1998). In the present study the respondents' (those allowing the access to their background data) study programmes and the departments provided the courses being evaluated were available, but not analysed. The main reason for this was that some of the students had changed their study programme after beginning their studies and interpreting the results would have thus been quite complicated.

When considering the actual responses collected from the students it should be also noted that the actual time the courses are assessed can affect the given estimations. Kuittinen (2004, p. 16) has argued that evaluations given any later than immediately after a course has been completed may be more useful in terms of the quality of the course and, especially, the usefulness of the course. However, it is also possible that the students would be more likely to remember their feelings about a course when they were participating in it than be able to objectively estimate the value of the course.

All in all, the words of Weiss (1996, p. 175) are worth keeping in mind: "Even with the best and most supportive data, models are never "proved". At best, they are not "disconfirmed" by the data. There may be alternative models that would provide equally plausible or better interpretations of the available facts. Scientific generalisations are built up by developing hypotheses and then submitting them to successive tests in an effort to disprove them or find the limits of their applicability."

7 CONCLUSIONS AND DISCUSSION

Evaluation is an investment in people and in progress.

– Guba & Lincoln 1989 –

Couper (2002) has stated that: "Each new technology enhances and extends the range of possibilities and opportunities for survey research, but also often introduces new challenges and issues for further research. Technology by itself is not inherently good or bad. It is how technology is harnessed in the service of human endeavour that determines its effect." This study has focused on the possibilities of online survey research in higher education provided by recent developments in technology. This field of study is especially important with the widespread use of student evaluations and the diverse ways of utilising the evaluation results: Student evaluations are used both in the development of a particular course and also for tenure decisions of a whole university faculty in addition to other purposes. The extensive use of evaluation results is also the reason for the active and often heated discussion of the validity and relevance of student evaluations.

7.1 Research Questions

The purpose of this study was to demonstrate the possibilities of utilising the connections between online ratings and currently existing databases in assessing the validity of student evaluations. This study has shown no new evidence of the existence of potential biases in student ratings. Instead it has demonstrated how easy it is to draw contradictory conclusions even from the same data. This is why the possibility to analyse student responses together with their individual background is much more informative than merely making statistical calculations of average background information.

The first research question of this study was presented in the following form: *"What kind of new information can be achieved by linking an online ratings system together with the university database and comparing individual students' responses to their individual background information, such as their actual course grades or the order in which they have proceeded in their studies?"*

This study has provided examples of issues which have been either actively researched but unsatisfactorily answered or which earlier were impossible to analyse at all. First, earlier analyses of the connection between students' survey responses and course grades have been based on students' own subjective expectations of the grades they were about to receive. This study showed that it is possible to analyse individual final course grades linked to the individual students' survey responses, even at the level of a single questionnaire item. Second, the relationship between the courses' workload, the courses' usefulness and the respondents' course grade has been one of the most widely researched topics in assessing the validity of student evaluations. This study showed that this connection can be analysed at an individual level, which had previously has been impossible. Third, the effect of the order in which individual students execute their courses connected to the same students' evaluation results and their individual final grades from the same courses was an example of an issue which, without this direct linkage, would have been impossible to investigate at all.

The second research question was the following: "*How can the linking of an online student evaluation to the other university's databases enhance the assessment of the validity of student evaluations, especially in search of potential biases such as the grading leniency?*

The results gained from an analysis of the results in the empirical part have demonstrated that an interlinked system can aid in assessing the validity of student evaluations in the following ways:

1) enabling the observations at individual respondent level instead of using averages and estimations, which in turn

2) reduce the need for complex statistical analyses which need expertise to be interpreted correctly.

The data received by the interlinked evaluation system demonstrated in this study can, naturally, also be analysed by means of high level statistical methods in order to yield more comprehensive results. In this study, however, the starting point for the research work was the finding that many lecturers question the validity of student evaluations arguing that student ratings are merely "popularity contests" (Felder 1995; Becker & Watts 1999; Emery et. al. 2003). With an interlinked evaluation system such as that demonstrated in the present study, any lecturer can easily view the survey results of his/her own course. At a glance of the figures where the responses and corresponding respondents' background data are presented together, he/she can see whether the results are valid and what limitations there may be.

7.2 Contribution of the Research

This study has demonstrated the potential to be gained when student evaluations are implemented online and connected [with respondents' permission] to other university data systems. The direct connection to the background data eliminates the need to insert into the questionnaire questions relating to items such as gender, major subject, and length of study, which can all be retrieved from the databases. The questionnaire can thus be shortened (increasing the students' willingness to respond to them) and response errors, either deliberate or accidental, are eliminated. As this study demonstrates, even more important is that by connecting the online survey results to other databases, each response together with the respondent's background information can be analysed anonymously at an individual level. This eliminates or at least reduces the need for using statistical algorithms based on class-averages or estimations which until now have been the most common ways to investigate the validity of the results of student evaluations.

Connecting online evaluation system to other data systems in the way demonstrated in this study will benefit both practitioners and researchers in the field of student evaluations in several ways:

- in everyday use it reduces the need for complex statistical analyses which, in turn, decreases the likelihood of human error caused by the use of statistical tools for gaining and interpreting the results.

- it enables analyses that have previously not been possible due to the lack of actual individual background information. In particular, the matrix presented in this study showing the students' opinions about the courses' usefulness and difficulty compared with the actual course grades they received, represents a new way to analyse a major topic in research into the validity of student evaluation. Prior research has been based on averages and estimations instead of actual individual data. In addition, the order and timing in which the respondents' have implemented their courses compared with their opinions on the same courses is another example of a useful product of the present study. Such issues can be analysed more thoroughly by the proposed construction to produce vital information for those personnel who are responsible for planning the curricula of degrees within universities.

- for scientific research, it opens up a range of new opportunities to achieve more valid and more reliable statistical analyses when the background data can be based on actual individual registered information instead of estimations and averages.

The last issue is of major importance, since as reported earlier (e.g., in Chapter 3.3.3) most of the scientific studies on the validity of student evaluations of teaching have been based on

factor analysis using class-averages as the unit of analysis. As a result, the individual variability occurring within each class has been ignored which has caused severe problems for single-level analyses.

7.3 Assessment of the Research

Assessment of research is an essential part of the research process. Criteria typically used for assessing the quality of research consist of four key elements: validity, reliability, generalisability and credibility (Gummesson 2000, p. 185). The following sections assess this thesis in terms of these four criteria.

7.3.1 Validity

As mentioned in the theoretical part of this study (see Chapter 3.3.1, *Validity*), the validity of an item can be assessed by examining, whether it measures what it purports to measure (Scriven 1991, 372). When assessing research work, this problem can also be formulated as a question: Does the evidence reflect the reality under examination? (Gummesson 2000, p. 185)

In this study, the question of validity has been the fundamental objective of the entire research project. To ensure that this study does not merely discuss validity as a topic, but also that validity is taken into the cognizance in the research process itself, the following actions were taken: The existing research materials concerning the validity of student evaluations were examined thoroughly to ascertain the deficiencies in current evaluation methods. In the development phase of the survey, the discovered deficiencies and the experiences of previous surveys were carefully considered when formulating the survey questions. To treat the researched topics as fully as possible both quantitative and qualitative questions were included. The online survey was designed together with TUT's system designer who implemented the survey and ensured that the linkages between the online survey and existing databases could be utilised to their fully potential. Analysis of the results included both the responses and results having the background connection from other databases and the responses and results without such a connection; also included were the percentages or numbers of respondents in both response groups. The received results have been compared with existing theoretical knowledge whenever possible.

To demonstrate the utility of the developed system several examples of survey results have been presented. As is well known, caution is necessary when generalising survey results based upon a general web sample that is intended to apply to the population at large (Schmidt

1997, pp. 277–279). Thus the presented results can not be generalised as such, but serve to demonstrate and concretize the vast potential of an interlinked evaluation system.

7.3.2 Reliability

The reliability of research work can be assessed by asking, whether other researchers would have obtained the same results had they carried out the same research (Gummesson 2000, p. 185). Another aspect of research reliability, especially when the research subject is student evaluations, concerns the consistency of the research results, i.e., whether an observation made at one point in time is likely to be similar to an observation made at another point in time (Braskamp & Ory 1994, p. 91).

In this study, as is the case with any research where the empirical data comes in the form of survey responses, it can be assumed that the responses of different surveys will not be identical. However, this study differs from traditional student evaluation studies in that the focus was on analysing the connection between students' responses and other background data and the ability to utilize such a connection rather than concentrating on the individual responses themselves. Thus, if this study were repeated by another researcher, it can be expected that whereas the single responses would be different, the analysis of the potential which the connection between the responses and other background data evokes would be quite similar. If the volume of survey data had been greater, it might also have been possible to analyse the statistical significance of individual survey response items. However, the data gathered was adequate for the original purpose of this study, which was to demonstrate examples of results that would have been unobtainable using earlier existing student evaluation methods.

7.3.3 Generalisability

Gummesson (2000, p. 185) assesses the generalisability of the research results by asking, what relevance the results have beyond the actual research. According to him (2000, p. 89) the traditional way of generalising from statistical samples is only one type of generalisation, and rarely applicable to a study where the empirical data comes from a single or only a few cases. Gummesson (2000, p. 90) emphasises that new data are never discomforting; instead, new data expand and improve existing theory.

The results of the interlinked online survey developed and analysed in this study have presented generic information about the students' perceptions both towards responding to an online survey and towards the whole curriculum they had experienced up to the time of

responding to the survey. The main general results of this survey showed that those students who consented to access of their background data when analysing their survey responses (29 % of all respondents) had considered the difficulty of the courses they had executed in the same way and the usefulness of the same courses slightly more positively than those students who had denied such access. Additionally, those respondents who had allowed access to their background data had completed the survey pages much more thoroughly than those respondents who had denied access to their background data.

However, more important than any of its particular survey findings, is that this study has demonstrated the usefulness of linking online survey results to other databases. By linking online survey results with the respondents' other background data any institute can implement evaluation systems that best support their needs. Moreover, each institute can concentrate on those validity (or other) issues which are of specific importance to them but have as yet remained unresolved. Thus the results of this study are not intended solely for researchers in the field, but also for everyday use within institutes of higher education.

7.3.4 Credibility

As a fourth element for assessing the quality of research, Gummesson (2000, p. 185) mentions credibility, which he formulates as questions: "Is there sufficient detail in the way the evidence was produced for the credibility of the research to be assessed? If we follow the researcher's journey – from questions to methods of data gathering, interpretation and answers – do we believe him or not?"

In the present study all the phases in the research have been documented as accurately as possible to make the entire research process visible. In addition, every attempt has been made to present the limitations of the gathered results as well as the limitations of the whole research.

To develop the new interlinked online survey presented in this study, a thorough examination was made of existing research findings. It was found that despite the fact that student evaluations are claimed to be one of the most widely researched topics in the field of higher education (Cashin 1995; McKeachie 1996), the potential of online surveys for assessing the validity of student evaluations has largely been overlooked. The research questions were formulated as a result of the insight that even though more universities have transferred their student evaluations from paper-based to web-based formats, the possibility to connect the evaluation systems to other information systems had not been recognised.

The survey executed in this study is not intended as an example of a pedagogically exhaustive questionnaire. However, the results gained from connecting the survey results to the same students' individual background information demonstrate that such access enables the evaluation results to be analysed more thoroughly.

7.4 Suggestions for Further Research

For future research, the constructed online evaluation method presented in this study opens up many new opportunities. For the ongoing research into the validity of student evaluation results and – more specifically – the potential existence of grading leniency, this proposed interlinked evaluation method developed here can bring new insights. The fact that all data can be accessed at an individual level eliminates the need for class-averages and students' estimations of their expected grades. Thus, any single course (given that there are enough students to make the analysis ethically sound) can be assessed at an individual level, anonymously to determine if biases exist, e.g., in the students' responses and the actual grades they awarded. Deeper statistical analyses (either using factor analyses or any other statistical methods) can also be expected to yield more accurate and valid results with this method.

Since practically all universities collect large amounts of information on their students in their databases, there already exists huge potential for the researchers on the field throughout the world. For the same reason, it is also possible for any single university to develop an interlinked evaluation system to gather and analyse student evaluations more thoroughly than has been previously possible.

In this study the causalities between background data and student responses were not analysed, only the possibilities of connecting background data to survey responses at the individual level. For future research, it would also be interesting to analyse further the student evaluations and individual background data to find why different students behave in different ways. However if such research is to be conducted, it is important that ethical aspects are also considered. The anonymity of students must not be compromised and the results must not be misused in any way which could jeopardise or hinder the students' study progress. Another interesting avenue for further research would be to conduct an interlinked follow-up survey of the respondents' opinions on the executed courses immediately on completion of the course and at later at the time of graduation. This would provide insight into whether the same respondents' evaluations of the same courses change as they gain more knowledge and experience.

When considering the significance of student evaluations – online or paper-based – it must be borne in mind that their utility depends upon how the evaluation results are being used. As Menges (2000) notes: no matter how valid and reliable the evaluation instrument is, consumers can and do misuse the results from it. And probably an even more common problem than a misuse of the evaluation results is that all too often the evaluation information remains unutilised. For evaluation to serve learning, both gathering and processing of information should be closely related to organisations' work processes (Seppänen-Järvelä 1999, p. 94).

