Quality in Statistics Education

Determinants of Student Outcomes in Methods & Statistics Education at Universities and Colleges

Pieternella S. Verhoeven

March 1, 2009

Verzorging omslag: Cunera, Amsterdam

ISBN 978 90 473 0090 8
NUR 741

www.boomonderwijs.nl

Quality in Statistics Education
Determinants of Student Outcomes in Methods & Statistics Education at Universities and Colleges

Kwaliteit in Statistiekonderwijs
Indicatoren van studieresultaten in Methodenleer & Statistiek aan Universiteiten en Colleges
(met een samenvatting in het Nederlands).

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 19 maart 2009 des middags te 2.30 uur

door

Pieternella Susanna Verhoeven

geboren op 31 maart 1961 te Utrecht

Promotoren:

Prof. dr. M.J. de Jong
Prof. dr. J.J. Hox

Dit proefschrift werd (mede) mogelijk gemaakt met financiële steun van Boom Onderwijs te Amsterdam.

Dankwoord / Acknowledgments

Een promotietraject kan nooit worden volbracht zonder de hulp van een grote groep mensen. Ook ik ben een aantal mensen veel dank verschuldigd! Echter, bij het samenstellen van het lijstje met namen voor mijn dankwoord bekroop mij het akelige gevoel dat ik nooit écht iedereen kan bedanken. Dat is altijd het risico bij het maken van lijstjes: dat je mensen vergeet. Toch wil ik hier een poging wagen om namen te noemen. Mocht je je naam niet terugvinden, jullie hulp staat in m'n hart gegrift!

Allereerst een woord van dank voor beide promotoren. Joop Hox verbaasde mij keer op keer met z'n positieve reacties op mijn analyseresultaten. Zeker als ik vermoedde dat de modellen te complex waren voor interpretatie. Mart-Jan de Jong slaagde er keer op keer in om mij op te beuren, tijdens de vele gesprekken die wij voerden. Hij liet me mijn werk vanuit een andere invalshoek bekijken, en met succes.

Aan dit promotie-onderzoek is door 11 instellingen van hoger onderwijs meegewerkt. Thierry Marchant, Jan Koster, Pieter Koele, Arie van Peet, Tjaart Imbos, Dirk Tempelaar, Geert Loosveldt, Jan Degadt en Bert Nijdam gaven interviews en verleenden hun medewerking bij het verzamelen van de data in hun colleges. Een speciaal woord van dank voor mijn 'nestor' in de Statistiek, Bert Nijdam. Jij leerde me om Statistiek als vak te waarderen! Ik zal je colleges nooit vergeten.

Vele nabije en verre collega's zijn behulpzaam geweest tijdens het phd-traject. Richard van den Doel, Dirk Vries en Henk Meijer hebben getracht me de zieleroerselen van LaTeX bij te brengen. Ik ben er nog steeds niet achter, maar zonder hen was het manuscript niet af gekomen. Joseph Resovsky en Marcin Sklad namen me werk uit handen tijdens de cursus, zodat ik aan m'n proefschrift kon werken. Alle collega's in 'Eleanor' en daarbuiten toonden van tijd tot tijd veel belangstelling voor mijn vorderingen; ze luisterden geduldig naar mijn klaagzang. Anya Luscombe, Gonny Pasaribu en Diederik van Werven adviseerden mij bij vertaalproblemen. Diederik en Ineke hielpen bij het 'proeflezen' van de Nederlandse samenvatting. Stef van Buuren in Leiden gaf kritisch en opbouwend commentaar op mijn voorlopige resultaten dat resulteerde in een frisse kijk op de methode- en resultatensectie (en uiteindelijk in het herschrijven ervan). Rolf Steyer reageerde altijd prompt en uitgebreid op mijn vragen over latente verschilmodellen. Ik heb veel van hem geleerd.

Veel steun ondervond ik van de collega's uit Maastricht en Jena met wie ik gedurende presentaties, workshops, conferenties en promoties discussieerde over de stand van het statistiekonderwijs. Hun interesse in mijn onderzoek en hun kritische vragen hielpen me de conclusies scherp te stellen. Carola Hageman, Martine Harsema, Barbara Kuiper, Ingrid Straten en Nico Buitendijk adviseerden bij alle zaken die het uitgeven van een manuscript met zich meebrengt.

Dan zijn er natuurlijk de studenten. Allereerst Zlata Mironova die geduldig de data van 2.555 studenten in diverse SPSS-spreadsheets invoerde. Mijn studenten in de 110, 210 en 310 van de afgelopen jaren luisterden geduldig naar

wéér zo'n voorbeeld over 'statsfobia' en 'houding ten aanzien van statistiek'; Jehonathan Ben, Mona Irrmischer, Yajing Zhu en Job van Tilburg die tijdens de cursus 'geavanceerde modelbouw' een aantal modellen succesvol voor me uittestten.

In moeilijke tijden leer je je echte vrienden kennen. Da's maar weer waarheid gebleken. Geduldig accepteerden ze een aantal malen 'nee' op een uitnodiging. Speciaal wil ik Candace Schau bedanken, die zich niet alleen een goede vriendin toonde, maar die haar huis, kantoor, school en expertise beschikbaar stelde zodat ik in de zomer van 2007 goede vorderingen met m'n analyse kon maken.

Dankwoorden zoals deze eindigen altijd met de familie. Dat is eigenlijk niet terecht, ze zouden ermee moeten beginnen. Tegelijkertijd besef ik dat het schrijven van zo'n dankwoord een fractie is van mijn werkelijke gevoel. Dus, hier gaat 'ie: mijn ouders die hielpen waar dat kon, met proeflezen, of simpelweg door te luisteren naar m'n verhalen. Sharon en Sander die onmiddellijk mijn paranimfen wilden worden. Sander die altijd z'n radio zachter zette als 'mama weer moest werken' en die als atletiekcoach fungeerde indien ik ontspanning zocht in het hardlopen.

Ten slotte is Jan Willem er altijd. Zonder hem was ik dit traject niet gestart; zonder hem had ik het zeker niet volbracht.

Middelburg, februari 2009.

Contents

Chapter 1

Introduction

1.1 Methods & Statistics in the Educational Program

Many freshmen at University take Methods & Statistics in their first year. This is partly due to the fact that Methods & Statistics is compulsory as a core course e.g. for students who major in Social Science. Most Methods & Statistics courses are offered within the framework of the bachelor / master program of Social Scientific Departments, such as Sociology, Psychology and Pedagogy. At most Liberal Arts & Sciences colleges, Introductory Statistics is compulsory for every freshman. This means that Methods & Statistics courses are not only taught to students who actually choose the course, but also to students from a broad spectrum of fields that must take the course in order to meet the requirements for the school they attend.

Universities consider mastery of Methods & Statistics to be part of essential academic skills. These skills involve the mastery of research processes such as the collection and analysis of data, interpretation and presentation of results and conclusions. Students 'learn to think' and 'learn to communicate' within these introductory courses. They learn how to develop a research question and a research plan and - most of all - they learn how to analyze (quantitative) data. Most of the course objectives are driven by 'learning to think', but part of the course is also based on 'learning how to communicate', as students have to present exercises and reports, and work together in groups (Weltje-Poldervaart et al., 2001). In this sense, acquiring 'statistical skills' can be looked upon as 'complex skills' in the definition of Van Merrienboer (1997; see also Hoogveld, 2003) according to the Theory of Instructional Design (Merrill, 1991).

The learning and teaching of these complex skills, especially methods & statistics, has become a grateful researchtopic over the years. It is also the topic for this thesis, and in this first chapter I will explain why. First, I will sketch the development of the teaching of statistics over the past decades, followed by a description of current statistics teaching and learning. After discussing some

contemporary expert views on statistics education, I will explain the focus of this study.

1.2 History of Introductory courses in Methods & Statistics

Technological developments in the past decades have resulted in ever faster information exchanges. This is one of the reasons why education in statistics has become more and more important. In this section the recent development with respect to Statistics Education will be discussed, especially the shift towards a more active learning & teaching attitude, both in the USA and the Netherlands. The developments in the USA are being described because they started earlier in the US than in Europe. Because similar developments took place on this side of the ocean, a lot of universities and colleges use the US developments on academic teaching as an example.

United States of America
The development of Introductory courses in Methods & Statistics in the United States can be divided into roughly two episodes, one during the first half of the 20th century with large groups of students and classic books such as Snedecor's book from 1937 '*Statistical Methods*'. The second episode starts towards the end of the 20th century with small groups and active learning processes.

During the first period the emphasis was on teaching scientists and lecturers assuming that their students were quantitatively skilled. During the 1960s the ideas on teaching introductory statistics began to change[1]. First of all data analysis became a more independent scientific activity, second a number of suitable analytical tools were introduced and students were no longer compelled to spend hours *'behind a mechanical calculator'*.

Towards the end of the seventies, a new episode started with the publication of two statistics books by Freedman (Statistics, 1978) and Moore (Statistics: Concepts and Controversies, 1978). This era is known as the era of the modern introductory statistics courses (Aliage et al., 2005). As Aliage states, two other trends coincided with these changes: first the growth in enrollment and second the introduction of placement tests. As the emphasis on statistical applications shifted from a conceptual to a more active approach, the importance of being able to understand and interpret statistical output increased and over the years the number of requirements for Introductory courses in Statistics rose. This also resulted in a shift from a highly motivated and quantitatively skilled student population to a population of students that only took the course in order to meet the departmental requirements and, furthermore, the latter were not quantitatively skilled. Cobbs' article on 'teaching statistics' (1992) is seen as

[1]Note that in most cases methods & statistics are taught as one course. In some cases however, statistics is taught as a separate course. In this study the emphasis lies on exploring ways and combinations in which statistics is taught. When relevant, a distinction between the two types of courses will be made.

the main driving force behind the changes of the past 10 to 15 years. Cobb recommends that the statistical curriculum should change to emphasize statistical thinking (see also Bryce, 2005), use fewer recipes and formulas, rely more on automatic computations and real-life data analysis and to make more use of active learning tools instead of lecturing (Cobb, 1992). His suggestions started a major change in statistics education.

In 2001, the American Statistical Association launched a set of curriculum guidelines that emphasized this constructivist way of looking at Introductory Statistics, as will be discussed in section 1.4.2. The guidelines are meant for undergraduate programs at universities and colleges, especially for students who major in statistics or who take a minor in statistics. These guidelines distinguish between mathematics and statistics, as statistics is considered to be a more practical education in statistical reasoning. Besides a number of skills e.g. computer skills, mathematical (to a certain extent) and statistical skills, students have to master more general academic skills such as writing and presenting, doing team projects et cetera. Furthermore, students also have to master a few methodological skills (research skills) in the field of study (AmStat, 2001). As these guidelines have been developed for students who major in statistics, they will not be further discussed, because this goes beyond the scope of the subject for this study. However, it is apparent that a constructivist approach is widely accepted and has taken a strong position in the current curricula. Before discussing this constructivist approach, an insight in the Dutch developments will be described.

Dutch developments

In the Netherlands, similar developments took place (i.e. a shift in pedagogy and learning), although the introduction of statistics in the curriculum was always much later than in the USA, hence the American student population is much younger. E.g in the Netherlands, it was not until the 1970s that Introductory courses in Statistics were introduced to the upper levels of the mathematics curriculum in secondary schools (Bakker, 2004) with a strong emphasis on mathematical formulas, concepts and computations. It mainly focused on probability and contrary to the United States, it was taught at a college level.

Recent developments in the field of Introductory Statistics show a shift towards a new content, new teaching and learning methods, the use of technology and the focus on descriptive tools and practical applications (Bakker, 2004, 10). The emphasis is placed more on active learning, on statistical reasoning, rather than a branch of mathematics, on working with real-life data and on computer software as a tool. This results in less computations and formulas, less theory and more applications, student projects, displaying and interpreting results. This new pedagogy is considered a more constructivist approach (Bakker, 2004). In the next section, this approach will be described. One of the results of the shift towards statistical reasoning instead of formulas and concepts is the use of graphics in an exploratory manner. Using real-life data, both instructor and students more often look at graphical displays of data in order to detect patterns and relationships, before actually computing a coefficient or test. The use of suitable software to accomplish this has become imperative.

Also, more emphasis is placed on the interpretation of real life results, although the mastery of key statistical concepts remains mandatory.

1.2.1 Implementation of a new structure - The 'Bologna Declaration'

After having signed the 'Bologna Declaration' in 1999, many European countries reformed their system of higher education, in order to increase transparency, international mobility and recognition.

The Netherlands
Currently, the implementation of the Bachelor-Master structure is considered one of the most drastic changes in Higher Education during the last decade in the Netherlands. Furthermore, in an attempt to establish a better connection with International Master- and PhD programs, a number of Liberal Arts & Sciences Colleges has been founded. These colleges offer a 3-year Bachelor's program on a number of majors and minors. After having completed this program, students graduate as Bachelor of Science or Bachelor of Arts, with which they can enroll in an International Master- or PhD-program.

Flanders
In Flanders, Higher Education consists of University Education and nonuniversity education (HOBU). University programs in most cases takes 4 years and non-university programs on average last 3 years. University programs tend to be theory oriented, while non-university programs are more likely to be professionally and vocationally oriented (Duchesne & Nonneman, 1998, 211).

The Flemish universities offer a three year Bachelor program and a one year Masters' program. In some cases (e.g. Civil Engineer) the program takes five years (3 years Bachelor and 2 years Masters). In the previous system the program was divided into a 2 year 'candidate', followed by a 2 year Flemish degree (licentiaat).

Non-universities in most cases offer two programs. First they offer a short, three year Bachelor degree such as nurse, secretary. The long program consists of offering a four year Master program (3 years Bachelor, 1 year Masters) like in University, but with a professional and vocational emphasis (e.g. Industrial Engineer). The difference between a University Master and a non-university Master is the focus on a more vocational side of the profession[2].

1.3 Systems of teaching statistics - an overview

Before looking at an appropriate theoretical model to explain the effects of student- and institutional determinants on student achievement, an overview of the current insights with regard to teaching introductory Statistics is offered.

[2]For instance, a Civil Engineer is a 5 Year University Master, and an Industrial Engineer is a 4 Year non-university Master.

After discussing a few paradigms regarding teaching statistics, I will give some examples of current stereotypes.

1.3.1 Paradigms in teaching Methods & Statistics

According to Roiter & Potecz (1996) there are four main paradigms when it comes to teaching statistics:

1. Statistics as a branch of Mathematics, with weekly lectures on combinatorics, probability theory, proof, derivations and lots of formulas. Assessment consists of mid-term exams and final exams.

2. Statistics as the Analysis of Data, with weekly lectures and lab-classes and much group interaction. This approach consists of statistical techniques (regression, correlation, hypothesis testing etc) and the assessment consists of lab-tests, assignments and exams.

3. Statistics as Experimental Design, with lots of discussion and group interaction. Usually it consists of critical reviews of existing literature and a real-life experiment (mostly done by small groups). Assessment takes place through papers, presentations, lab tests and sometimes exams (without mathematical content).

4. Statistics as a problem-based subject, again with a lot of group discussion, project work, more coaching than lecturing. Students work in small groups, solving problems within their own field of interest (their major). Assessment takes place through report writing, presentations, essays, journals (Roiter & Potecz, 1996).

In sum these paradigms take a model-driven or a data-driven approach to teaching statistics, the main difference being that the first places more emphasis on mathematical concepts than the latter, and the latter emphasizes 'learning by doing'.

It is a great challenge to make, or better yet to keep Methods & Statistics attractive and learnable for all students who enroll (Sowey, 1995, 2001; Ahlgren, 2001). Many teachers have already tried to make Methods & Statistics more attractive, by using 'motivational tools' (Chance, 1997) such as:

- using 'real-life' examples and experiments
- having students perform their own research project
- having students do projects applicable in their own field
- using interactive forms of teaching
- combining lectures with computer labs, group discussions, and other interactive forms of teaching

Moreover, during the past decades, views on teaching introductory statistics have changed focus from instruction to learning. This means that instead of seeing learning as 'receiving lecture material from the instructor' the view shifted towards 'experiencing the material by students'(Steinhorst & Keeler, 1995).

Garfield et al. (2000) summarized recommendations for Statistics teaching on the basis of best practices. They came to the conclusion that the emphasis should lie on the mastering of Statistical Thinking by means of active learning methods, usage of real life examples and less usage of recipes and derivations.

Strategies for teaching complex subjects like 'Methods & Statistics' are constantly changing (Moore, 1997). As described above, university teachers find new and better didactic models, student populations change, educational and organizational settings change, as do government regulations. However, the most important goal of teaching Methods & Statistics remains the same: the development of Statistical Skills and Knowledge (competences) by students at University.

1.3.2 Stereotypes in Statistics Education

Methods & Statistics is not very popular with many students (Garfield et al., 2002a, 2002b). Students find the course very difficult, it scares them to work with statistical software or formulas, they think it is tedious, the examples or the books used do not appeal to them. A fair amount of Methods & Statistics still consists of reading a fair amount of statistical text, calculating variables, computing methods and interpreting test results (Aliage et al., Diamond, 2002). Little emphasis is put on teaching statistical thinking or reasoning (Melton, 2004). Furthermore, compulsory courses such as Methods & Statistics often encounter problems with students who major in a completely different field. Hence, many students show signs of statistics anxiety or 'statsphobia', i.e. *a lack of confidence in quantitative abilities* (Bradsheet, 1996). There are several current stereotypes regarding Methods & Statistics.

Statistics literacy?