There are already hundreds of studies attempting to resolve the debate over student ratings. This study is not to replace the existing research results but rather to show that by utilising modern technology and universities databases it is possible to create new, more informative means of gathering information. Couper (2002) has stated that "To the extent that we effectively exploit the advantages that technology may bring, we can potentially improve the quality of survey data, or at least reduce the costs or time involved in producing such data". This study has demonstrated that connecting online evaluation to already existing databases constitutes a solid basis for interpreting the evaluation results. With appropriate tools it is possible to attain more, yet asking less.

REFERENCES

Abrami, P.C. 1989. *How Should We Use Student Ratings to Evaluate Teaching?* Research in Higher Education, Vol. 30 (2), pp. 221–227.

Abrami, P.C. 2001. *Improving Judgments About Teaching Effectiveness Using Teacher Rating Forms.* New Directions for Institutional Research, Issue 109 (Spring 2001), pp. 59–87.

Adams, J.V. 1997. *Student Evaluations: The Ratings Game.* Inquiry, Vol. 1 (2), pp. 10–16.

Adams, M., Neumann, R. & Rytmeister, C. 1996. *Is it a Level Playing Field? Factors Which Influence Student Evaluation of Teaching.* Paper presented at the 1996 ERA/AARE Joint Conference, Singapore. [http://www.aare.edu.au/96pap/adamm96084.txt]. Visited 19.08.2005.

Afonso, N., Cardozo, L., Mascarenhas, O., Aranha, A. & Shah, C. 2005. *Are Anonymous Evaluations a Better Assessment of Faculty Teaching Performance? A Comparative Analysis of Open and Anonymous Evaluation Processes.* Family Medicine, Vol. 37 (19), pp. 43–47.

Ala-Mutka, K. 2005. *Automatic Assessment Tools in Learning and Teaching Programming*, Doctoral Thesis, Tampere, Tampere University of Technology, 102 [+67] p.

Aleamoni, L.M. 1999. *Student Rating Myths Versus Research Facts from 1924 to 1998.* Journal of Personnel Evaluation in Education, Vol. 13 (2), pp. 153–166.

Alexander, F.K. 2000. *The Changing Face of Accountability: Monitoring and Assessing Institutional Performance in Higher Education.* The Journal of Higher Education, Vol. 71 (4), pp. 411–431.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G. & Spooner, F. 2004. *Student Evaluation of College Teaching. A practice in search of principles.* College Teaching, Vol. 52 (4), pp. 134–141.

Andrews, S. & Feinberg, S. 1999. *Developing and Implementing Effective Web-based Surveys.* An article in the STC 1999 Proceedings. [http://www.stc.org/confproceed/1999/PDFs/046.PDF]. Visited 22.12.2004.

Ballantyne, C. 2000. *Why Survey Online? A Practical Look at Issues in the Use of the Internet for Surveys in Higher Education.* Paper presented at the annual conference of the American Evaluation Association, Honolulu, November 2000. [http://www.tlc.murdoch.edu.au/pubs/docs/aea-2000.html]. Visited 13.12.2004.

Ballantyne, C. 2003. *Online Evaluations of Teaching: An Examination of Current Practice and Considerations for the Future*. New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 103–112.

Ballantyne, C. 2004. *Online or on paper: An examination of the differences in response and respondents to a survey administered in two modes.* A paper presented at the Australasian Evaluation Society 2004 International Conference 13-15 October, Adelaide, Australia. 10 p.

Ballantyne, R., Borthwick, J. & Packer, J. 2000. *Beyond Student Evaluation of Teaching: identifying and addressing academic staff development needs.* Assessment & Evaluation in Higher Education, Vol. 25 (3), pp. 221–236.

Barrie, S. 2001. *Reflections on student evaluation of teaching: Alignment and congruence in a changing context.* In: Santhanam, E. (ed.). Student Feedback on Teaching: reflections and projections. Refereed Proceedings of Teaching Evaluation Forum, 28-28 August 2000, Perth, Australia, pp. 1–16.

Beaty, L. 2001. *Teaching Evaluation and Accreditation.* New Directions for Teaching and Learning, Issue 88, Winter 2001, pp. 75–85.

Becker, W.E. & Watts, M. 1999. *How Departments of Economics Evaluate Teaching.* American Economic Review, Vol. 89 (2), pp. 344–349.

Becker, W.E. 2000. *Teaching Economics in the 21st Century.* Journal of Economic Persepectives, Vol. 14 (1), pp. 109–119.

Benett, Y. 1993. *The Validity and Reliability of Assessments and Self-assessments of Work-based Learning.* Assessment & Evaluation in Higher Education, Vol. 18 (2), pp. 83–94.

Berk, R.A. 2005. *Survey of 12 Strategies to Measure Teaching Effectiveness*. International Journal of Teaching and Learning in Higher Education, Vol. 17 (1), pp. 48–62.

Betoret, F.D. & Tomás, A.D. 2003. *Evaluation of the University Teaching/Learning Processes for the Improvement of Quality in Higher Education.* Assessment & Evaluation in Higher Education, Vol. 28 (2), pp. 165–178.

Bhola, H.S. 2003. *Social and Cultural Contexts of Educational Evaluation: A Global Perspective.* In: Kellaghan, T. & Stufflebeam, D.L. (eds.). International Handbook of Educational Evaluation, Part One: Perspectives, Cornwall, Kluwer Academic Publishers, pp. 397–415.

Biggs, J. 1979. *Individual differences in study processes and the quality of learning outcomes.* Higher Education, Vol. 8 (4), pp. 381–394.

Boffo, S. & Moscati, R. 1998. *Evaluation in the Italian Higher Education System: many tribes, many territories … many godfathers.* European Journal of Education, Vol. 33 (3), pp. 349–360.

Borden, V.M.H. 2005. *Using alumni research to align program improvement with institutional accountability*. New Directions for Institutional Research, Issue 126, pp. 61–72.

Bosnjak, M. & Tuten, T.L. 2001. *Classifying Response Behaviors in Web-based Surveys.* Journal of Computer-Mediated Communication, Vol. 6 (3), [http://www.ascusc.org/jcmc/vol6/issue3/boznjak.html], Visited 24.01.2005.

Bothell, T.W. & Henderson, T. 2003. *Do Online Ratings of Instruction Make $ense?* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 69–79.

Bradley, N. 1999. *Sampling for Internet Surveys. An examination of respondent selection for Internet research.* Journal of the Market Research Society, Vol. 41 (4), pp. 387–395.

Brakke, D.F. & Brown, D.T. 2002. *Assessment to Improve Student Learning.* New Directions for Higher Education, Number 119, Fall 2002, pp. 119–122.

Braskamp, L.A. 2000. *Toward a More Holistic Approach to Assessing Faculty as Teachers.* New Directions for Teaching and Learning, Issue 83, Fall 2000, pp. 19–33.

Braskamp, L.A. & Ory, J.C. 1994. *Assessing Faculty work: Enhancing Individual and Institutional Performance*. Jossey-Bass Inc., Publishers, San Francisco, 333 p.

Brennan, J., Brighton, R., Moon, N., Richardson, J., Rindl, J. & Williams, R. 2003. *Collecting and using student feedback on quality and standards of learning and teaching in higher education.* The Open University Centre for Higher Education Research and Information. 131 p. [http://www.hefce.ac.uk/Pubs/rdreports/2003/rd08_03/]. Visited 03.11.2005.

Brennan, J. & Williams, R. 2004. *Collecting and using student feedback. A guide to good practice*. Learning and Teaching Support Network, York, 58 p.

Broder, J.M. & Dorfman, J. H. 1994. *Determinants of Teaching Quality: What's Important to Students?* Research in Higher Education, Vol. 35 (2), pp. 235–249.

Brown, G., Bull, J. & Pendlebury, M. 1997. *Assessing Student Learning in Higher Education.* Routledge. London. 317 p.

Bull, J. 1999. *Computer-Assisted Assessment: Impact on Higher Education Institutions.* Educational Technology & Society. Vol 2 (3), pp. 123–126.

Bull. J. & Danson, M. 2004. *Computer Assisted Assessment (CAA).* LTSN Generic Centre, Assessment Series No 14, 24 p. Available also in Internet: [http://www.heacademy.ac.uk/embedded_object.asp?id=20388&file], Visited 22.04.2005.

Bullock, C.D. 2003. *Online Collection of Midterm Student Feedback.* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 95–102.

Candoli, C. & Stufflebeam, D.L. 2003. *The Context of Educational Program Evaluation in the United States*. In: Kellaghan, T. & Stufflebeam, D.L. (eds.). International Handbook of Educational Evaluation, Part One: Perspectives, Cornwall, Kluwer Academic Publishers, pp. 417–428.

Cannon, R. 2001a. *Broadening the Context for Teaching Evaluation.* New Directions for Teaching and Learning, Issue 88, Winter 2001, pp. 87–97.

Cannon, R. 2001b. *Evaluating learning or evaluating teaching: Is there a difference and does it matter?* Refereed Proceedings of Teaching Evaluation Forum, Perth, Australia, 28th – 29th August 2000, The University of Western Australia, pp. 81- 92.

Carey, G.W. 1993. *Thoughts on the Lesser Evil: student evaluations.* Perspectives on Political Science, Vol. 22 (1), pp. 17–20.

Carini, R.M., Hayek, J.C., Kuh, G.D., Kennedy, J.M. & Ouimet, J.A. 2003. *College Student Responses to Web and Paper Surveys: Does Mode Matter?* Research in Higher Education, Vol. 44 (2), pp. 1–19.

Cashin, W.E. 1995. *Student Ratings of Teaching: The Research Revisited*. Idea Paper no. 32, September 1995. [http://www.idea.ksu.edu/papers/Idea_Paper_32.pdf]. Visited 28.12.2004.

Centra, J.A. 2000. *Evaluating the Teaching Portfolio: A Role for Colleagues.* New Directions for Teaching and Learning, Issue 83, Fall 2000, pp. 87–93.

Centra, J.A. 2003. *Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work?* Research in Higher Education, Vol. 44 (5), pp. 495–518.

CETL. 2004. *Online Student Evaluation of Instruction.* 2 p. [http://www.csufresno.edu/cetl/FacInfo/Technlogy/OnlineEval.pdf]. Visited 10.12.2004.

Chambers, B.A. & Schmitt, N. 2002. *Inequity in the Performance Evaluation Process: How You Rate Me Affects How I Rate You*. Journal of Personnel Evaluation in Education. Vol. 16 (2), pp. 103–112.

Chelimsky, E. 1997. *The Coming Transformations in Evaluation.* In Chelimsky, E. & Shadish, W.R. (eds.). Evaluation for the 21st Century. Thousand Oaks, SAGE Publications, pp. 1–26.

Chen, Y. & Hoshower, L.B. 2003. *Student Evaluation of Teaching Effectiveness: an assessment of student perception and motivation.* Assessment & Evaluation in Higher Education, Vol. 28 (1), pp. 71–88.

Cheng, D.X. 2001. *Assessing Student Collegiate Experience: where do we begin?* Assessment & Education in Higher Education. Vol 26 (6), pp. 525–538.

Cho, H. & LaRose, R. 1999. *Privacy Issues in Internet Surveys*. Social Science Computer Review, Vol. 17 (4), pp. 421–434.

Cody, A. 1999. *Evaluation via the web.* Teaching and Education News, Vol. 9 (4). [http://www.tedi.uq.edu.au/TEN/TEN_previous/TEN4_99/index.html]. Visited 16.12.2004.

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. Vol. 26 (1), pp. 89–93.

Conn, C. 2003. *Using the Internet for Surveying: Techniques for Designing , Developing & Delivering*. Office of Academic Assessment, Northern Arizona University. Internet Article. [http://www4.nau.edu/assessment/main/research/responserates.htm]. Visited 05.11.2004.

Conn, C. & Norris, J. 2004. *Investigating Strategies for Increasing Student Response Rates to Online-Delivered Course Evaluations.* Poster session presented at the 2nd Annual NAU

Assessment Fair, Flagstaff, AZ. [www4.nau.edu/assessment/oaainfo/presentations/ ResponseRates/ConnNorrisproceedingspaper%5B1%5D.pdf]. Visited 13.12.2004.

Couper, M.P. 2000. *Web surveys: a review of issues and approaches.* Public Opinion Quarterly, Vol. 64 (4), pp. 464–494.

Couper, M. 2002. *New Technologies and Survey Data Collection: Challenges and Opportunities.* Keynote speech at the International Conference on Improving Surveys, ICIS, Copenhagen, August 25-28 2002, [http://www.icis.dk/ICIS_papers/Keynote1_0_3.pdf]. Visited 06.04.2005.

Couper, M.P., Traugott, M.W. & Lamias, M.J. 2001. *Web survey design and administration.* Public Opinion Quarterly, Vol. 65 (2), pp. 230–253.

Cranton, P. 2001. *Interpretive and Critical Evaluation.* New Directions for Teaching and Learning, Issue 88, Winter 2001, pp. 11–18.

Crawford, S., McCabe, S., Couper, M. & Boyd, C. 2002. *From Mail to Web: Improving Response Rates and Data Collection Efficiency*. A paper presented at the International Conference on Improving Surveys, ICIS, Copenhagen, August 25–28 2002, [http://www.icis.dk/ICIS_papers/B_2_2.pdf]. Visited 06.04.2005.

Creswell, J. 2003. *Research Design. Qualitative, Quantitative, and Mixed Methods Approaches.* Thousand Oaks USA, Sage Publications Inc., 246 p.

Cummings, R. & Ballantyne, C. 1999. *Student Feedback on Teaching: Online! On Target*? Paper presented at the Australasian Society Annual Conference, October 1999. [http://www.tlc.murdoch.edu.au/pubs/docs/AES_1999_Conference_Final.rtf]. Visited 13.12.2004.

Cummings, R., Ballantyne, C. & Fowler, L. 2001. *Online student feedback surveys: Encouraging staff and student use.* Refereed Proceedings of Teaching Evaluation Forum, Perth, Australia, 28th – 29th August 2000, The University of Western Australia, pp. 29–37.

d'Apollonia, S. & Abrami, P.C. 1997. *Navigating Student Ratings of Instruction*, American Psychologist, Vol. 52 (11), pp. 1198–1208.

Deasy, J. 2004. *Course and Instructor Evaluations: Why, When and How.* A Presentation at 2004 APAP Education Forum Syllabus, November 3–7 2004, Nashville, USA. An Internet Document [http://www.apap.org/2004syllabus/pdf/1.pdf]. Visited 17.11.2004.

Dieks, D. 1992. *Doomsday - - Or: The Dangers of Statistics.* The Philosophical Quarterly, Vol. 42, No. 166, pp. 78–84.

Dietel, R.J. , Herman, J.L. & Knuth, R.A. 1991. *What Does Research Say About Assessment?* North Central Regional Educational Laboratory, U.S. Internet Article. [http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm]. Visited 03.01.2006.

Dill, D.D. 1998. *Evaluating the 'Evaluative State': implications for research in higher education.* European Journal of Education, Vol. 33 (3), pp. 361–377.

Dillman, D.A. 2000. *Mail and Internet Surveys: The Tailored Design Method.* New York, USA, John Wiley & Sons, Inc., 464 p.

Dillman, D.A. & Bowker, D.K. 2001. *The Web Questionnaire Challenge to Survey Methodologists.* In Reips, U. & Bosnjak. M. (eds.) Dimensions of Internet Science, Pabst Science Publishers, Lengerich, Germany, pp. 159–178.

Dillman, D., Tortora, R. Conradt, J. & Bowker, D. 1998. *Influence of plain vs. fancy design on response rates for web surveys.* A paper presented at Joint Statistical Meetings, Dallas, Texas, August 1998. [http://survey.sesrc.wsu.edu/dillman/papers/asa98ppr.pdf]. Visited 13.12.2004.

Doerfel, M.L. & Ruben, B.D. 2002. *Developing More Adaptive, Innovative, and Interactive Organizations.* New Directions for Higher Education, No. 118, Summer 2002, pp. 5–27.

Dommeyer, C.J., Baum, P., Chapman, K.S. & Hanna, R.W. 2002a. *Attitudes of Business Faculty Towards Two Methods of Collecting Teaching Evaluations: Paper vs. Online.* Assessment & Evaluation in Higher Education, Vol. 27 (5), pp. 455–462.

Dommeyer, C.J., Baum, P. & Hanna, R.W. 2002b. *College Students' Attitudes Toward Methods of Collecting Teaching Evaluations: In-Class Versus On-Line.* Journal of Education for Business, Vol. 78(1), pp. 11–15.

Dommeyer, C.J., Baum, P., Chapman, K.S. & Hanna, R.W. 2003. *An Experimental Investigation of Student Response Rates to Faculty Evaluations: the effect of the online method and online treatment*. Proceedings of the 34th Annual Meeting of the Decision Sciences Institute, November 22–25 2003, Washington D.C. [http://www.sbaer.uca.edu/research/dsi/2003/procs/451-7916.pdf]. Visited 04.03.2005.

Dommeyer, C.J., Baum, P., Hanna, R.W. & Chapman, K.S. 2004. *Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations.* Assessment & Evaluation in higher Education, Vol. 29 (5), pp. 611–623.

Dressel, P.D. 1976. *Handbook of Academic Evaluation.* San Francisco, Jossey-Bass Publishers, 518 p.

Driscoll, M.P. 2001. *Computers for What? Examining the Roles of Technology in Teaching and Learning.* Educational Research and Evaluation. Vol. 7 (2–3), pp. 335–349.

Dunegan, K.J. & Hrivnak, M.W. 2003. *Characteristics of Mindless Teaching Evaluations and the Moderating Effects of Image Compatibility*, Journal of Management Education, Vol. 27 (3), pp. 280–303.

Emery, C. R., Kramer T. R. & Tian, R. G. 2003. *Return to academic standards: a critique of student evaluations of teaching effectiveness.* Quality Assurance in Education. Vol 11 (1), pp. 37–46.

Erwin, T.D. 1991. *Assessing Student Learning and Development.* A Guide to the Principles, Goals, and Methods of Determining College Outcomes. San Francisco, Jossey-Bass Publishers, 208 p.

Eskola, J. & Suoranta, J. 2001. *Johdatus laadulliseen tutkimukseen*. Tampere, Osuuskunta Vastapaino. 266 p.

Felder, R.M. 1995. *What Do They Know Anyway?* TRC Newsletter 6.2 (Spring 1995). [http://www.indiana.edu/~teaching/felder.html]. Visited 20.12.2004.

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, Vol. 9 (3), pp. 199–242.

Felton, J., Mitchell, J. & Stinson, M. 2004. *Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness.* Assessment & Evaluation in Higher Education, Vol. 29 (1), pp. 91–108.

Fenwick, T.J. 2001. *Using Student Outcomes to Evaluate Teaching: A Cautious Exploration.* New Directions for Teaching and Learning, Issue 88, Winter 2001, pp. 63–74.

Fetterman, D.M. 1997. *Empowerment Evaluation and Accreditation in Higher Education.* In Chelimsky, E. & Shadish, W.R. (eds.). Evaluation for the 21st Century. Thousand Oaks, SAGE Publications, pp. 381–395.

Feynman, R.P. 1989. *"Surely You're Joking, Mr. Feynman!" Adventures of a Curious Character.* New York, Bantam Books, 322 p.

Forsman, G. & Varedian, M. 2002. *A Cost and Response Rate Comparison in a Study of Students Housing Conditions*. A paper presented at the International Conference on Improving Surveys, ICIS, Copenhagen, August 25–28 2002, [http://www.icis.dk/ICIS_papers/C2_2_3.pdf], Visited 06.04.2005.

Fowler, F.J. Jr. 2002. *Survey Research Methods*. Applied Social Research Methods Series, Volume 1. Thousand Oaks, SAGE Publications Inc. 178 p.

Franklin, J. 2001. *Interpreting the Numbers: Using a Narrative to Help Others Read Student Evaluations of Your Teaching Accurately.* New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 85–100.

Freedman, D., Pisani, R. & Purves. R. 1998. *Statistics.* New York, W.W.Norton & Company, 578 p.

Gamliel, E. & Davidovitz, L. 2005. *Online versus traditional teaching evaluation: mode can matter*, Assessment & Evaluation in Higher Education, Vol. 30 (6), pp. 581–592.

Gardner, L. , Sheridan, D. & White, D. 2002. *A web-based learning and assessment system to support flexible education.* Journal of Computer Assisted Learning, Vol. 18 (2), pp. 125–136.

Geis, G.L. 1991. *The Moment of Truth: Feeding Back Information About Teaching.* New Directions for Teaching and Learning, No. 48, Winter 1991, pp. 7–20.

Goldstein, R.J. 1993. *Some Thoughts about Standardized Teaching Evaluations.* Perspectives on Political Science, Vol. 22 (1), pp.8–10.

Goodman, A. & Campbell, M. 1999. *Developing Appropriate Administrative Support For Online Teaching With An Online Unit Evaluation System.* Paper presented at ISIMADE'99, International Symposium on Intelligent Multimedia and Distance Education, August 1999, Germany.

Gramlich, E.M. & Greenlee, G.A. 1993. *Measuring Teaching Performance.* Journal of Economic Education, Winter 1993, pp. 3–13.

Gray, P.J. 1991. *Using Assessment Data to Improve Teaching.* New Directions for Teaching and Learning, No.48, Winter 1991, pp. 53–64.

Greenwald, A.G. 1997. *Validity Concerns and Usefulness of Student Ratings of Instruction.* American Psychologist, Vol. 52 (11), pp. 1182–1186.

Greenwald, A.G. & Gillmore, G.M. 1997a. *Grading Leniency Is a Removable Contaminant of Student Ratings.* American Psychologist, Vol. 52 (11), pp. 1209–1217.

Greenwald, A.G. & Gillmore, G.M. 1997b. *No Pain, No Gain? The Importance of Measuring Course Workload in Student Ratings of Instruction.* Journal of Educational Psychology, Vol. 89 (4), pp. 743–751.

Greimel-Fuhrmann, B. & Geyer, A. 2003. *Students' Evaluation of Teachers and Instructional Quality – Analysis of Relevant Factors Based on Empirical Evaluation Research.* Assessment & Evaluation in Higher Education, Vol. 28 (3), pp. 229–239.

Griffin, B.W. 2001. *Instructor Reputation and Student Ratings of Instruction.* Contemporary Educational Psychology, Vol. 26 (4), pp. 534–552.

Griffin, B.W. 2004. *Grading leniency, grade discrepancy, and student ratings of instructionn.* Contemporary Educational Psychology, Vol. 29(4), pp. 410–425.

Guba, E.G. & Lincoln, Y.S. 1989. *Fourth Generation Evaluation.* Newbury Park, SAGE Publications Inc. 294 p.

Gummesson, E. 1993. *Case Study Research in Management.* Methods for Generating Qualitative Data. Stockholm University, Department of Business Administration, 63 p.

Gummesson, E. 2000. *Qualitative Methods in Management Research.* Second Edition. Thousand Oaks, Sage Publications Inc. 250 p.

Gunn, H. 2002. *Web-based Surveys: Changing the Survey Process.* First Monday, Vol. 7(12). [http://www.firstmonday.dk/issues/issue7_12/gunn/]. Visited 13.12.2004.

Ha, T.S. & Marsh, J. 1998. *Using the Web for Student Evaluation of Teaching. (COSSET & OSTEI).* In: James, J. (ed.). Quality in Teaching and Learning in Higher Education: A collection of refereed papers from the first conference on Quality in Teaching and Learning in Higher Education. The Hong Kong Polytechnic University, Educational Development Centre, pp. 237–244.

Ha, T.S., Marsh, J. & Jones, J. 1998. *A Web-based System for Teaching Evaluation.* Paper presented at the Lingnam College 30th Anniversary Conference on "New Challenges and Innovations in Teaching and Training into the 21st Century, May 1998, Hong Kong.

Ha, T. S., Marsh, J. & Jones, J. 2000. C*entralized On-line System for Student Evaluation of Teaching.* (Online document). [http://home.ust.hk/~eteval/cosset/]. Visited 10.1.2005.

Hanson, J., Millington, C. & Freewood, M. 2001. *Developing a Methodology for Online Feedback and Assessment.* Proceedings of the Fifth International Computer Assisted Assessment Conference. Loughborough University, United Kingdom. July 2nd – 3rd. Online proceedings. [http://s-d.lboro.ac.uk/caanew/pastConferences/2001/proceedings/n2.pdf]. Visited 22.11.2005.

Hardy, N. 2003. *Online Ratings: Fact and Fiction.* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 31–38.

Harman, G. 1998. *Quality Assurance Mechanisms and Their Use as Policy Instruments: major international approaches and the Australian experience since 1993*. European Journal of Education, Vol. 33 (3), pp. 331–348.

Harvey, L. 2001. *Student Feedback. A report to the Higher Education Funding Council for England. Centre for research into Quality.*
[http://uce.ac.uk/crq/publications/studentfeedback.pdf]. Visited 01.12.2004.

Harvey, L. 2004. *Analytic Quality Glossary*. Quality Research International.
[http://www.qualityresearchinternational.com/glossary/]. Visited 19.09.2005.

Harvey, L. & Askling, B. 2003. *Quality in Higher Education*. In: Begg, R. (ed.). 2003. The dialogue between higher education research and practice. Kluwer Academic Publishers, Dordrecht, pp. 69–83.

Harvey, L. & Green, D. 1993. *Defining Quality.* Assessment & Evaluation in Higher Education, Vol. 18 (1), pp. 9–34.

Heerwegh, D. & Loosveldt, G. 2002. *An evaluation of the effect of response formats on data quality in Web surveys*. A paper presented at the International Conference on Improving Surveys, ICIS, Copenhagen, August 25–28 2002, [http://www.icis.dk/ICIS_papers/A2_3_2.pdf], visited 06.04.2005.

Heikkilä, T. 2001. *Tilastollinen tutkimus.* Helsinki, Edita. 328 p.

Hendry, G.D., Cumming, R.G., Lyon, P.M. & Gordon, J. 2001. *Student-centred Course Evaluation in a Four-year, Problem Based Medical Programme: issues in collection and management of feedback.* Assessment & Evaluation in Higher Education. Vol. 26 (4), pp. 327–339.