First of all, many students do not have Statistics literacy. Statistics literacy (sometimes referred to as 'numerical literacy, see Bradsheet, 1996) can be defined as the *ability and propensity to interpret, critically evaluate, and communicate about statistical information, data-related claims, or chance-related phenomena which they may encounter in diverse life contexts* (Shield, 2002). Students at a later stage are expected at least to be well informed about basic statistics and, furthermore, be able to understand basic statistical information. In this respect, it is important that students at least acquire some basis statistical skills. Hence, it is expected that many students are actually statistical illiterates. This influences the way in which methods & statistics must be taught. Hence, it must be made learnable for students at every level of statistical knowledge and skills.

Statistics as 'service teaching'

Second Methods & Statistics courses are usually offered as 'service teaching'

(Sowey, 1995). This means that students are required to take this course in order to meet the requirements for a major or minor. However, the methods and examples used in this course are not commonly applied in their field of interest. In many cases, Methods & Statistics is characterised by formulas, theory, methods and computations, it is so-called 'theory-driven' (Moore, 1992). Applications of formulas are usually presented in one specific field, and they are not adaptable to every-day life. It is difficult to find subject matter that is relevant to the entire student group (Bradsheet, 1996)

Daily use of statistics
Third, the use of Statistics in everyday life seems unclear to many students. This argument links to the last one, because the teaching of Methods & Statistics is often not guided by everyday applications (Yilmaz, 1996). Instructors use notorious examples from the 'great researchers' regarding experimental settings, validation of questionnaires et cetera. So, because students do not learn how to apply Statistics to every day life by learning from every-day examples, they often feel that Statistics is something they will not use again when they leave university.

My primary aim is to analyze to what extent these stereotypical ideas hold true and, if so, to make recommendations regarding what can be done to change them. Before being able to give any recommendations as to changing these attitudes, I need to determine what causes these attitudes and to what extent this is affecting student outcomes, such as final grade.

1.4 Teaching and Learning Statistics - is there a 'best way'?

Let us move on to the next question: if Methods & Statistics has to be taught to (and learned by) students in other majors and fields of interest and expertise, what is the best way to do this according to experts? In this section, I will describe the ideas of a number of experts in the field and shed some light on the ongoing discussion between statistics literacy, reasoning and thinking.

1.4.1 Introductory Statistics - a subject in itself

How is Introductory Methods & Statistics best taught? Over the past decades numerous articles have been published on this subject, as many instructors and researchers described their best way of teaching 'Introductory Methods & Statistics. There probably is no 'best way' to teach Methods & Statistics. Numerous assumptions can be made as to the nature of 'good Introductory Statistics'. What is considered the best way to transfer knowledge and skills with respect to Methods & Statistics, especially to freshmen?

Moore (1992) describes a number of key points with respect to teaching Methods & Statistics. He starts off with a comparison between Statistics and

Mathematics. The assumed resemblance between the two is considered one of the main reasons why some students are so anxious in the Introductory Statistics course. Statistics is a distinct subject, with its own substance leaning on mathematical formulas.

Most colleges and universities treat their Introductory Methods & Statistics courses as service teaching (Moore, 1992). This means that the courses are not a field of interest in themselves, but they serve other fields such as economics, biology, sociology and psychology. That is why these courses are taught within the Academic Core departments of many Liberal Arts & Science Colleges. This means that many students do not choose to take the introductory course themselves, but it is a prerequisite for other fields of interest. The result of all this is that teachers are often confronted with students who did not want to take Statistics, who are 'statistics illiterate', or who have 'statsphobia'.

This places special emphasis on the choices of didactic methods and teaching approaches. That's why in most cases a theory driven approach does not work. Theory driven means that the teacher explains statistical models by using mathematical formulas as a starting point. Teachers should take a more practical approach and only add formulas and calculations if it is absolutely essential for students in order to understand a statistical topic.

But what is so different between Statistics and Mathematics? Moore (1992) explains the difference as follows:

- Statistics has its own subject matter. Statistics is looked upon as the 'science of data'. Students learn to structure data in such a way, that they can detect a pattern in those data and report this pattern in a quantitative manner. They learn how to look at data in their own context.
- Statistics is a distinct discipline that does not originate in mathematics. The origins lie in the attempt to combine observations in surveying and astronomy (Moore, 1992, 2).
- However, there is a relationship between mathematics and statistics, as statistics uses mathematical tools from mathematics.

Hence, an important educational factor to take into account is the extent to which the didactical approach has a mathematical focus. The assumption is that the less focus is placed on mathematics, the more students in the introductory course will be able to acquire methodological and statistical knowledge and skills, and the more suitable an introductory course of that kind will be as 'service teaching'(Moore, 1992; also see Moore & Cobb, 2000).

1.4.2 A constructivist viewpoint

Moore (1992, 1997) has a few suggestions. In his model, he proposes a shift from a theoretical approach (called 'information transfer') towards a more direct way of learning by students. This constructivist view means that students work actively towards constructing their own knowledge and combining it with the

knowledge they already have (Moore, 1997). This means that students actually learn by doing, by performing research. This educational view that emerged during the first half of the 1990s not only changes the perspective for students, it also changes the teachers' perspective. Instead of telling students what to learn, teachers discuss the content and coach them in doing class exercises (Moore, 1997; Bryce, 2005), a shift towards a **variety** of learning activities; a shift from teacher to student, from teaching to learning. Hence, statistics is best learned when explained in an applied environment. This view was adopted by the American Statistical Association in 2001.

This viewpoint is in line with the constructivist approach to 'Instructional Design' where the teaching and learning of new concepts is founded on existing knowledge. Students 'construct' their own knowledge on what they already know about a topic. This approach has advantages for both teaching and learning such complex skills as 'statistics' (see also Hoogveld, 2003; Karagiorgi & Symeou, 2005).

With regard to teaching, Cobb & Moore (1997, 820), for instance, oppose the more 'probability driven approach to statistics education' that probability theory should not be a part of Introductory courses in Methods & Statistics and mathematical formulas should be kept to a minimum. This approach is also known as the 'data driven' approach to statistics education (Roiter & Petocz, 1996).

Where the content of the course is concerned, a number of suggestions have been made over the years:

1. Do not use too many formulas. Teach concepts first, then methods.

2. The 'cookbook' approach does not work. Instead, use real-life data and place it into context, as realistic data or simulations motivate students to learn (Bradsheet, 1996).

3. The emphasis is on reasoning, the interpretation of results. Of course, some calculating is involved, but because of all the automated systems the calculations can be be kept to a minimum.
   Moore divided the content of Introductory Methods & Statistics into three parts

   (a) Organization and summary of data, where students learn to look at data and to uncover patterns using statistical descriptions and graphic displays.

   (b) Production of data, where students familiarize themselves with techniques such as sampling and setting up data collection methods.

   (c) Inference, where students learn how to interpret tests and how to look upon (causal) relations between data using concepts such as significance.

4. Move from simple to complex, first use descriptions and graphical displays before you turn to probability and testing. Step by step, students

will learn how to choose and use methodological and statistical tools and instruments.

5. Only use more complex theory when it is absolutely necessary with respect to probability and inference. Those two main statistical topics are essential but their complexity often makes it difficult for students to understand. Teachers need to make these subject understandable by using interesting and easy real-life examples and data (also see Cobb & Moore, 1997; Moore, 1992, 7-11).

In the constructivist approach students do not learn individually but their learning is embedded in the social environment. Learning therefore is considered a collaborative process, where students develop complex skills by comparing their own perspective to that of other students (Karagiorgi & Symeou, 2005).

Bradsheet (1996) summarizes constructivist ideas by suggesting that students in introductory courses in Methods & Statistics should be taught to learn statistical reasoning before computing methods and learning complex formulas. Nonstatisticians tend to feel the link between mathematics and statistics, although it is kept to a minimum. Statistical reasoning is defined as *the way people reason with statistical ideas and make sense of statistical information*(Garfield & Gal, 1999; Garfield, 2002). The main question that remains is '*What should students learn and how should they learn it?* (Moore, 1997)'. This means that technology serves pedagogy, and that mathematics serves statistics only when it is absolutely necessary to understand statistical concepts. In that way, Introductory Methods & Statistics becomes more feasible, understandable and enjoyable for all students, not only for a small mathematically oriented group.

1.4.3 Applying Statistical Ideas in Educational settings

A distinction has to be made between statistical literacy, reasoning and thinking. Why? Because developers and lecturers of Introductory Methods & Statistics, scientists discover the special place that these courses have in the college-curriculum (ARTIST, 2002a, 2002b). First of all, students who enroll in a mandatory Introductory Statistics course come from various educational backgrounds. This means that in high school they could have undertaken a 'science program', or a more 'social scientific' background. In any case, Dutch students have to show proficiency in at least a basic level of mathematics. Unfortunately, Statistics is not a part of this basic course. Having chosen a higher level of mathematics means that Dutch students will have had some introduction in the very basics of Statistics, but no more than that. However, previous studies show that a certain experience in mathematics helps understanding statistics at the undergraduate level (see for instance Gal & Garfield, 1997). In Flanders, pupils at secondary schools also get acquainted with statistics (Vlaams ministerie van Onderwijs, 2008)[3].