Hendry, G.D. & Dean, S.J. 2002. *Accountability, evaluation of teaching and expertise in higher education.* The International Journal for Academic Development. Vol. 7 (1), pp. 75–82.

Henkel, M. 1998. *Evaluation in Higher Education: conceptual and epistemological foundations*, European Journal of Education, Vol. 33 (3), pp. 285–297.

Heywood, J. 1988. *Assessment in Higher Education*. Chichester, Wiley. 448 p.

Hiironniemi, O. & Tuunainen, L. 1995. *Välttämätön paha vai nuorempi kollega?* Opetuksen arviointiprojekti, Tampereen yliopisto.
[http://www.uta.fi/opiskelu/opetuksen_tuki/arviointi/opiskys.htm] Visited 03.11.2004.

Hmieleski, K. & Champagne, M.V. 2000. *Plugging in to Course Evaluation*. Assessment, September/October 2000.
[http://distance.wsu.edu/facultyresources/savedfromweb/pluggingin.htm]. Visited 3.10.2005.

Hoey, J.J. & Gardner, D.C. 1999. *Using Surveys of Alumni and Their Employers to Improve an Institution*. New Directions for Institutional Research, No. 101, Spring 1999, pp. 43–58.

Hoffman, K.M. 2003. *Online Course Evaluation and Reporting in Higher Education.* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 25–29.

Hofstetter, C.H. & Alkin, M.C. 2003. *Evaluation Use Revisited.* In: Kellaghan, T. & Stufflebeam, D.L. (eds.). International Handbook of Educational Evaluation, Part One: Perspectives, Cornwall, Kluwer Academic Publishers, pp. 197–222.

Hubball, H., Clarke, A. & Beach, A.L. 2004. *Assessing Faculty Learning Communities.* New Directions for Teachnig and Learning, Number 97, Spring 2004, pp. 87–100.

Huemer, M. 2007. *Student Evaluations: A Critical review.*
http://home.sprynet.com/~owl1/sef.htm. Visited 04.01.2008.

Husbands, C.T. 1998. *Implications for the Assessment of the Teaching Competence of Staff in Higher Education of Some Correlates of Students' Evaluations of Different Teaching Styles.* Assessment & Evaluation in Higher Education, Vol. 23 (2), pp. 117–136.

Ince, M. 2006. *Oxbridge closes on Harvard in rankings. (cover story)* Times Higher Education Supplement Issue 1763, pp. 1–4.

Ingham, J. 2000. *Data Warehousing: A Tool for the Outcomes Assessment Process.* IEEE Transactions on Education, Vol. 43 (2), pp. 131– 136.

Jackson, D.L., Teal, C.R., Raines, S.J., Nansel, T.J., Force, R.C. & Burdsal, C.A. 1999. *The Dimensions of Students' Perceptions of Teaching Effectiveness.* Educational and Psychological Measurement, Vol. 59 (4), pp. 580–596.

Job, P.A. 2004. *The Relationship between Personality, Occupation and Student Evaluations of Teaching Effectiveness of Adjuct Faculty.* Dissertation Thesis. Portland State University, 222 p.

Johnson, T.D. 2003. *Online Student Ratings: Will Students Respond*? New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 49–68.

Johnson, T.D. & Ryan, K.E. 2000. *A Comprehensive Approach to the Evaluation of College Teaching*. New Directions for Teaching and Learning, Issue 83, Fall 2000, pp. 109–123.

Johnston, B. 2004. *Summative assessment of portfolios: an examination of different approaches to agreement over outcomes.* Studies in Higher Education, Vol. 29 (3), pp. 395–412.

Jokinen, T., Malinen, H., Mäki, M., Nokela, J., Pakkanen, P. & Kekäläinen H. 2007. *Tampereen teknillisen yliopiston laadunvarmistusjärjestelmän auditointi*. Korkeakoulun arviointineuvoston julkaisuja 4:2007. Tampere, Tammer-Paino Oy, 58 p.

Kaikkonen, V. 1996. *Johdatus yrityksen taloustieteelliseen ajatteluun ja tutkimukseen.* Tampereen yliopisto, Tampere, Jäljennepalvelu, 112 p.

Kallio, T. J. 2006. *Teoreettinen tutkimus ja liiketaloustieteet*. Liiketaloudellinen aikakauskirja, 4/2006, pp. 510–538.

Kasanen, E., Lukka, K. & Siitonen, A. 1993. *The Constructive Approach in Management Accounting Research*, Journal of Management Accounting Research, Vol. 5, pp. 243–264.

Kauranen, I., Aaltonen, M., Naumanen, M. & Kaila, M.M. 1992. *A Guidebook for Writers of Research Papers in Industrial Management.* Helsinki University of Technology, Institute of Industrial Management, Otaniemi, 70 p.

Kehoe, C.M. & Pitkow, J.E. 1996. *Surveying the Territory: GVU's Five WWW User Surveys.* The World Wide Web Journal, Vol. 1 (3), pp. 77–84.

Kellaghan, T., Stufflebeam, D.L. & Wingate L. A. 2003. *International Handbook of Educational Evaluation,* Part One: Perspectives, Cornwall, Kluwer Academic Publishers, 481 p.

Kelly, M. & Marsh, J. 1999. *Going on-line with student evaluation of teaching.* Evaluation of the student experience project, Vol. 6. City University of Hong Kong, Centre for the Enhancement of Learning and Teaching, 70 p. [http://teaching.polyu.edu.hk/datafiles/R54.pdf]. Visited 08.03.2005.

Kember, D., Leung, D.Y.P. & Kwan, K.P. 2002. *Does the Use of Student Feedback Questionnaires Improve the Overall Quality of Teaching?* Assessment & Evaluation in Higher Education. Vol. 27 (5), pp. 411–425.

Kerka, S. & Wonacott, M.E. 2000. *Assessing Learners Online*. Practitioner File, pp. 1–12. [http://www.cete.org/acve/docs/pfile03.pdf]. Visited 17.1.2005.

Kerridge, J.R. & Mathews, B.P. 1998. *Student Ratings of Courses in HE: further challenges and opportunities.* Assessment & Evaluation in Higher Education. Vol. 23 (1), pp. 71–83.

Kiesler, S. & Sproull, L.S. 1986. *Response Effects in the Electronic Survey*. Public Opinion Quarterly, Vol. 50 (3), pp. 402– 413.

Knapper, C. 2001. *Broadening Our Approach to Teaching Evaluation.* New Directions for Teaching and Learning, Issue 88, Winter 2001, pp. 3–9.

Kolari, S. 2003. *An Active Role for Students in the Learning Process in Engineering Education: Interactive Teaching Methods in Promoting Understanding,* Doctoral Thesis, Tampere, Tampere University of Technology, TTY-Paino, 55 [+75] p.

Konttinen, R. 1996. *Arvostelusta näyttöön – koulutuksen arvioinnin kehityspiirteitä Suomessa*. In: Takala, S. (ed.) Arviointi ja koulutuksen laadun kehittäminen. Kasvatustieteiden tutkimuslaitos, Jyväskylän yliopisto, Jyväskylä, Jyväskylän yliopistopaino, 257 p.

Kreber, C. & Brook, P. 2001. *Impact evaluation of educational development programmes.* The International Journal for Academic Development. Vol. 6 (2), pp. 96–108.

Kuittinen. M. 2004. *Opetuksessa ei voi kaikkia miellyttää.* Peda-forum. 1/2004, pp.16–18.

Kulik, J.A. 2001. *Student Ratings: Validity, Utility, and Controversy.* New Directions for Institutional Research, Issue 109 (Spring 2001), pp. 9–25.

Kurz, L. & Banta, T.W. 2004. *Decoding the Assessment of Student Learning.* New Directions for Teaching and Learning, Number 98, Summer 2004, pp. 85–94.

Kwan, K. 1999. *How Fair are Student Ratings in Assessing the Teaching Performance of University Teachers?* Assessment & Evaluation in Higher Education, Vol. 24 (2), pp. 181–195.

Kärkkäinen, J. 2005. *Ohjaus opintojen edistäjänä? Tampereen teknillisen yliopiston opiskelijoiden kokemuksia opintojen etenemisen esteistä, opintojen ohjauksesta ja yliopistoon integroitumisesta*. Tampereen teknillinen yliopisto, Yliopistopalvelut, Raportti 1. 116 p.

Lappalainen, M. 1997. *Arvioinnin merkitys yliopiston opetuksen ja oppimisen osina.* In: Lappalainen, M. (ed.). Opetus, oppiminen ja arviointi. Turun yliopiston arviointijärjestelmän rakentaminen. Turun yliopisto, Hallintoviraston julkaisusarja 4/97. Turku, Unipaps, pp. 7–37.

Layne, B.H., DeCristoforo, J.R. & McGinty, D. 1999. *Electronic versus traditional student ratings of instruction.* Research in Higher Education, Vol. 40 (2), pp. 221–232.

Leathwood, C. & Phillips. D. 2000. *Developing curriculum evaluation research in highereducation: Process, politics and practicalities.* Higher Education, Issue 40, pp. 313–330.

Lehtimäki, S. 2006. *Valmistumisen vauhdittaminen –projekti. Palautekyselyn tulokset*. Oulun yliopisto, Sähkö- ja tietotekniikan osasto. 24 p. [http://www.ee.oulu.fi/opiskelu/materiaali/PalautekyselyR_tiivistelma.pdf]. Visited 02.05.2007.

Lewis, K.G. 2001a. *Making Sense of Student Written Comments.* New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 25–32.

Lewis, K.G. 2001b. *Using Midsemester Student Feedback and Responding to it.* New Directions for Teaching and Learning, No. 87, Fall 2001, pp. 33–44.

Lieberman, D., Bowers, N. & Moore, D.R. 2001. *Use of Electronic Tools to Enhance Student Evaluation Feedback.* New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 45–54.

Liegle, J. & McDonald, D.S. 2004. *Lessons Learned From Online vs. Paper-based Computer Information Students' Evaluation System.* Proceedings of ISECON, Newport, November 4-7, pp. 1–12.

Linnakylä, P. 2002. *Kansainvälisten ja kansallisten oppimistulosten arviointien välisestä suhteesta*. In: Olkinuora, E., Jakku-Sihvonen, R. & Mattila, E. (eds.). Koulutuksen arviointi. Lähtökohtia, malleja ja tilannekatsauksia. Turun yliopiston kasvatustieteiden tiedekunnan julkaisuja B:70. Turku, Painosalama Oy, pp. 31–51.

Lizzio, A., Wilson, K. & Simons, R. 2002. *University Students' Perceptions of the Learning Environment and Academic Outcomes: implications for theory and practice.* Studies in Higher Education. Vol. 27 (1), pp. 27–52.

Llewellyn, D.C. 2003. *Online Reporting of Results for Online Student Ratings.* New Directions for Teaching and Learning, No. 96, Winter 2003, pp. 61–68.

Lubinescu, E.S., Ratcliff, J.L. & Gaffney, M.A. 2001. *Two Continuums Collide: Accreditation and Assessment.* New Directions for Higher Education, No. 113, Spring 2001, pp. 5–21.

Lönnqvist, A. 2004. *Measurement of Intangible Success Factors: Case Studies on the Design, Implementation and Use of Measures*. Doctoral Thesis. Tampere, Tampere University of Technology. Publication 475. TTY-Paino. 255 p.

MacElroy, B. 1999. *Comparing seven forms of on-line surveying*. Quirk's Marketing Research Review. [http://www.quirks.com/articles/article_print.asp?arg_articleid=510]. Visited 15.12.2004.

MacLeod, C. 2001. *Student Feedback on Teaching: An Institutional Perspective.* In: Santhanam, E. (ed.). Student Feedback on Teaching: reflections and projections. Refereed Proceedings of Teaching Evaluation Forum, 28–28 August 2000, Perth, Australia, pp. iv–xii.

Manfreda, K.L., Batagelj, Z. & Vehovar, V. 2002. D*esign of Web Survey Questionnaires: Three Basic Experiments.* Journal of Computer-Mediated Communication, Vol. 7 (3). [http://www.ascusc.org/jcmc/vol7/issue3/vehovar.html]. visited 22.12.2004.

Manwaring, G. & Calverley, C. 1998. *Directing your evaluation.* In Harvey, J. (ed.). Evaluation Cookbook. Edinburgh, Institute for Computer Based Learning, pp. 9–11. [http://www.icbl.hw.ac.uk/ltdi]. Visited 11.12.2002.

Marks, R.B. 2000. *Determinants of Student Evaluations of Global Measures of Instructor and Course Value*. Journal of Marketing Education, Vol. 22 (2), pp. 108–119.

Marsh, H.W. 1987. *Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research.* International Journal of Educational Research, Vol 11 (3), pp. 253–388.

Marsh, H.W. 2001. *Students' Evaluations of University Teaching*. Presented as part of an invited lecture at Minho University, Braga Portugal on 13 June 2001. 26 p.

Marsh, H.W. & Bailey, M. 1993. *Multidimensional Students' Evaluations of Teaching Effectiveness: A Profile Analysis.* Journal of Higher Education, Vol. 64 (1), pp. 1–18.

Marsh, H.W. & Hocevar, D. 1991. *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, Vol 7 (1), pp. 9–18.