[3]In Flanders the level of statistics at 'ASO' is also basic, as pupils learn some univariate descriptive statistics, inferential statistics, concepts such as 'validity' and basic bivariate ana-

Another specific feature of some of the institutions in this study is, that they offer education in the 'Liberal Arts and Sciences'. This entails a broad introduction into the whole scale of Academic courses, resulting in a bachelor degree with a particular major. Every first year student has to take statistics, whether they major in Arts & Humanities or in Science. This obligation results in a very unequal entrance level at the start of the Introductory Statistics course. It takes special skills and preparation, in order to motivate such a diverse group of students, especially in Statistics.

Another reason why preparations for such a course need special skills is that due to the mandatory nature of this service course, it has always been looked upon by students as unpleasant and difficult (Garfield et al.. 2000). This is one of the main reasons why the teaching method and the (possible) change in attitudes are so important, with a focus on statistical thinking - literacy or -reasoning. But which of the three focal points should teachers take? Or is there another concept that teachers should focus on?

1.4.4 Statistical reasoning, thinking or literacy?

Garfield et al. (2000) propose a focus on statistical thinking. They define 'statistical thinking' as *'the thought process that recognizes that variation is all around us and present in everything we do*' (Snee, 1990). Snee defined 'statistical thinking' as a series of processes, such as identifying, quantifying, controlling and reducing. As a result, students learn how to solve a (statistical) problem. This view, however valuable, is considered too narrow for the approach in this study, because it refers to student thinking, and not so much to doing. In my view, statistical thinking is a very important prerequisite for a broader understanding of statistics.

Schield (2002) favors a focus on 'statistical literacy'. He defines statistical literacy as 'critical thinking about statistics as evidence for inferences (Schield, 2004). In his view instructors should focus on descriptive statistics and modeling, interpreting tables and graphs, inductive inference, et cetera. Although this definition focuses on 'literacy' the term 'thinking' covers a large part of the definition. Again, I think it is an important prerequisite for obtaining skills and knowledge in statistics, but the concept does not cover all I want to show in this study.

Statistical reasoning is considered important, as Garfield & Gal (1999) define it as *'the way people reason with statistical ideas and make sense of statistical information'*. This last concept best describes what I think is important in statistics education, because it covers the way in which I want students to work with statistical information (see also Tempelaar, Gijselaers & Schim van der Loeff, 2006). In a broader sense even, teachers want students to develop 'statistics competency', meaning *'the ability to critically process statistical information and appropriately making use of that information.'*. This closely resembles the 'complex skills' introduced in section 1.1 of this thesis (Hoogveld, 2003).

lysis. This is primarily done in a theoretical rather than an applied manner (Vlaams Ministerie van Onderwijs & Vorming, 2008).

Instructors should encourage students to learn statistical concepts and to critically review results, interpret those results and link technical outcomes to real-life research questions. By making students enthusiastic and involving them in real-life projects, chances are that students will retain these skills and knowledge for a much longer time than when having been taught in a more traditional, unidirectional setting.

It is therefore necessary to analyze attitudes toward statistics and their effect on student outcomes and possibly recommend on more successful teaching methods. So far, literature has shown (Garfield & Gal, 1999) that helping students to verbally discuss statistical results challenges students to participate in research projects and statistical analysis, and as a result improve student achievement.

This study examines how statistics' attitudes can be modeled for Dutch and Flemish universities and, if necessary, give directions for improvements. Firstly, in chapter 2, I will introduce the Expectancy Value Model that clarifies the link between institutional and individual factors and attitudes its the effect on student outcomes.

1.5 Focus of this study

This study focuses on determinants of student outcomes with respect to introductory courses in Methods & Statistics. For this study they can be divided into two groups of factors. Educational factors (also referred to as institutional or school factors), include school factors, teaching methods, class size, teacher skills, existing rules and regulations (see Schau, 2000). These indicators are derived from the school system. Student or individual factors include previous school careers, background characteristics, statistics literacy, prior knowledge to learning statistics (Tempelaar, Van der Loeff & Gijselaers, 2002), mathematical skills, study habits, attitudes toward statistics and self-confidence.

Schau (2000) presented an additional model, where educational factors mediate the effect from learner and institutional characteristics on student (course) outcomes. Schau (2000) decomposed educational factors into school and course components. A broad outline of the implications of both groups of factors for this study, will be described in the methods part of this thesis.

Lastly, getting acquainted with 'Methods & Statistics' is considered to be a complex process of knowledge construction, where students process information actively and construct knowledge through experience (Verhoeven, Brand-Gruwel & Joosten-ten Brinke, 2004; Simons, Van de Linden & Duffy, 2000). This study focuses on the best way to teach Methods & Statistics at University level in The Netherlands and Flanders.

1.5.1 Research Purpose

This study has a twofold purpose. Firstly, it aims to analyze attitudes toward statistics and any changes in the course of the semester. The effect of several in-

stitutional and individual determinants on the learning process will be analyzed. It will present a model for determining the effect of individual, institutional factors and attitudes on student outcomes. These effects will be compared across colleges and universities throughout the Netherlands and Flanders. Finally, with the outcome of this study I aim to provide a number of didactical recommendations for teaching Methods & Statistics at Universities and Colleges, with a focus on making Methods & Statistics memorable. The approach taken in this study is empirical-theoretical Sociology.

Secondly, this study aims to make a contribution to the development of statistical tools to analyse any changes in attitudes and provide tools to determine the influence of individual and institutional covariates on these changes in attitudes and, indirectly, on student outcomes.

With the results of this study I hope to contribute to the ongoing development and innovation of course material for Methods & Statistics. After all, the aim of every university and college teacher is (or should I say 'should be') - apart from getting across statistical knowledge and skills - to inspire his / her students and to make them enthusiastic for the subjects taught. Furthermore, the knowledge, applications and examples taught during an Introductory course in Methods & Statistics should make statistics memorable (Sowey, 1995).

Chapter 2

Determining Course Outcomes in Statistics Education - Theoretical approach

2.1 Introduction

This chapter will present a model of the determinants of course outcomes for Introductory courses in Methods & Statistics. The theory in this chapter will be described by means of the funneling principle, i.e. from 'general' to 'specific'. I will start with a social scientific view on education and move via social psychology toward a model that includes both institutional and individual factors in an attempt to predict student behavior and course outcomes, especially student achievement.

2.2 Teaching Methods & Statistics - A perspective from Social Science

First an overview will be given of sociological ideas on education and, more especially, on teaching Methods & Statistics.

Effective schooling - a debate on the use of institutional factors
Until recently, the discussion of 'effective schooling' has primarily focused on K-12 systems of education. Coleman (1961) developed an almost Durkheimian model for School Effects, because he sees social systems as fundamental determinants of individual social action. Coleman is the main representative of the rational choice theory. He describes the movement from the individual to a

collective level. According to Coleman, a school's academic program has the greatest influence on effective learning (Sorensen, 2000, 148).

In his Theory of School effects (in: Sorensen, 2000, 141) Coleman focuses on social systems of learning, such as peer groups, families etc. as a condition for individual growth and achievement. According to Coleman, institutional factors (such as 'school of choice') explain very little variance in 'course outcomes', but family background of students does (Sorensen, 1996, 207; 2000). This is the main conclusion of the first Coleman Report (1966). In his later work, Coleman contradicts the previous findings by stating that 'the choice of school does matter'. He looks back at his original work claiming that the social systems associated with schools matter a great deal when it comes to norms and values of their students, along with their work attitudes.

What causes this contradiction? According to Sorensen (1996) the choice of variables plays an important role here, since Coleman included economic predictors in his first study and they did not contribute to the explanation of 'study achievement'. However, other variables such as 'private vs. public' and 'discipline' could play an important role. Another reason for these conflicting conclusions is the choice of methodology, as Coleman solely focused on explained variance and others (after him) focused more on the existence of an effect.

Murphy (1985) concluded that schools make a significant difference and these findings also seem to contradict Coleman's conclusions. However, as Murphy puts it, the British (Rutter et al., 1979) and the American approach actually do not contradict each other at all, they simply make different claims. As Murphy states, Coleman and Rutter each analyzed different predictors for the effect on academic achievement.

Although they made an important contribution to the discussion of institutional factors, these studies focus on K-12 systems (primary and secondary schools). In teaching a topic like Statistics to adolescents of 18 years of age, the influence of institutional factors cannot be ignored, but increasingly the influence of individual factors must be taken into account.

Institutional factors in this study

To what extent can the outcome of this debate be used to the benefit of this study? Although Coleman's study concentrated on primary and secondary school systems, this study can build on Coleman's claim that schools do make a difference. As will be announced in chapter 3, schools in this study will be looked upon as 'separate systems' each with their own organization, educational method, policies, norms & values. As a result of these offerings, student populations across universities are quite different.