Marton, F. 1982. *Towards a Phenomenography of Learning I. Integrating experiental aspects.* University of Göteborg, Department of Education, N:o 06, 17 p.

Mason, P.M., Steagall, J.W. & Fabritius, M.M. 1995. *Student Evaluations of Faculty: A New Procedure for Using Aggregate Measures of Performance.* Economics of Education Review, Vol. 14 (4), pp. 403–416.

Mc Caig, R.D. 2002. *Publisize or Perish? College Faculty Perspectives on the Validity and Public Access to Student Ratings of Teaching Effectiveness.* Doctoral Thesis. Philadelphia, Temple University. 206 p.

McAlpine, L. & Harris, R. 2002. *Evaluating teaching effectiveness and teaching improvement: A language for institutional policies and academic development practices.* International Journal for Academic Development. Vol. 7 (1), pp. 7–17.

McCormack, C. 2005. *Reconceptualizing student evaluation of teaching: an ethical framework for changing times.* Assessment & Evaluation in Higher Education, Vol. 30 (5), pp. 463–476.

McCormack, C., Applebee, A. & Donnan, P. 2003. *Opening a Can of Worms: A Conversation about the Ethics of Online Student Evaluation of Teaching*, The Technology Source July/August 2003. [http://technologysource.org/article/opening_a_can_of_worms/], Visited 11.11.2005.

McGhee, D.E. & Lowell, N. 2003. *Psychometric Properties of Student Ratings of Instruction in Online and on-Campus Courses.* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 39–48.

McGourty, J., Scoles, K. & Thorpe, S. 2002a. *Web-Based Student Evaluation of Instruction: Promises and Pitfalls.* Paper presented at the 42nd Annual Forum of the Association for Institutional Research, Toronto, June 2–4, 11 p.

McGourty, J., Scoles, K. & Thorpe, S. 2002b. *Web-based Course Evaluation: comparing the experience at two universities*. Paper presented at the 32nd ASEE/IEEE Frontiers in Education Conference, November 6–9 2002, Boston, 6 p.

McGourty, J., Shuman, L., Besterfield-Sacre, M., Hoare, R., Wolfe, H., Olds, B. & Miller, R. 2001. *Using Technology to Enhance Outcome Assessment in Engineering Education.* 31st ASEE/IEEE Frontiers in Education Conference, October 10–13, Reno. [http://fie.engrng.pitt.edu/fie2001/papers/1267.pdf]. Visited 08.04.2005.

McGuire, M.D. & Casey, J.P. 1999. *Using Comparative Alumni Data for Policy Analysis and Institutional Assessment*. New Directions for Institutional Research, Issue 101, pp. 81–99.

McKeachie, W.J. 1996. *Student Ratings of Teaching.* In: England, J., Hutchings, P. & McKeachie, W.J. The Professional Evaluation of Teaching. American Council of Learned Societies, Occasional Paper No. 33, [http://www.acls.org/op33.htm#McKeachie], Visited 26.01.2005.

McKeachie, W.J. & Kaplan, M. 1996. *Persistent Problems in Evaluating College Teaching.* AAHE Bulletin 48 (6), pp. 5–8.

McKenzie, J., Sheely, S. & Trigwell, K. 1998. *Drawing on Experience: an holistic approach to student evaluation of courses.* Assessment & Evaluation in Higher Education. Vol. 23 (2), pp. 153–164.

McKinney, K. 1997. *What Do Student Ratings Mean?* The National Teaching & Learning Forum, Vol. 7 (1), pp. 2–4.

McKinnon, J. 1988. *Reliability and Validity in Field Research: Some Strategies and Tactics.* Accounting, Auditing & Accountability Journal, Vol. 1 (1), pp. 34–54.

McKone, K.E. 1999. *Analysis of student feedback improves instructor effectiveness.* Journal of Management Education. Vol. 23 (4), pp. 396–415.

Mehrens, W.A. 1998. *Consequences of Assessment: What is the Evidence?* Education Policy Analysis Archives, Vol. 6 (13). [http://epaa.asu.edu/epaa/v6n13.html]. Visited 5.12.2005.

Menges, R.J. 1991. *The Real World of Teaching Improvement: A Faculty Perspective*. New Directions for Teaching and Learning, No. 48, Winter 1991, pp. 21–38.

Menges, R. 2000. *Shortcomings of Research on Evaluating and Improving Teaching in Higher Education*. New Directions for Teaching and Learning, No. 83, Fall 2000, pp. 5–11.

Mertler, C. 2003. Patterns of Response and Nonresponse from Teachers to Traditional and Web Surveys. Practical Assessment, Research & Evaluation, Vol. 8 (22), [http://pareonline.net/getvn.asp?v=8&n=22], Visited 28.04.2005.

Millea, M. & Grimes, P.W. 2002. *Grade Expectations and Student Evaluations of Teaching.* College Student Journal, Vol. 36 (4), pp. 582–560.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G. & Johnson, L. 2002. *Making Sense of Data From Complex Assessments.* Applied Measurement in Education, Vol. 15 (4), pp. 363–389.

Moss, J. & Hendry, G. 2002. *Use of electronic surveys in course evaluation.* British Journal of Educational Technoloy. Vol. 33 (5), pp. 583–592.

Muffo, J.M., Sinclair, A. & Robson, V. 2003. *A Comparison of Web Versus Paper Alumni Surveys*. Paper presented at the Annual Forum of the Association for Institutional Research, Tampa, Florida, May 20 2003, [http://www.aap.vt.edu/Alumni_Assessment/Web_Survey_paper.htm], Visited 17.03.2005.

Mutch, A. 2002. *Thinking Strategically about Assessment.* Assessment & Evaluation in Higher Education, Vol. 27 (2), pp. 163–174.

Nasser, F. & Fresko, B. 2002. *Faculty Views of Student Evaluation of College Teaching.* Assessment & Evaluation in Higher Education, Vol. 27 (2), pp. 187–198.

Nasser, F. & Fresko, B. 2006. *Predicting student ratings: the relationship between actual student ratings and instructors' predictions.* Assessment & Evaluation in Higher Education, pp. 1–18.

Neilimo, K. & Näsi, J. 1980. *Nomoteettinen tutkimusote ja suomalainen yrityksen taloustiede. Tutkimus positivismin soveltamisesta.* Yrityksen taloustieteen ja yksityisoikeuden laitoksen julkaisuja, Sarja A2: 12. Tampereen yliopinston keskusmonistamo, Tampere, 82 p.

Nevgi, A. & Lindblom-Ylänne, S. 2003. *Opetuksen suunnittelun työkalut.* In: Lindblom-Ylänne, S. & Nevgi, A. (eds.). Yliopisto- ja korkeakouluopettajan käsikirja. Helsinki, WSOY, 505 p.

Nulty, D. 2001. *Web On-Line Feedback (WOLF): Intentions and evaluation.* Refereed Proceedings of Teaching Evaluation Forum, Perth, Australia, 28th – 29th August 2000, The University of Western Australia, pp.38–41.

Ogier, J. 2005. *Evaluating the Effect of a Lecturer's Language Backgroung on a Student Rating on a Teaching Form,* Assessment & Evaluation in Higher Education, Vol. 30 (5), pp. 477–488.

Olivares, O.J. 2001. *Student Interest, Grading Leniency, and Teacher Ratings: A Conceptual Analysis*. Contemporary Educational Psychology, Issue 26, pp. 382–399.

Oliver, M. & Conole, G. 1998. *Selecting a methodology*. In: Harvey, J. (ed.). 1998. *Evaluation Cookbook.* Edinburgh, Institute for Computer Based Learning.pp. 12–13. [http://www.icbl.hw.ac.uk/ltdi]. Visited 11.12.2002.

Oliver, R.L. & Sautter, E.P. 2005. *Using Course Management Systems to Enhance the Value of Student Evaluations of Teaching.* Journal of Education for Business, Vol. 80 (4), pp. 231–234.

Olkkonen, T. 1993. *Johdatus teollisuustaloudelliseen tutkimustyöhön*. Report No 152, Teknillinen korkeakoulu, Teollisuustalous ja työpsykologia. TKK Offset, Otaniemi, 143 p.

OnSET. 2005. *Online Student Evaluation of Teaching in Higher Education.* [http://onset.byu.edu/]. Visited 19.05.2005.

Ory, J.C. 2000. *Teaching Evaluation: Past, Present, and Future*. New Directions for Teaching and Learning, Issue 83, Fall 2000, pp. 13–18.

Ory, J.C. 2001. *Faculty Thought and Concerns About Student Ratings.* New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 3–15.

Ory, J.C. & Ryan, K. 2001. *How Do Student Ratings Measure Up to a New Validity Framework?* New Directions for Institutional Research, Issue 109 (Spring 2001), pp.27–44.

Ouimet, J.A., Bunnage, J.C., Carini, R.M., Kuh, G.D. & Kennedy J. 2004. *Using focus groups, expert advice and cognitive interviews to establish the validity of a college survey.* Research in Higher Education, Vol. 45 (3), pp. 233–250.

Pajarre, E. 1997. *Tietokoneistetun kurssiarvioinnin kehittäminen TTKK:ssa*. Master of Science Paper, Tampere University of Technology. 107 p.

Pajarre, E. 2001. *Tutkinnon sisältö, rakenne ja läpäistävyys – raportti TTKK:sta vuonna 2000 valmistuneiden antamasta palautteesta*. TTKK Laatutyöryhmä, Tampere, 31 p.

Pajarre, E. 2002. *Raportti tutkinnon suorittamista hidastavista tekijöistä TTKK:ssa*. TTKK Laatutyöryhmä, Tampere, 28 p.

Parjanen, M. 2003. *Amerikkalaisen opiskelija-arvioinnin soveltaminen suomalaiseen yliopistoon.* Korkeakoulujen arviointineuvoston julkaisuja 8:2003. Helsinki, Edita Publishing Oy. 49 p.

Paswan, A.K. & Young, J.A. 2002. *Student Evaluation of instructor: A Nomological Investigation Using Structural Equation Modeling.* Journal of Marketing Education, Vol. 24 (3), pp. 193–202.

Patton, M.Q. 1996. *A world larger than formative and summative.* Evaluation Practice, Vol. 17 (2), pp. 131– 144.

Patton, M.Q. 1997. *Utilization–Focused Evaluation. The New Century Text.* Thousand Oaks, Sage Publications, 431 p.

Pedersen, E.L. 1983. *Computer-assisted evaluation of student papers: I can write anything you can write – faster and better.* CALICO Journal Vol. 1 (2), pp. 39–42.

Pellegrino, J.W., Chudowsky, N. & Glaser, R. (eds.). 2001. *Knowing what Students Know, The Science and Design of Educational Assessment*. Washington DC, National Academy Press, 366 p.

Platt, M. 1993. *What Student Evaluations Teach.* Perspectives on Political Science, Vol. 22 (1), pp. 29–40.

Popham, W.J. 1981. *Modern Educational Measurement.* Englewood Cliffs, Prentice-Hall, 441 p.

Porter, S.R. & Whitcomb, M.E. 2003. *The Impact of Lottery Incentives on Student Survey Response Rates.* Research in Higher Education, Vol. 44 (4), pp. 389–407.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J.M. & Singer, E. 2004. *Methods for testing and evaluating survey questions.* Public Opinion Quarterly, Vol. 68 (1), pp. 109–130.

Punch, K.F. 2005. *Introduction to Social Research. Quantitative and Qualitative Approaches.* Wiltshire, SAGE Publications, 320 p.

QAA. 2000. *Code of practice for the assurance of academic quality and standards in higher education. Section 6: Assessment of students – May 2000,* [http://www.qaa.ac.uk/academicinfrastructure/codeOfPractice/default.asp], Visited 23.03.2005.

Raivola, R. 2000. *Tehoa vai laatua koulutukseen?* Juva, WSOY. 237 p.

Ramsden, P. 1991. *A Performance Indicator of Teaching Quality in Higher Education: the Course Experience Questionnaire*. Studies in Higher Education, Vol. 16 (2), pp. 129–150.

Ramsden, P. 2003. *Learning to Teach in Higher Education.* London, RoutledgeFarmer, 288 p.

Rando, W.L. 2001. *Writing Teaching Assessment Questions for Precision and Reflection.* New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 77–83.

Ranki, A. 1999. *Vastaako henkilöstön osaaminen yrityksen tarpeita?* Kauppakaari. Gummerus. Jyväskylä. 174 p.

Rautopuro, J. & Väisänen, P. 2000. *Mikä vie ilon opiskelusta? Opintojen kokeminen Joensuun yliopistossa.* In Honkimäki, S. & Jalkanen, H. 2000. Innovatiivinen yliopisto? Koulutuksen tutkimuslaitos, Jyväskylän yliopisto. Jyväskylän yliopistopaino, Jyväskylä. 265 p.

Recker, M.M. & Greenwood, J. 1995. *An Interactive Student Evaluation System*. Hypermedia Proceedings of AUUG95 and Asia-Pacific World Wide Web ’95 Conference & Exhibition, Sydney, 19–21 September 1995, [http://www.csu.edu.au/special/conference/apwww95/papers95/mrecker/mrecker.html]. Visited 29.12.2005.