Actually in the nineteen sixties and seventies, when this 'school effect' debate was taking place in the Anglo-Saxon academic communities, choice of schools was a matter of background and finances of parents, i.e. individual background factors. What Coleman found was little effect from 'school resources, funding and housing' on achievement. In that way, Rutter even had to agree with Coleman, since he had analyzed these factors as well.

Coleman and Rutter would agree about one thing: the individual intellec-

tual abilities of the student as well as the personal background form important factors in predicting school achievement. When looking at additional institutional factors, the educational profiles of schools do make a difference, along with the occupational prospects for students. Besides analyzing the effect of institutional factors on course outcomes, in this study I want to generalize the notion of 'school differences' to 'differences across school systems'.

In the next section, a model for the combined effect of individual and institutional determinants on student behavior, and therefore on course outcomes will be presented.

2.3 Expectancy-Value Theory and Attitudes towards Introductory Statistics

Introduction

Two main types of determinants of course outcomes in statistics education will be looked at: institutional and individual determinants. First, a model will be presented that predicts (student)behavior toward statistics and achievement as a result of characteristics of the institutions, of the students themselves and their expectations, of course components and attitudes. After describing the general Expectancy Value model, an application for Statistics Education and for this study will be presented. Then, assumptions about specific institutional and individual factors and achievement will be made.

How does this model work?

The Expectancy Value Model is a model for explaining achievement related choices (Wigfield, Tonks & Eccles, 2004). The development of Expectancy Value Models started in the nineteen thirties by Lewin and Tolman, the general model was developed by Atkinson and Feather (Wigfield et al., 2004). This theory has many applications in predicting achievement-like behavior, among which learning behavior.

The general model is depicted in fig. 2.1. According to the basic theory, achievement behavior can be looked upon as a function of the expectancies a student has, the goals toward which he/she is working and the task value of the student. When the student has more than one choice, he or she will choose the option with the best possible combination of expected success and value. In other words, you can look at this model as a form of Rational Choice (Schunk, Pintrich & Meece, 2008, 64). In turn, the students' motivations and beliefs are influenced by cognitive processes, such as experiences with past events and perceptions of expectations of others. Lastly, these processes are linked to previous events, upbringing and cultural stereotypes.

Next, this Expectancy Value theory will be applied to Attitudes in Statistics Education. This application is based on the notions developed by Wigfield & Eccles (2002, 2000; see also Wigfield, Tonks & Eccles, 2004); I will explain how the model can be linked to 'attitudes toward statistics'.

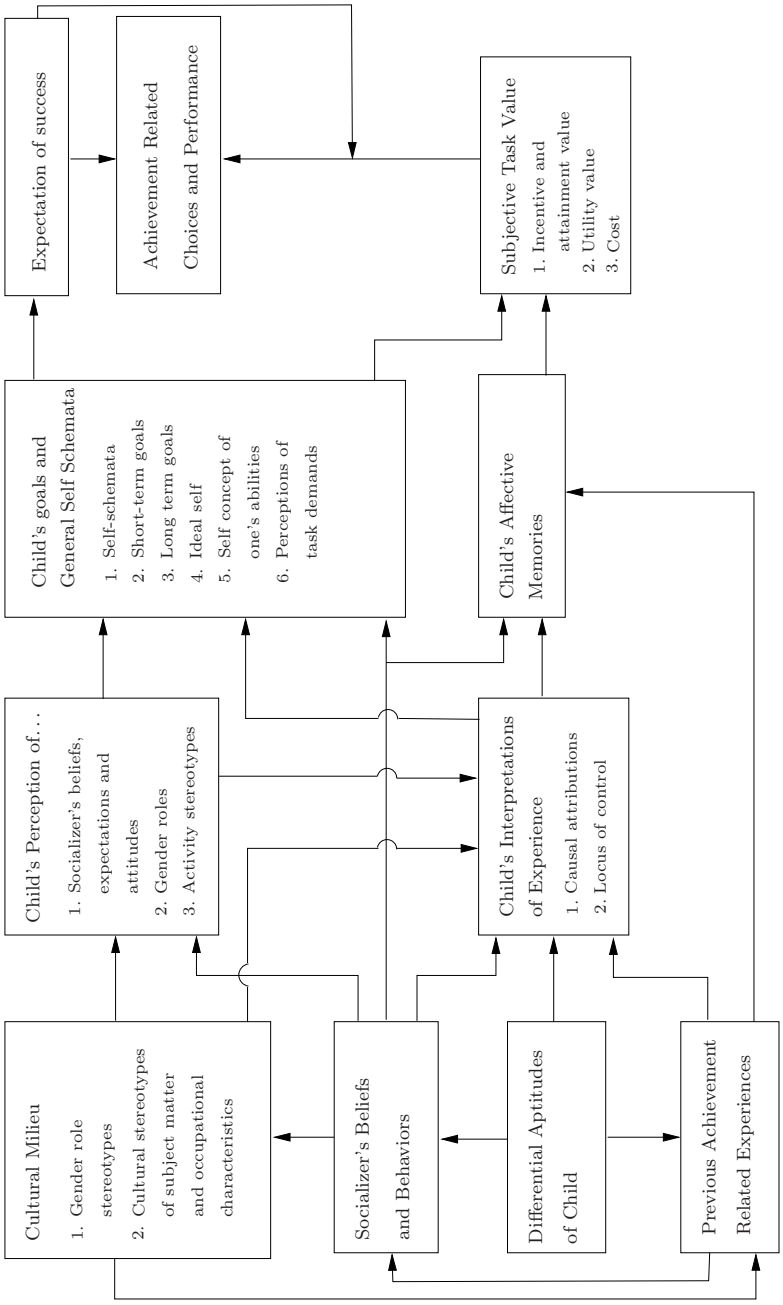

Figure 2.1: General model Expectancy Value Theory

2.3.1 Applying the Expectancy-Value Model to Statistics Attitudes

Student achievement in statistics is assumed to be linked to the perception of being successful in this course. In other words, if students perceive their ability to be successful in statistics positively, they will have higher expectations and values toward their achievement and they will be better motivated to work hard. What I need to know is the students' perception of their ability to do statistics (cognitive competence), their perception of difficulty of the statistics tasks and their feelings towards the course (more positive experiences result in more positive feelings about the course). Furthermore, I need to know what task value students attach to the task before them.

The model predicts expected achievement (student outcomes as part of course outcomes) using four types of attitudes toward a task / course, such as affective feelings, perception of difficulty and cognitive competency and task value. They will be elaborated below. These attitudes derive from the motive for success and the motive to avoid failure, and from the evaluation of the success rate in a given situation. As the motive to be successful derives from experiences in achieving success in a previous learning situation (such as high school mathematics), the motive to avoid failure derives from similar learning situations where the student was unsuccessful. Experiences of success result in high achievement motivation, and lack of success result in motives to avoid failure (Motivation, 2006).

In turn, these motives are shaped by (the interpretation of) previous experiences and the expectancies of the social surroundings (peer pressure and / or expectation, socialization et cetera). Lastly, the larger cultural milieu plays an important role in affecting those beliefs and motives.

Schau's application

Schau (1992) applied the Expectancy Value Model to attitudes towards Statistics (Schunk, Pintrich & Meece, 2008; Sorge & Schau, 2002; Wigfield & Eccles, 2000). The model predicts the influence of expectancy and value factors on student achievement. Schau (1992) proposes a model for predicting student outcomes using institutional -, learner - and course factors (Schau, 2000; Garfield et al., 2003), as is shown in fig. 2.2. Although student learning is considered the main predictor for achievement, a positive attitude towards statistics is also considered to be a major determinant of course outcomes (Hilton, Schau & Olsen, 2004, 92-93).

The model shows both direct and indirect effects. Course factors such as didactic methods, goals, assessments and the teacher have a direct effect on course outcomes. Institutional and learner characteristics, however, produce both an indirect and a direct effect on course outcomes, more especially on student outcomes. According to Schau (2003), instructors and programs need to take into account institutional and learner characteristics when course components - and, what is more, student learning - are developed. This would be in line with the constructivist approach that has been discussed in chapter 1 (for examples,

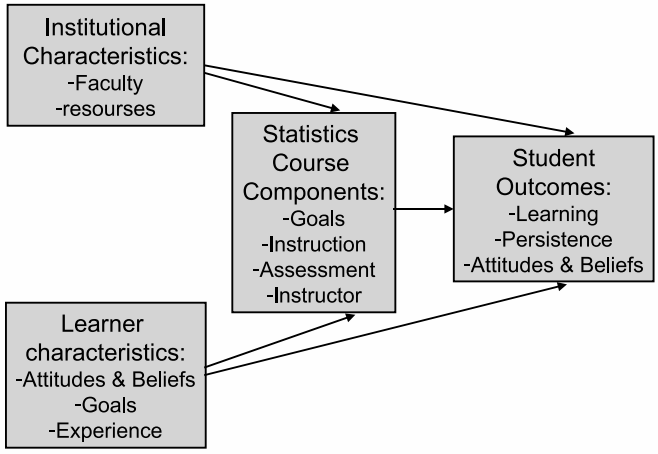

Figure 2.2: Model by Schau (2003)

see Garfield, Hogg, Schau & Whittinghill, 2002). However, Schau assumes that learner characteristics still rarely play an important role when the desired outcomes are formulated by the educationalists, as the primary focus in most cases still is teacher rather than student oriented, exemplified by a question like: *What do I want to teach them*, instead of *What do they need to know?*. Furthermore, it is important how the courses are shaped, how the teacher communicates with students and how teacher-learner relations are manifested. I assume that all these factors play a role as far as student outcomes are concerned.