Remmers, H.H., Gage, N.L. & Rummel, J.F. 1966. *A Practical Introduction to Measurement and Evaluation*. A Harper International Student Reprint, 390 p.

Rigsby, B. & Smith, C. 1999. *Evaluation using the Anonymous Feedback system.* Teaching & Education News, Vol. 9 (4). [http://www.tedi.uq.edu.au/TEN/TEN_previous/TEN4_99/index.html]. Visited 16.12.2004.

Romano, M.F. & Himmelmann, M. 2002. *Determinants of Web mode choice in a "Web and paper" survey in a high education population.* A paper presented at the International Conference on Improving Surveys, ICIS, Copenhagen, August 25–28 2002, [http://www.icis.dk/ICIS_papers/C2_6_2.pdf], Visited 06.04.2005.

Ruskai, M.B. 1997. *Evaluating Student Evaluations*. Notices of the American Mathematical Society, Vol. 44 (3), p. 308.

Rust, C. 2001. *Basic assessment issues and terminology*. The Higher Education Academy. 6 p. Internet article, [http://www.heacademy.ac.uk/resources.asp?process=full_record&section=generic&id=436], Visited 17.08.2005.

Ryan, K. 2002. *Shaping Educational Accountability Systems*. American Journal of Evaluation, Vol. 23 (4), pp. 453–468.

Salminen, R.T. & Mirola, T. 2001. *Is it Worth Developing University Courses Pedagogically? Some Pedagogical Experiments in the Marketing Course Business Relationships and Networks*. Proceedings of ANZMAC2001 Conference, Massey University, Albany Campus, Auckland, New Zealand 3rd – 5th December 2001.

Saroyan, A. & Amundsen. C. 2001. *Evaluating University Teaching: time to take stock.* Assessment & Evaluation in Higher Education. Vol. 26 (4), pp. 341–353.

Savander-Ranne, C. 2003. *An Active Role for Students in the Learning Process in Engineering Education: A Means to Develop Conceptual Understanding.* Doctoral Thesis, Tampere, Tampere University of Technology, TTY-Paino, 55 [+75] p.

Sax, L. J., Gilmartin, S.K. & Bryant, A.N. 2001. *Is Online Survey Administration the Answer? Findings from a National Study of Survey Methodologies.* Annual meeting of the Association for the Study of Higher Education (ASHE), Richmond, VA, USA. November, 2001. 45 p.

Sax, L.J., Gilmartin, S.K. & Bryant, A. 2003. *Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys.* Research in Higher Education, vol. 44 (4), pp. 409–432.

Schmidt, W.C. 1997. *World-Wide Web Survey Research: Benefits, Potential Problems, and Solutions.* Behavior Research Methods, Instruments & Computers, Vol. 29 (2), pp. 274–279.

Schnell, R. 2002. *Antworten auf Nonresponse.* Vortrag auf dem XXXVII Kongress der Deutschen Marktforschung am 7. Mai 2002 in Wolfsburg. 14 p.

Scoles, K. 2000. *A New Course Evaluation Process*. IEEE Transactions on Education, Vol. 43 (2), pp. 125–131.

Scriven, M. 1991. *Evaluation Thesaurus.* Fourth Edition. Newbury Park, Sage Publications, 391 p.

Scriven, M. 1995. *Student ratings offer useful input to teacher evaluations.* Practical Assessment, Research & Evaluation, Vol. 4 (7), [http://www.ericdigests.org/1997-1/ratings.html], Visited 03.12.2004.

Scriven, M. 1996. *Types of Evaluation and Types of Evaluator*. Evaluation Practice, Vol 17 (2), pp. 151–161.

Scriven, M. 2001. *Evaluation: Future Tense.* American Journal of Evaluation, Vol. 22 (3), pp. 301–307.

Scriven, M. 2003. *Evaluation Theory and Metatheory*. In: Kellaghan, T. & Stufflebeam, D.L. (eds.). International Handbook of Educational Evaluation, Part One: Perspectives, Cornwall, Kluwer Academic Publishers, pp. 15–30.

Seldin, P. 1993. *The Use and Abuse of Student Ratings of Professors*. Chronicle of Higher Education, July 21, p. A40.

Seppänen-Järvelä, R. 1999. *Kehittämistyö ja arviointi*. In: Eräsaari, R., Lindqvist, T., Mäntysaari, M. & Rajavaara, M. Arviointi ja asiantuntijuus. Gaudeamus. Tampere, Tammer-Paino Oy, pp. 90–105.

Serva, M.A. & Fuller, M.A. 2004. *Aligning What We Do and What We Measure in Business Schools: Incorporating Active Learning and Effective Media Use in the Assessment of Instruction*, Journal of Management Education, Vol. 23 (1), pp. 19–38.

Shannon, D.M., Johnson, T.E., Searcy, S. & Lott, A. 2002. *Using Electronic Surveys: Advice from Survey Professionals.* Practical Assessment, Research & Evaluation, Vol. 8 (1). [http://pareonline.net/getvn.asp?v=8&n=1]. Visited 15.12.2004.

Sheehan, K. 2001. *E-mail Survey Response Rates: A Review.* Journal of Computer-Mediated Communication 6(2). [http://www.ascusc.org/jcmc/vol6/issue2/sheehan.html]. Visited 15.12.2004.

Sheehan, K.B. 2002. *Online Research Methodology: Reflections and Speculations*. Journal of Interactive Advertising, Vol. 3 (1), [http://www.jiad.org/vol3/no1/sheehan/], Visited 17.03.2005.

Shevlin, M., Banyard, P., Davies, M. & Griffiths, M. 2000. *The Validity of Student Evaluation of Teaching in Higher Education: love me, love my lectures?* Assessment & Evaluation in Higher Education, Vol. 25 (4), pp. 397–405.

Smith, M.C. 1990. *Students' Perceptions of the Teaching Evaluation Process.* Paper presented at the annual meeting of the American Educational Research Association, Boston. 11p.

Soininen, M. 1997. *Kasvatustieteellisen evaluaation perusteet.* Turun yliopiston täydennyskoulutuskeskuksen julkaisuja A:56. Turku, Pallosalama Oy, 156 p.

Sojka, J., Gupta, A.K. & Deeter-Schmelz, D.R. 2002. *Student and Faculty Perceptions of Student Evaluations of Teaching. A Study of Similarities and Differences.* College Teaching, Vol. 50 (2), pp. 44–49.

Solomon, D.J. 2001. *Conducting Web-Based Surveys*. Practical Assessment, Research & Evaluation, Vol. 7 (19). [http://pareonline.net/getvn.asp?v=7&n=19]. Visited 15.12.2004.

Sorenson, D.L. & Reiner, C. 2003. *Charting the Uncharted Seas of Online Student Ratings of Instruction.* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 1–24.

Spencer, K.J. & Schmelkin, L.P. 2002. *Student Perspectives on Teaching and its Evaluation.* Assessment & Evaluation in Higher Education, Vol. 27 (5), pp. 397–409.

Sproule, R. 2000. *Student Evaluation of Teaching: A Methodological Critique of Conventional Practices.* Educational Policy Analysis Archives, Vol 8 (50), [http://epaa.asu.edu/epaa/v8n50.html], Visited 03.01.2006.

Stake, R.E. & Cisneros-Cohernour, E.J. 2000. *Situational Evaluation of Teaching on Campus.* New Directions for Teaching and Learning, Issue 83, Fall 2000, pp. 51–72.

Stapleton, R.J. & Murkison, G. 2001. *Optimizing the Fairness of Student Evaluations: A Study of Correlations between Instructor Excellence, Study Production, Learning*

Production, and Expected Grades. Journal of Management Education, Vol. 25 (3), pp. 269–291.

Stephens, D., Bull, J. & Wade, W. 1998. *Computer-assisted Assessment: suggested guidelines for an institutional strategy.* Assessment & Evaluation in Higher Education. Vol. 23 (3), pp. 283–294.

Stufflebeam, D. 2003. *The CIPP Model for Evaluation.* In: Kellaghan, T. & Stufflebeam, D.L. (eds.). International Handbook of Educational Evaluation, Part One: Perspectives, Cornwall, Kluwer Academic Publishers, pp. 279–302.

Svinicki, M.D. 2001. *Encouraging Your Students to Give Feedback.* New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 17–24.

Takala, M. 1993. *Koulutusohjelman palautejärjestelmä*. M. Sc. Thesis. Helsinki University of Technology. 107 p.

Theall, M. 2000. *Electronic Course Evaluation Is Not Necessarily the Solution.* The Technology Source, November/December 2000. [http://ts.mivu.org/default.asp?show=article&id=823]. Visited 13.12.2004.

Theall, M. 2001. *Can We Put Precision into Practice? Commentary and Thought Engendered by Abrami's "Improving Judgments About Teaching Effectiveness Using Teacher Rating Forms".* New Directions for Institutional Research, Issue 109 (Spring 2001), pp.89–96.

Theall, M. & Franklin, J. 1991. *Using Student Ratings for Teaching Improvement.* New Directions for Teaching and Learning, No. 48, Winter 1991, pp. 83–98.

Theall, M. & Franklin, J. 2000. *Creating Responsive Student Ratings Systems to Improve Evaluation Practice.* New Directions for Teaching and Learning, Issue 83, Fall 2000, pp. 95–107.

Theall, M. & Franklin, J. 2001a. *Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?* New Directions for Institutional Research, Issue 109 (Spring 2001), pp.45–56.

Theall, M. & Franklin, J. 2001b. *Using Technology to Facilitate Evaluation*. New Directions for Teaching and Learning, Issue 88, Winter 2001, pp. 41–50.

Theall, M., Scannell, N. & Franklin, J. 2000. *The eye of the beholder: individual opinion and controversy about student ratings.* Instructional Evaluation and Faculty Development, Vol. 20 (1), [http://www.umanitoba.ca/academic_support/uts/sigfted/iefdi/spring00/matrix.htm]. Visited 01.02.2007.

Thorpe, S.W. 2001. *Linking Learning Outcomes to Student Course Evaluation.* Paper presented at 28th Annual Conference of Northeast Association for Institutional Research, Boston, November 2001.

Thorpe, S.W. 2002. *Online Student Evaluation of Instruction: An Investigation of Non-Response Bias*. Paper presented at the 42nd Annual Forum of the Association for Institutional Research, Toronto, Canada, [http://www.drexel.edu/provost/ir/conf/bias.pdf]. Visited 05.11.2004.

Tiberius, R. 2001. *Making Sense and Making Use of Feedback from Focus Groups*. New Directions for Teaching and Learning, Issue 87, Fall 2001, pp. 63–75.

Toegel, G. & Conger, J.A. 2003. *360-Degree Assessment: Time for Reinvention*. Academy of Management Learning and Education, Vol. 2 (3), pp. 297–311.

Toivonen, T. 1999. *Empiirinen sosiaalitutkimus. Filosofia ja metodologia.* WSOY, Helsinki, 449 p.

Toland, M.D. & de Ayala, R.J. 2005. *A Multilevel Factor Analysis of Students' Evaluations of Teaching.* Educational and Psychological Measurement, Vol. 65 (2), pp. 272–296.

Tomsic, M. L., Hendel, D.D. & Matross, R.P. 2000. *A World Wide Web Response to Student Satisfaction Surveys: Comparisons Using Paper and Internet Formats*. Paper presented at the 40th Annual Meeting of the Association of Institutional Research, Cincinnati, May 21–24, 19 p.

Truell, A.D. 2003. *Use of Internet Tools for Survey Research.* Information Technology, Learning, and Performance Journal. Vol. 21 (1), pp. 31–37.

Tucker, B., Jones, S., Straker, L. & Cole, J. 2003. *Course Evaluation on the Web: Facilitating Student and Teacher Reflection to Improve Learning.* New Directions for Teaching and Learning, Issue 96, Winter 2003, pp. 81–93.

Turner, G.M. 2004. *The Role of Cognitive Schemas in a Web-Based Student Evaluation of Teaching system: Usability Issues of Design and Implementation.* Doctoral Thesis. The University of Texas at Austin. 95 p.

Turner, W.L. & Stylianou, A.C. 2004. *The IT advantage assessment model: Applying an expanded value chain model to academia*. Computers & Education. Vol. 43 (3), pp. 249–272.

TUT. 2007. *TUT in a Nutshell.* [http://www.tut.fi/public/]. Visited 30.03.2007.

Tynjälä, P., Salminen, R., Sutela, T., Nuutinen, A. & Pitkänen, S. 2004. *Tärkeimmät opintomenestykseen yhteydessä olevat tekijät Lappeenrannan teknillisessä yliopistossa – Oppimisen kokonaismallin pohjalta toteutettu kyselytutkimus.* Lappeenrannan teknillinen yliopisto, Hallinnon julkaisuja 146. 22 p.

Tynjälä, P., Salminen, R. T., Sutela, T., Nuutinen, A. & Pitkänen, S. 2005. *Factors related to study success in engineering education*. European Journal of Engineering Education, Vol. 30 (2), pp. 221–231.

Upcraft, M.L. & Wortman, T.I. 2000. *Web-based Data Collection and Assessment in Student Affairs.* Student Affairs On-Line, Vol. 1 (3).
[http://www.studentaffairs.com/ejournal/Fall_2000/art1.html]. Visited 16.12.2004.

Uusitalo, H. 1995. *Tiede, tutkimus ja tutkielma. Johdatus tutkielman maailmaan*. Juva, WSOY, 121 p.

Vartiainen, P. 2005. *Institutional Tendencies of Legitimate Evaluation: A Comparison of Finnish and English Higher Education Evaluations.* Higher Education in Europe, Vol. 30 (3–4), pp. 371–384.

Vuorenmaa, M. 2001. *Ikkunoita arvioinnin tuolle puolen. Uusia avauksia suomalaiseen koulutusta koskevaan evaluaatiokeskusteluun.* Dissertation, Jyväskylän yliopisto, 266 p.

Välimaa, J. 2000. *Ulkoinen itsearviointi ja käyttötypologiat.* In: Honkimäki, S. & Jalkanen, H. Innovatiivinen yliopisto? Koulutuksen tutkimuslaitos, Jyväskylän yliopisto, Jyväskylän yliopistopaino, 265 p.

Wachtel, H.K. 1998. *Student evaluation of college teaching effectiveness: a brief review.* Assessment & Evaluation in Higher Education. Vol. 23 (2), pp. 191–211.

Watt, S. & Simpson, C., McKillop, C. & Nunn, V. 2002. *Electronic Course Surveys: does automating feedback and reporting give better results?* Assessment & Evaluation in Higher Education, Vol. 27 (4), pp. 325–337.

Weiss, C.H. 1996. *Excerpts from Evaluation Research: Methods of Assessing Program Effectiveness[1]*, Evaluation Practice, Vol. 17 (2), pp. 173–175.

Weissberg, R. 1993. *Standardized Teaching Evaluations*. Perspectives on Political Science, Vol. 22 (1), pp. 5–7.

Welsh, J.F. , Alexander, S. & Dey, S. 2001. *Continuous Quality Measurement: restructuring assessment for a new technological and organisational environment*. Assessment & Evaluation in Higher Education, Vol 26 (5), pp. 391–401.

Wholey, J.S. 1996. *Formative and Summative Evaluation: Related Issues in Performance Measurement.* Evaluation Practice, Vol. 17 (2), pp. 145–149.

Wicks, A.C. & Freeman, R.E. 1998. *Organization Studies and the New Pragmatism: Positivism, Anti-positivism, and the Search for Ethics.* Organization Science, Vol. 9 (2), pp. 123–140.

Wilhelm, W.B. 2004. *The Relative Influence of Published Teaching Evaluations and Other Instructor Attributes on Course Choice*. Journal of Marketing Education, Vol. 26 (1), pp. 17–30.

Wilson, R. 1998. *New Research Casts Doubt on Value of Student Evaluations of Professors*, The Chronicle of Higher Education, Vol 44 (19), pp. A12–A14.

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, Vol. 27 (1), pp. 49–64.

Xenos, M. 2004. *Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks.* Computers & Education. Vol. 43 (4), pp. 345–359.

Yao, Y. 2001. *How Do Faculty Use Student Evaluation Feedback? A multiple case study.* Doctoral Thesis. Lincoln, The Graduate College at the University of Nebraska. 202 p.

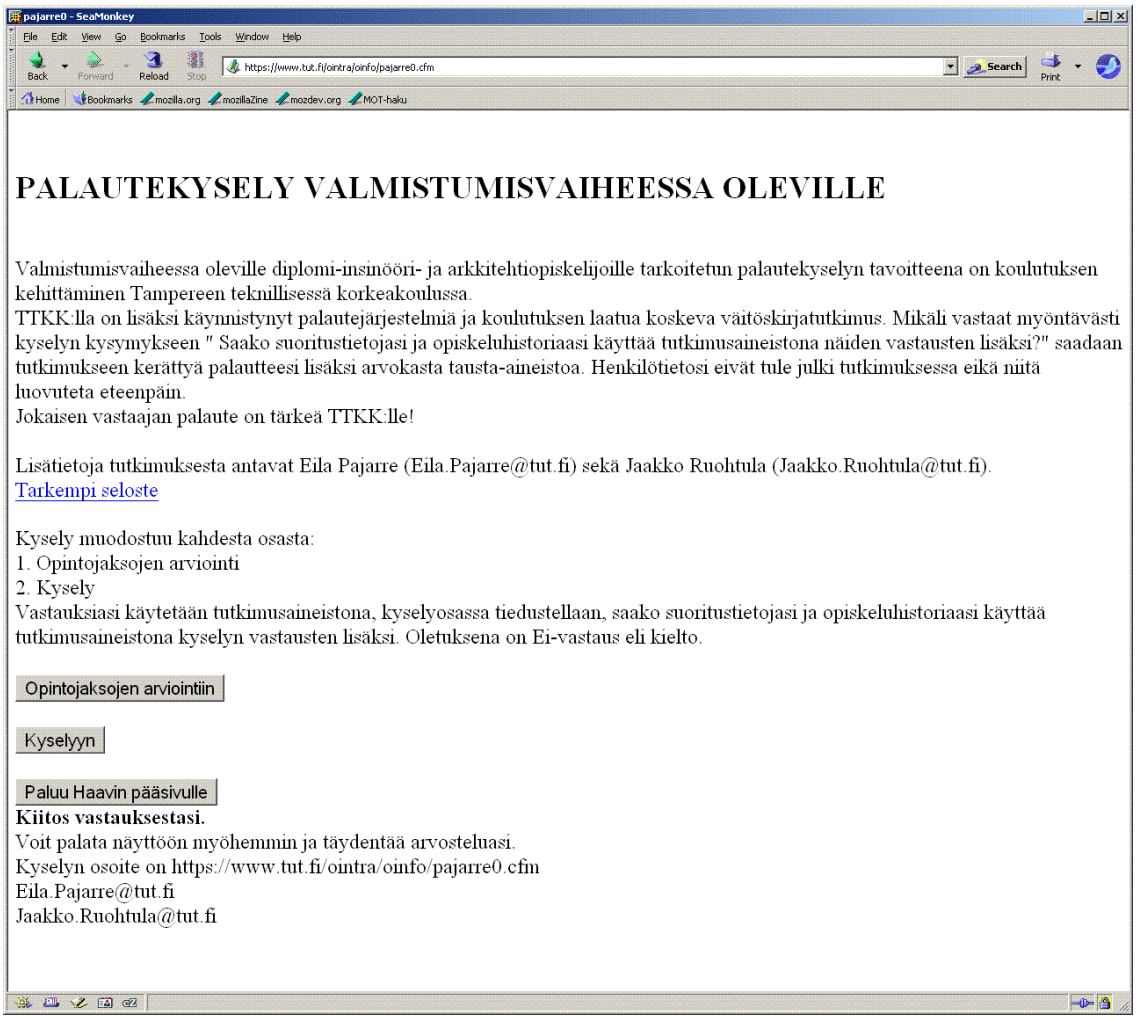
Young, R.D. 1993. *Student Evaluation of Faculty: A Faculty Perspective.* Perspectives on Political Science, Vol. 22 (1), pp. 12–17.

Young, I.P, Delli, D.A., Johnson, L. 1999. *Student Evaluation of Faculty: Effects of Purpose on Pattern.* Journal of Personnel Evaluation in Education, Vol. 13 (2), pp. 179–190.

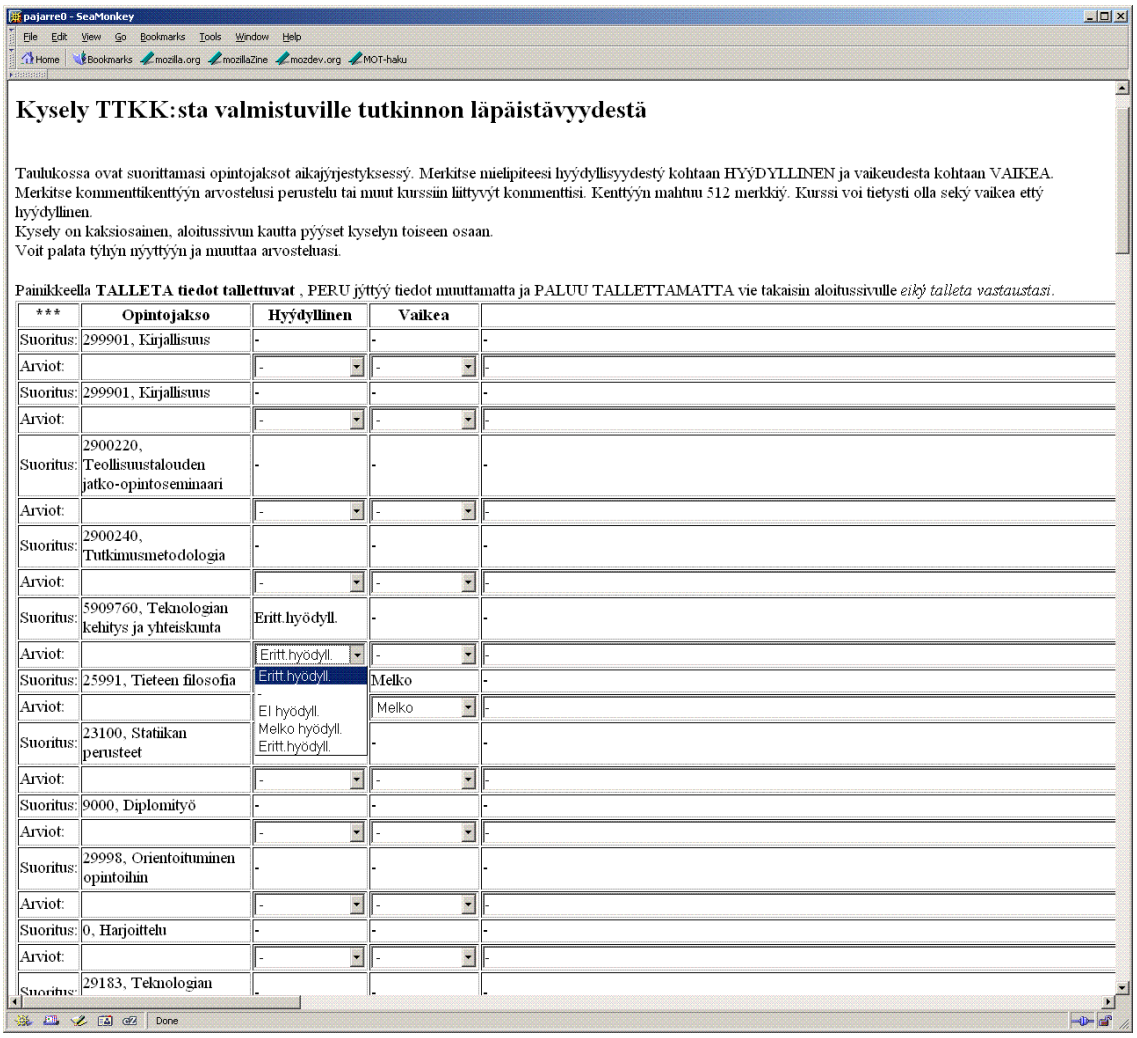
Yorke, M. 1998. *The Management of Assessment in Higher Education.* Assessment & Evaluation in Higher Education, Vol. 23 (2), pp. 101–116.

Zimitat, C. & Crebert, G. 2002. *Conducting online research and evaluation.* 2002 Annual International Conference of the Higher Education Research and Development Society of Australasia. pp. 761–769.

The front page of the case survey



An example of the survey's course specific questions (each responding student received a unique questionnaire page where those and only those courses she/he had executed were listed)



pajarre0 - SeaMonkey

File Edit View Go Bookmarks Tools Window Help

Home Bookmarks mozilla.org mozillaZine mozdev.org MOT-haku

Kysely TTKK:sta valmistuville tutkinnon läpäistävyydestä

Taulukossa ovat suorittamasi opintojaksot aikajärjestyksessä. Merkitse mielipiteesi hyödyllisyydestä kohtaan HYÖDYLLINEN ja vaikeudesta kohtaan VAIKEA. Merkitse kommenttikenttään arvostelusi perustelu tai muut kurssiin liittyvät kommenttisi. Kenttään mahtuu 512 merkkiä. Kurssi voi tietysti olla sekä vaikea että hyödyllinen.
Kysely on kaksiosainen, aloitussivun kautta pääset kyselyn toiseen osaan.
Voit palata tähän näyttöön ja muuttaa arvosteluasi.

Painikkeella **TALLETA tiedot tallettuvat** , PERU jättää tiedot muuttamatta ja PALUU TALLETTAMATTA vie takaisin aloitussivulle *eikä talleta vastaustasi*.