Development of attitudes towards Statistics
For the development of the Survey of Attitudes Toward Statistics, Schau & Stevens (1995) formed a panel of experts (both students and instructors). All members had expertise in enrolling in or teaching introductory courses in (Methods &) Statistics. During a mind mapping session they came up with a number of words and concepts that could best describe 'attitudes toward statistics'. They also reviewed previous measurement instruments for proper descriptions of attitudes toward statistics, as follows:

1. *Difficulty* of the tasks, i.e. perceived difficulty for a particular student
2. *Cognitive competence*, i.e. students' perceptions whether they could master the necessary knowledge and skills
3. *Affect*, i.e. the students' positive and negative feelings about the course
4. *Value*, students' individual motives and beliefs about the importance of

fulfilling a task (the usefulness for one's future job, intrinsic interest in the task and whether the cost balances the effort). This component, also known as 'Task Value' was added for three reasons:

- student's affective feelings toward statistics may not be the same as their attitudes about the value of statistics.
- this component has often been added as a result of 'statistics anxiety'.
- it is believed that students' affect toward statistics is important in it's own right (Sorge & Schau, 2002).

Besides the aforementioned attitudes Schau added two more components at a later stage of the development of this model (Schau, 2005):

- *Effort*, the Effort a student plans to put in, in order to achieve a good grade.
- *Interest*, the students' level of individual interest in statistics (Schau, 2003).

2.3.2 Application by Prosser & Trigwell (1999)

Prosser & Trigwell (1999; see also Prosser, Trigwell, Hazel & Callagher, 1994) developed an earlier application of the EV Model, depicted in fig. 2.3.

They added students' perceptions of the context and students' learning approaches separately, whereas in Schau's model they are represented by 'Effort'. Students' perceptions of the context mediate the effect of student approaches on learning outcomes, according to Prosser & Trigwell (1999). Educational factors both directly and indirectly influence course outcomes, whereas student characteristics only play an indirect role in the model.

Course - or learning outcomes

Schau (1992) defines 'course outcomes' as the outcome of learning, persistence, attitudes and beliefs. Prosser & Trigwell define 'learning outcomes' as the quantity and/or quality of what students learn (see also Garfield et al., 2002). For my study, however, 'course outcomes' are defined as 'student outcomes', or as the individual course result and they are operationalized as the final grade (see also section 2.5). This measure has been chosen for two reasons. Firstly it forms an objective and independent measure of students' achievement at the end of a course (Shachar & Neumann, 2003), irrespective of students' opinion and perception. Secondly, despite the fact that 'final grade' might be subject to variation due to differences in teaching quality and other institutional variables, it can be compared across institutions, as the grading systems in most cases have a 10-point scale (or a scale that can be linked to this).

However, as will also be argued in section 5.3.3, the relative importance of obtaining a higher grade might differ across institutions, causing the dependent variable to be flawed. Therefore, a more subjective additional measure of

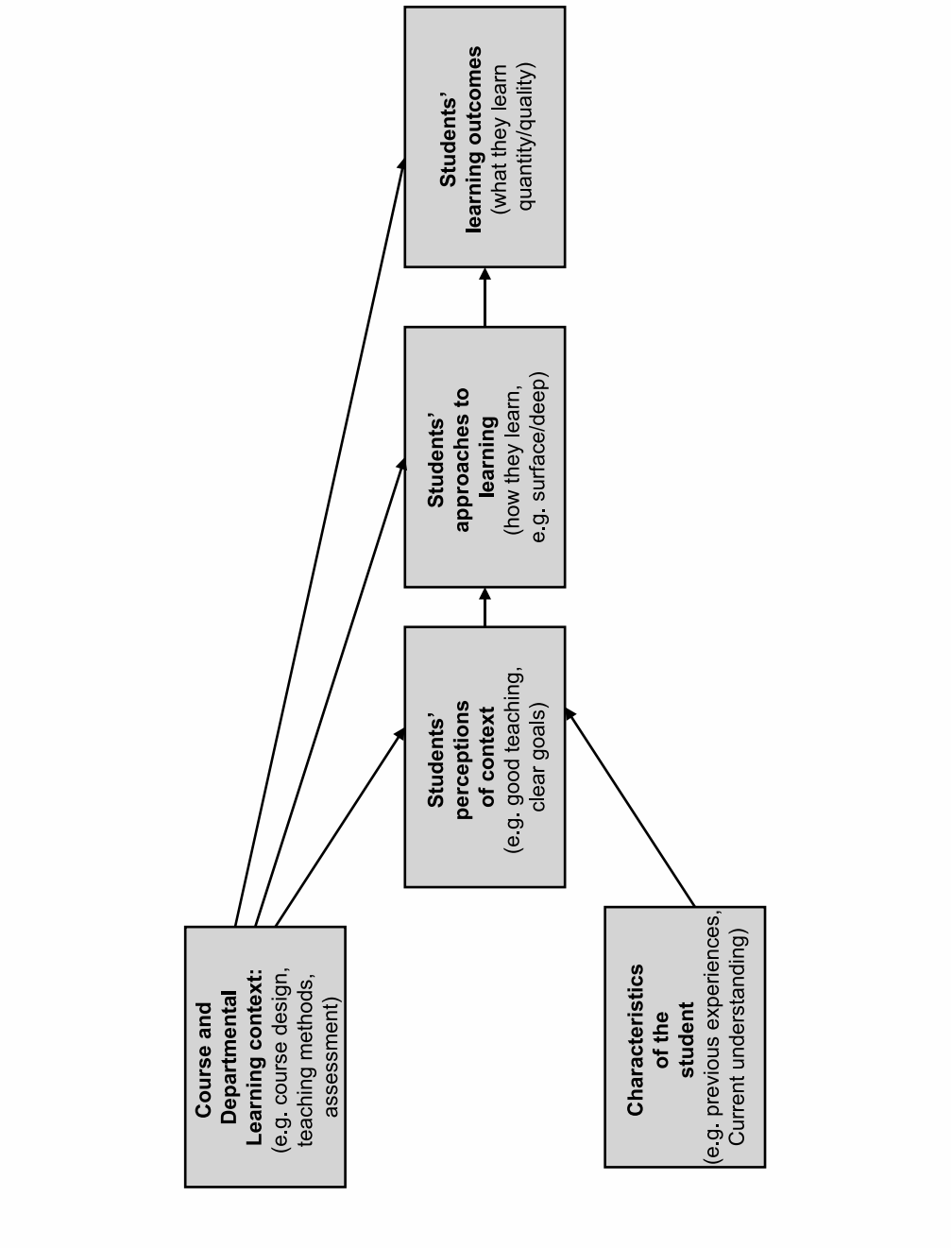

Figure 2.3: Expectancy value model by Prosser & Trigwell (1999)

student outcomes is introduced by means of the students' expectancies of their own achievement (Pascarella & Terenzini, 1991). Pace & Pike (Pace, 1984; Pike, 1995, 1996) found growing evidence of the reliability of this additional measure of 'self-report' results. Expected grade is a reflection of the combination of opinions and perceptions of the individual and institutional factors. For instance, it is expected that a student who lacks self confidence would expect a lower grade than a student who is very confident.

2.3.3 The application for this Study

For the model in this study elements from both Prosser & Trigwell and from the Schau model are used. The model is depicted in fig 2.4. In this model it is assumed that attitudes toward statistics reflect upon a certain learning motivation. In that sense attitudes influence student outcomes. Individual characteristics (such as previous math experience, self confidence and background) and educational characteristics (didactical and assessment methods and organisation of the course) in turn influence these attitudes and, indirectly, outcomes (see section 2.4).

Attitudes toward statistics obviously already exist before any course starts. As a result of taking this course, this attitude could change. In fig. 2.4, the attitude component is shown as one indicator. It should be considered an indicator of 'change'. Therefore, I have chosen to measure these attitudes twice, once at the start of the course and once at the end of the course, and then to analyze the attitude change. A more detailed description of this method will be given in chapters 5 and 8.

The model acts as a simplified version of the Prosser & Trigwell- model, because I will not test separately for 'learning approaches', but measure effort as an indication of learning approach.

Additionally to Schau's model, attitudes not only act as an outcome variable, but part of it could also mediate the effect of institutional (educational) and individual factors on 'final grade'. Next, I will explain that I expect this to be especially true for 'Effort' as this construct has a special place in the model.

The special position of 'Effort' in the model

'Effort' is defined as 'the amount of work a student plans to expend (or has expended) to learn statistics' (Schau, 2005). Firstly, Effort is considered the 'work attitude' that a student has at the *start* of the course: does he or she plan to work hard in order to pass the course? Secondly, Effort in hindsight is the subjective amount of work the student perceives he/she actually spent (at the end of the course). This perception is different from merely asking the 'number of hours' a student put in (or observing 'time on task'), because Effort as an attitude consists of a combination of motivation, interest, energy put in, time spent et cetera.

Effort takes a special position in the 6 attitude components that Schau mentioned in her model (2003). First of all Effort can be the result of certain attitudes, for instance feeling more competent or perceiving statistics to be more

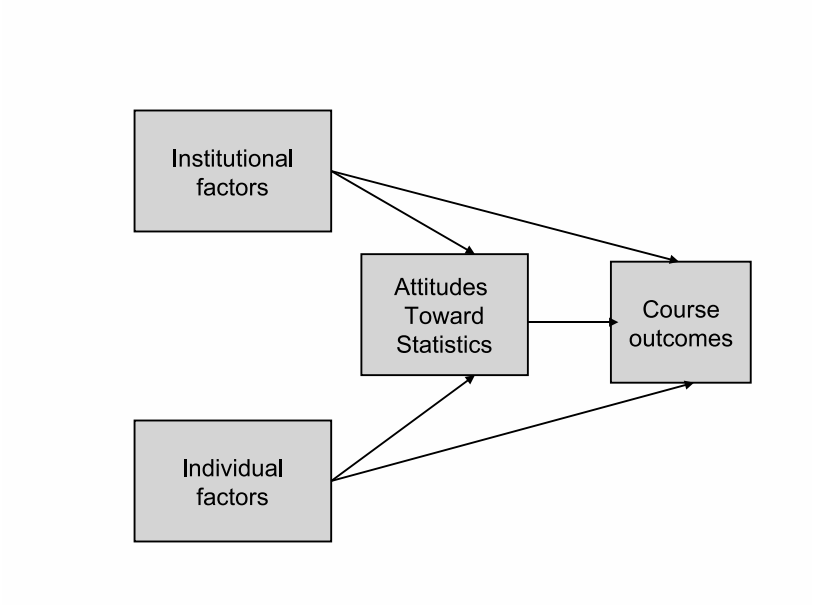

Figure 2.4: Simplified Expectancy Value Model for this study

interesting could motivate a student to work harder and therefore get a higher grade. This interest in the topic causes the student to take a deep approach to learning (Prosser & Trigwell, 1999; Tempelaar, 2007a). Secondly, Effort can operate on the level of the other attitudes toward statistics, and in relation to these other attitudes affect course outcomes. Lastly, Effort could exhibit two separate processes:

- a **surface learning approach**. In this case, students only focus on passing the course, on memorizing the most important concepts to do so. They will not retain that information very long and in general it is assumed that they are not interested in the topic.
- a **deep learning approach**. A 'deep' learning approach is a learning style where the student takes a critical look at the material, linking it to already obtained knowledge (Biggs, 2003). Students who are really interested in a certain topic, such as Statistics, are willing to invest time and effort in order to really focus on the meaning and application of statistical concepts (see also Tempelaar, 2007a; Prosser & Trigwell, 1999).

Effort can act in different directions. Students who are not interested in the topic (and are considered not so competent in the subject matter), will only put in so much energy and time as to just pass the course, based on extrinsic motivation. On the other hand students who are interested in the topic (and probably more competent in the subject matter) will put in more Effort, because of intrinsic motivation. Effort could be an indicator, and/or a mediator.

Effort could have an influence on course outcomes, as more Effort could lead to a better grade. Effort could also be a result of motives and beliefs (e.g. positive Affect) and therefore act as a mediating factor between other attitudes and course outcomes. As it is currently not clear what the position of Effort in the SATS©-model should be, I decided to first use Effort as it was originally intended, as one of the attitudes in the prediction of 'course outcomes'. Fig. 2.5 shows this. Additionally Effort might come out as a mediating factor, creating an extra layer in the model.

Effort and 'time on task'

Literature has shown that the more time a student spends on a given task, the more he will learn (Brophy, 1988), provided the instructor is skilled enough and the course is organized effectively. Perceived Effort however is expected to give a better picture of student learning than 'time on task' alone, especially taking the aforementioned two approaches to learning into account. After all, putting in more Effort could indicate two things: either the student takes a deep interest in the topic or he struggles to pass. Therefore, I consider 'Effort' to be a more appropriate measure of students' learning approach relative to student achievement than mere 'time on task'. However, in order to compare 'number of hours put in' to 'Effort', in this study, additional measurement of the 'number of hours studied' will be added. This is done by asking students to indicate in hindsight how many hours they put in.

Assumptions regarding 'Effort'

If a deep learning approach is taken by the students, then there is a positive relation between cognitive competency and effort notwithstanding the influence of other factors. Hence, the more competent a student perceives himself to be, the more effort he will put in. The degree of difficulty, interest or value will not make a difference.

If a surface learning approach is taken by the students, then a negative relation between cognitive competency and effort is expected, meaning that the more competent a student believes he is, the less effort he will put in to pass the course. The degree of interest, value and difficulty a student reports in this 'surface learning approach' will clarify this correlation.

Teaching quality

Student learning may be influenced by the quality of the teacher and the program offered. However, teaching quality has not been added to the model in this study. Ethical arguments prevented some institutions to give permission to use these evaluations. This is a pity, because the lack of a uniform cross-institutional evaluation system complicates the comparison of teaching quality across institutions. However teaching and teacher quality do affect study behavior and therefore student achievement, as was found by Den Brok, Brekelmans & Wubbels (2004). Den Brok et al. (2004) showed that interpersonal teacher behavior and course outcomes are related. Analyses within universities could produce relevant data, but the differences across systems and the only partial

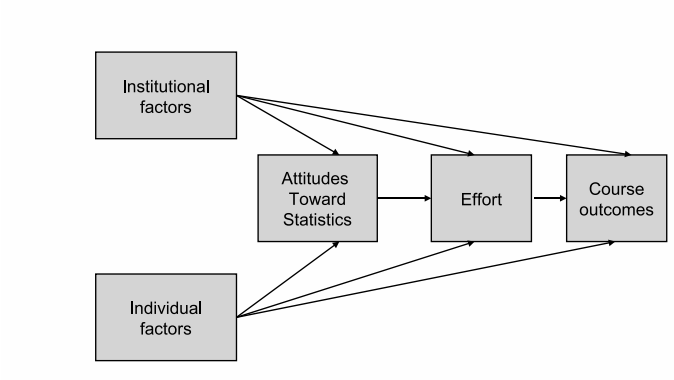

Figure 2.5: Expectancy model for this study, with Effort as mediating factor

permission to use evaluation results prevented me from performing a multilevel analysis that would provide at least a partial contribution to this issue. There is a distant possibility that the unexplained part of my results may point to teacher behavior. For the present this study is limited to analyzing student attitudes toward statistics.

2.4 Predicting outcomes in Statistics Education

As institutional (educational) predictors, didactical and assessment methods will be considered, as well as course organisation factors (duration, reward system and size). Individual factors that play a role are social background, previous experience, study habits and self confidence. Both groups of indicators will be discussed in this section.

2.4.1 Institutional factors

School & class size

Research into the effect of school size on student achievement shows different results. A few studies show a negative relationship between school size and student achievement. This means that smaller school size causes higher student achievement (Howley & Bickel, 1999; Lee & Smith, 1997). However, this relationship is confounded by the effect of background characteristics, such as socio-economic status. As Howley & Bickel state (1999), effective school size may well vary across communities, as the circumstances in some communities may lead to larger effective school size, whereas in other communities small size schools would be better.

Others state that in small sized schools, more teacher collaboration takes place (team teaching) and the teachers feel that they have a more direct influence on their work (Palardy & Rumberger, 2002). Other factors relating to small school size are (Howley & Bickel, 1999; Lee & Smith, 1997):

- more frequent interaction between and among teachers and students in smaller schools
- more intense interaction between and among teachers and students in smaller schools
- a greater sense of community in smaller schools compared to larger schools
- higher motivation and satisfaction in smaller schools
- dependency on the level at which teaching takes place: primary, secondary and tertiary school.

However, these results do not provide a solid assumption for this study. First of all most studies concentrate mainly on schools of primary and secondary education; second, they sometimes refer to 'school size' instead of 'class size'. School size may be small, but class size may still be large. Lastly, most results mentioned come from the United States, and the generalizability to the Dutch and Flemish situation is questionable.

Class size

Research shows that class size does not conclusively affect student achievement (Gilbert, 1995; Martins & Walker, 2006). It is what goes on in the classroom that matters, teacher's and teaching quality, frequency of interaction and course organization. No matter what size a group is, if the teaching is great, the outcomes will be positive.