***	Opintojakso	Hyödyllinen	Vaikea	
Suoritus:	299901, Kirjallisuus	-	-	-
Arviot:		-	-	-
Suoritus:	299901, Kirjallisuus	-	-	-
Arviot:		-	-	-
Suoritus:	2900220, Teollisuustalouden jatko-opintoseminaari	-	-	-
Arviot:		-	-	-
Suoritus:	2900240, Tutkimusmetodologia	-	-	-
Arviot:		-	-	-
Suoritus:	5909760, Teknologian kehitys ja yhteiskunta	Eritt.hyödyll.	-	-
Arviot:		Eritt.hyödyll.	-	-
Suoritus:	25991, Tieteen filosofia	-	Melko	-
Arviot:		Ei hyödyll.	Melko	-
Suoritus:	23100, Statiikan perusteet	Melko hyödyll.	-	-
Arviot:		Eritt.hyödyll.	-	-
Suoritus:	9000, Diplomityö	-	-	-
Arviot:		-	-	-
Suoritus:	29998, Orientoituminen opintoihin	-	-	-
Arviot:		-	-	-
Suoritus:	0, Harjoittelu	-	-	-
Arviot:		-	-	-
Suoritus:	29183, Teknologian	-	-	-

Done

Contents of the survey's general part

Kysely TTKK:sta valmistuville tutkinnon läpäistävyydestä

Vastauksiasi käytetään tutkimusaineistona, viimeisenä kysymyksenä tiedustellaan, **saako** suoritustietojasi ja opiskeluhistoriaasi käyttää tutkimusaineistona kyselyn vastausten lisäksi. Oletuksena on **Ei**-vastaus eli kielto.
Kysely on kaksiosainen, aloitussivun kautta pääset kyselyn suoritusten arviointiosaan.
Voit palata tähän näyttöön ja muuttaa arvosteluasi.

Painikkeella **TALLETA tiedot tallettuvat**, PERU jättää tiedot muuttamatta ja PALUU TALLETTAMATTA vie takaisin aloitussivulle *eikä talleta vastaustasi.*

Taustaa

***	***
K:	Koulutusta tukeva työkokemus kk.
V:	-
K:	Muu työkokemus kk.
V:	-
K:	Onko tutkinnon rakenteessa jotain, joka hankaloittaa opiskelun etenemistä? Esim. perusopinnot, koulutusohj.-kohtaiset opinnot, ammattiaineet, muuta?
V:	-
K:	Ehdotus tilanteen korjaamiseksi?
V:	-

Pääaine

***	**Vastaus max. 500 merkkiä**
K:	Jos valitsisit uudelleen pääaineesi, mikä olisi valintasi nyt?
V:	-
K:	Miksi?
V:	-

Kieliopinnot

***	***
K:	Oliko sinulla mahdollisuus valita riittävästi kieliopintoja?
V:	-

Opintoja vaikeuttavat tekijät

***	***
K:	Lukujärjestykset (luennot ti-to, päällekkäisyydet...)
V:	-
K:	Tentit, tenttitulosten viivästyminen, harjoitustöiden palautusten viipyminen.
V:	-
K:	Opintoneuvonnan puute tai vähäisyys.
V:	-
K:	Opettajien tavoitettavuus.
V:	-
K:	Muu palvelujen saatavuus.
V:	-
K:	Muualla suoritettujen opintojen hyväksilukeminen.
V:	-
K:	Muut asiat.
V:	-

Esimerkkejä vaikeuttavista tekijöistä:

***	**Vastaus max. 500 merkkiý**
K:	Kerro esimerkkejä:
V:	-

Muuta ja aineiston käyttölupa

***	**Vastaus max. 500 merkkiä**
K:	Mitä muuta haluaisit tuoda esiin opiskelusta TTKK:lla?
V:	-
K:	Minkä kokonaisarvosanan antaisit saamastasi opetuksesta TTKK:ssa?
V:	-

***	***
K:	Saako suoritustietojasi ja opiskeluhistoriaasi käyttää tutkimusaineistona näiden vastausten lisäksi?
V:	Ei

Talleta Peru

Paluu tallettamatta

Kiitos vastauksestasi.

Voit palata näyttöön myöhemmin ja täydentää arvosteluasi.

Eila.Pajarre@tut.fi
Jaakko.Ruohtula@tut.fi

Basic statistics

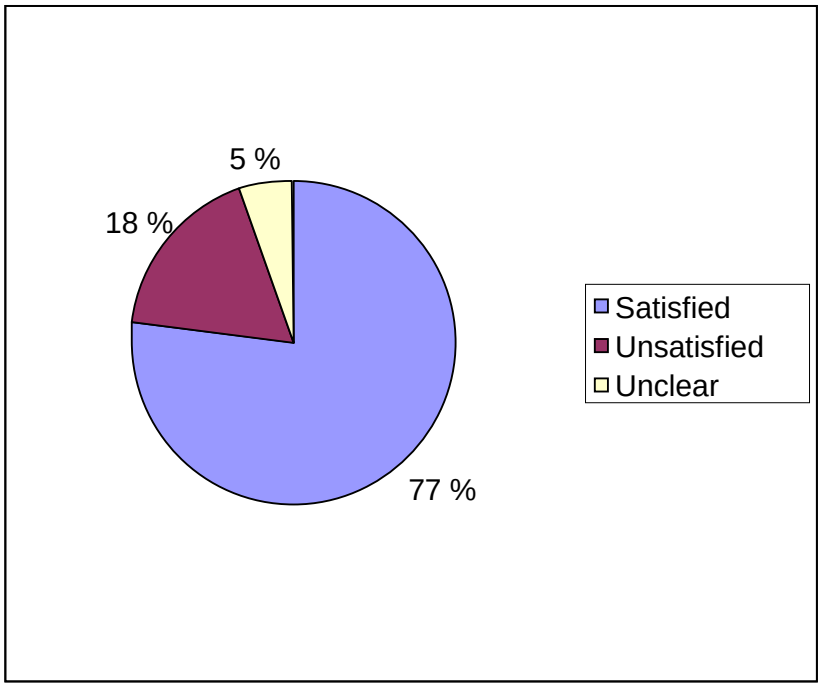
Deviation of respondents' (R_{bg}) study programme and gender.

Study Programme	Number of respondents allowing use of their background information	Male (R_{bg})	Female (R_{bg})	Male (all students graduating during survey execution time)	Female (all students graduating during survey execution time)
Architecture	1	0 % (n=0)	100 % (n=1)	58 % (n=19)	42 % (n=14)
Automation Engineering	8	87 % (n=7)	13 % (n=1)	90 % (n=76)	10 % (n=8)
Civil Engineering	3	67 % (n=2)	33 % (n=1)	70 % (n=23)	30 % (n=10)
Electrical Engineering	25	80 % (n=20)	20 % (n=5)	83 % (n=160)	17 % (n=32)
Environmental and Energy Technology	7	29 % (n=2)	71 % (n=5)	42 % (n=20)	58 % (n=28)
Fibre, Textile and Clothing Science	4	0 % (n=0)	100 % (n=4)	9 % (n=2)	91 % (n=20)
Industrial Engineering and Management	18	83 % (n=15)	17 % (n=3)	66 % (n=78)	34 % (n=40)
Information and Knowledge Management	-	-	-	67 % (n=2)	33 % (n=1)
Information technology	34	85 % (n=29)	15 % (n=5)	91 % (n=234)	9 % (n=93)
Materials Engineering	10	50 % (n=5)	50 % (n=5)	49 % (n=32)	51 % (n=33)
Mechanical Engineering	15	87 % (n=13)	13 % (n=2)	95 % (n=105)	5 % (n=6)
Science and Engineering	6	33 % (n=2)	67 % (n=4)	53 % (n=10)	47 % (n=9)
Total	**131**	**73 % (n=95)**	**27 % (n=36)**	**77 % (n=761)**	**23 % (n=225)**

Distribution of estimations given by respondents (R_{bg}) on the usefulness of each course they had executed, connected with their actual final grades from the same courses retrieved from university's database (M.Sc. theses included).

	Final course grade	**Number of given estimations**
Very useful	1	104
	2	210
	3	446
	4	571
	5	465
	Approved (no grade)	265
Quite useful	1	303
	2	509
	3	776
	4	716
	5	579
	Approved (no grade)	348
Not useful	1	164
	2	184
	3	234
	4	202
	5	126
	Approved (no grade)	178

Distribution of satisfaction to the selected main subject among respondents (n= 152).



Chi-Square tests analysing the differences in respondent groups R_{bg} and R_{den} responses:[1]

Difficulty:

	Column 1	Column 2	Column 3	Total
Row 1	2339	3117	921	6377
Row 2	1411	1769	518	3698
Total	3750	4886	1439	10075

Degrees of freedom: 2
Chi-square = 2.20619523442031
For significance at the .05 level, chi-square should be greater than or equal to 5.99.
The distribution is not significant. *p* is less than or equal to 1.

[1] Calculated by a web Chi-Square calculator available at http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html.

Usefulness:

	Column 1	Column 2	Column 3	Total
Row 1	1088	3231	2061	6380
Row 2	800	1790	1157	3747
Total	1888	5021	3218	10127

Degrees of freedom: 2
Chi-square = 28.8161171038188
p is less than or equal to 0.001.
The distribution is significant.

Example of respondents' attitudes toward the usefulness and difficulty of a single obligatory course (Fundamental University Physics Part I) in TUT.

Grade	*Amount of given estimations*	*Amount of given estimations*	*Amount of given estimations*
5	not useful, n=2 easy, n=8	quite useful, n=11 quite difficult, n=12	very useful, n=7 very difficult, n=0
4	not useful, n=6 easy, n=9	quite useful, n=4 quite difficult, n=4	very useful, n=3 very difficult, n=0
3	not useful, n=2 easy, n=3	quite useful, n=14 quite difficult, n=14	very useful, n=2 very difficult, n=1
2	not useful, n=2 easy, n=1	quite useful, n=10 quite difficult, n=9	very useful, n=1 very difficult, n=3
1	not useful, n=4 easy, n=0	quite useful, n=3 quite difficult, n=7	very useful, n=1 very difficult, n=1

List of courses most often evaluated as being both very difficult and very useful

- Software Engineering Project
- Introductory Course on Programming
- Data Structures and Algorithms
- Introduction to Software Engineering
- Basic Swedish
- Management Accounting
- Software Engineering Methodology

Factors hindering the respondents (Rbg) progressing in their studies analysed according to gender.

Factor hindering studies	**Gender of Respondent (Rbg)**	**Did hinder**	**Did NOT hinder**	**Not answered**
Schedules	Female n=36	75 %	25 %	-
	Male n=95	74 %	26 %	-
Exams, delays in receiving results from exams, …	Female n=36	47 %	53 %	-
	Male n=95	46 %	54 %	-
Lack or insufficiency in student counselling	Female n=36	47 %	53 %	-
	Male n=95	25 %	75 %	-
Reachability of lecturers	Female n=36	31 %	69 %	-
	Male n=95	20 %	80 %	-
Availability of other services	Female n=35	17 %	83 %	n=1
	Male n=95	5 %	95 %	-
Receiving approval for studies executed in other universities…	Female n=32	16 %	84 %	n=4
	Male n=78	9 %	91 %	n=17
Other issues	Female n=23	30 %	70 %	n=13
	Male n=72	21 %	79 %	n=23

Respondent groups' R_{bg} and R_{den} opinions of factors hindering their progress in their studies in TUT.

Factor hindering the studies	Respondent group	did hinder	did NOT hinder
Schedules	R_{bg} (n=131) R_{den} (n=45)	74 % 69 %	26 % 31 %
Exams, delays in receiving results from exams, …	R_{bg} (n=131) R_{den} (n=45)	47 % 49 %	53 % 51 %
Lack or insufficiency in student counselling	R_{bg} (n=131) R_{den} (n=45)	31 % 27 %	69 % 73 %
Reachability of lecturers	R_{bg} (n=131) R_{den} (n=44)	23 % 18 %	77 % 82 %
Availability of other services	R_{bg} (n=130) R_{den} (n=45)	8 % 18 %	92 % 82 %
Receiving approval for studies executed in other universities…	R_{bg} (n=110) R_{den} (n=38)	11% 11 %	89 % 89 %
Other issues	R_{bg} (n=95) R_{den} (n=36)	23 % 39 %	77 % 61 %

Distribution of respondents (R_{bg}) all given course estimations (usefulness and difficulty) and the respective final grades from the same estimated courses (N=6306).

Usefulness & Difficulty, n	%	Grade	f (grade)	% (grade)
Not useful & very difficult	2.85	1	65	36
n=180		2	44	24
		3	39	22
		4	20	11
		5	7	4
		Approv.	5	3
Not useful & easy	8.21	1	23	4
n=518		2	55	11
		3	98	19
		4	112	22
		5	84	16
		Approv.	146	28
Not useful & quite difficult	6.14	1	75	19
n=387		2	87	22
		3	96	25
		4	69	18
		5	35	9
		Approv.	25	6
Very useful & very difficult	4.58	1	39	3
n=289		2	54	19
		3	61	21
		4	70	24
		5	43	15
		Approv.	22	08
Very useful & easy	9.32	1	10	2
n= 588		2	27	5
		3	106	8
		4	158	27
		5	155	26
		Approv.	132	22
Very useful & quite difficult	17.81	1	55	5
n=1123		2	129	11
		3	271	24
		4	307	27
		5	251	22
		Approv.	110	10
Quite useful & very difficult	6.90	1	102	23
n=435		2	114	26
		3	95	22
		4	64	15
		5	38	9
		Approv.	22	5
Quite useful & easy	19.41	1	31	3
n=1224		2	111	9
		3	274	22
		4	302	25
		5	276	23
		Approv.	230	19
Quite useful & quite difficult	24.77	1	170	11
n=1562		2	284	18
		3	405	26
		4	345	22
		5	263	17
		Approv.	95	6