Previous research, however, also shows evidence of the contrary. An experiment by Bressoux, Kramarz & Prost (n.d.) showed a significant improvement on student achievement when class sizes were reduced. Other experiments (Dillon & Kokkelenberg, 2002) show contradicting influences of class sizes on student achievement, indicating that other factors might play a role. Hattie's overview of meta-analyses (2005) shows that if reduction of class size influences student achievement, the reduction should be to at least 15 students. Even if class size plays a role in predicting student achievement, other factors could play an important role, such as teachers' qualities, the frequency and nature of interaction, disruptive students. *'Class size reduction can lead to improvements, provided certain conditions are met.'*(Hattie, 2005). Most meta-analyses however concentrate on pupils at primary and secondary schools and the results cannot simply be extrapolated to college and university students. This study aims at analyzing the possible effect of class size on attitudes toward statistics and, thereafter on learning outcomes for instance final grades.

Course Duration

Presumably it should make a difference whether a student has 8 weeks, 15 weeks or even 26 weeks to study a number of introductory statistics topics. Let us assume that every student in a mandatory statistics course has to study approximately the same number of topics in order to obtain a certain level of statistical knowledge. Almost every course starts with methodological principles

such as setting up a research design, operationalization, validity- and reliability issues, after which a number of statistical topics have to be mastered. There is hardly any curriculum in the Netherlands and Flanders that has different topics on their lists. So, having to study those topics in a time span of 8 weeks or 26 weeks should make a difference.

Budé (2007) analyzed the level of conceptual understanding (and for that matter, final exam grade) for several groups of students, among which students who took a long course (with so-called 'distributed practice') and some students took a short course (with so-called 'massed practice'). The latter means that students had to concentrate all meetings and study in a short period of time. Budé found that conceptual understanding and thus exam grade was significantly lower for students who took the short course compared to students who took the long course (2007, 72). It is assumed that because students have more opportunity to practice statistical skills, longer lasting courses yield better student achievement than shorter courses. In this study I will analyze whether this assumption holds true.

Teaching methods

Presumably, teaching methods influence student achievement. A more direct and interactive way of teaching statistics motivates students to work actively towards constructing their own knowledge (Moore, 1992; see also Den Brok et al., 2004). The more a teacher interacts with students, the more motivated students (can) become.

Lecturing is a uni-directional teaching method that is often negatively evaluated both by students and teachers. However, due to cutbacks in university budgets, lectures often provide a solution to teaching a certain topic to a large group of students. For this reason many universities still use this teaching method for teaching a large course such as Introductory Statistics. Besides lecturing, many universities offer the option to get some extra explanation or do extra exercises in small groups. If this method is applied parallel to lectures, it would partially make up for the loss of motivation during the lectures, since a student can still get extra coaching during the small group sessions.

The teaching method that is considered the most interactive is the student project. In statistics this usually means that students set up and perform a small research project and discuss this with a coach, usually the teacher, in regular meetings. Not only do students feel they can learn more from one-to-one coaching, it is also evaluated positively because students can do real life research projects, and experience what it is to set up their own data collection and analysis. It is assumed that students learn more from performing their own research projects (in small research groups) than just from unidirectional lectures (Moore, 1992), hence more interactive teaching methods are expected to have a positive effect on student achievement.

Assessment methods

Although there is a range of assessment methods in statistics education, the one most used is still the multiple choice exam, closely followed by the exam with

open questions. Assessing students' statistics competencies by means of only one test is not considered good enough (Garfield, 1994), because many different skills are developed during the statistics course. Many courses have additional assessment methods such as papers, presentations, attendance, active participation and projects. As the most interactive teaching methods are assumed to have a positive influence on course outcomes, assessment methods related to those interactive methods are assumed to also have a positive effect. Besides papers, presentations and projects, continuous assessment methods could play a role, such as keeping the students informed on their progress and about the learning goals the student may expect to have obtained by the end of the course (Biggs, 1993; Garfield, 1994).

ECTS

Finally, the number of ECTS that is at stake is expected to have little effect on achievement. ECTS is part of the European regulations for higher education (Realizing the European Higher Education Area, 2004), and it is not really expected to influence student motivation, as ECTS is a fixed reward system. However, I do believe that taking part in a course with more ECTS motivates students to study, as there is more 'to gain'.

Differences between institutions

Every institution offers a different combination of teaching methods, assessment, frequency of lectures and work groups, course duration, ECTS, and class/group size. They are 'different systems' in every respect.

Moreover, I expect to find a 'selection effect', i.e. different student populations reflected in each institution. Students often base their choice of university on social factors, such as the city where the university resides, the student life and the availability of housing, but also on different offerings that universities have (Astin, 2003). Furthermore, universities receive different student populations based on certain paradigms that are established in universities, or the research possibilities. Where some universities offer problem based learning, others offer interactive projects, or a certain methodological direction, some are catholic or protestant by origin.

Assumptions about institutional factors

The fact that the social environment is perceived to be an important predictor in course outcomes leads to the assumptions that class size, assessment method, didactical approach, course duration and other specifics play a role in predicting course outcomes.

2.4.2 Individual factors

Not only does socialization have an effect on how people interpret experiences, it also shapes the way people deal with cultural (peer) pressure, how self confident they are, and what motives and beliefs they have toward achievement behavior. A number of individual factors are assumed to have an influence on attitudes

and, in turn, on student achievement.

Background characteristics
Individual factors such as gender, nationality, language, major and nationality are expected to influence attitudes. Age is expected to influence attitudes toward statistics, because as one gets older, the outlook on priorities changes, and students are less bothered with negative feelings towards courses. Therefore I expect a more positive attitude and more positive changes as students are older.

Math skills
Starting level with respect to Methods & Statistics, also known as math skills can be divided into:

- the extent to which students had Mathematics at Secondary School = *mathematics experience*
- mathematics grades at Secondary school and *perception of ones own competencies at Secondary school*
- placement test Methods & Statistics. In most cases the placement test will contain a number of algebraic questions testing the extent to which a student is capable of solving quantitative questions. As this way of testing is not customary in the Netherlands and Flanders, it is not taken into account in this study.

Study habits
Additional to Effort, study-related behavior can be seen as efficiency in use of (study) time, willingness to work for the course, readiness to think about a problem instead of giving up. 'Number of hours studied' is defined as the number of hours studied outside class hours. This of course can only be measured after the course has ended.

Test anxiety
This can be described as the lack of confidence in one's own abilities to do a test. The result often is a so-called 'black out' and students fail the test because of it. A special case of this type of anxiety is described as Statistics anxiety. This is the lack of confidence in one's own statistical abilities.

Research by Musch & Broder (1999) shows that both Math skills and Test anxiety can separately explain student achievement with respect to Introductory Methods & Statistics. According to Musch and Broder (1999, 114), academic performance can be affected by test anxiety directly, by lack of knowledge at the time when it is needed, and indirectly by distraction at the time of the actual test. Math skills seem to be the best predictor of statistics outcomes, followed by test anxiety. However, since the constructs were only measured at one moment in time, there is no clear evidence for the unique contribution to the explanation of student outcomes by math skills, because test anxiety might have influenced math skills in an earlier stage of the respondents' educational

career and this potential intervening influence was not measured.

Global attitudes
These attitudes regard the students' general beliefs of themselves as learners (of statistics), beliefs about future use (for statistics) and self-esteem. These general self-perceptions are assumed to play a role in most 'competency- and achievement driven' tasks, not only with students but more in general in jobs, family and personal life (Schau, 2003, 2005; Schau et al., 1992, 1995, 1999). In this study they have been applied to 'Statistics Education':

- how does a student perceive his/her mathematics competencies
- does one expect to use statistics in future jobs
- does the student have any math and / or stats experience? As many students, especially in the Netherlands, do not have much experience in that discipline, experience in Mathematics is also considered a good predictor of student outcomes. Math experience can be divided into the length of this experience and the self report on how well the student did during those math classes. Although the same goes for Statistics experience, not many students can reliably report about stats experience, because the introductory course for this study is the first course they are taking. Students from different national backgrounds usually show differences in the amount of statistics that was incorporated in the high school curriculum. Therefore I expect a difference in outcomes with regard to nationality.
- self confidence. It is assumed that self confidence has an effect on the motivation to learn and/or study. Self confidence is looked upon as the self-perception of ability or competence. Schunk, Pintrich & Meece (Schunk et al., 2008) consider them as part of motivational beliefs, the beliefs that lead to actual achievement, involvement in the task and effort and persistence (Schunk et al., 2008).

  Linked to '*self confidence*' are measurements concerning *expected outcomes*. Expected outcomes are considered to be subjective measurements of course outcomes (Shachar & Neumann, 2003). Students' self confidence would in general result in higher expected outcomes. Additionally, higher expectations could be affected by positive attitudes and therefore affect student outcomes, as is shown in fig. 2.6. This model represents the final stage of the application of the Expectancy Value Theory for this study.

Assumptions about individual factors
In sum, the two main assumptions derived from the Expectancy Value Theory are:

1. the more positive students are about their own competencies and their expectancies about the outcome, the better their result (i.e. course outcome) will be.

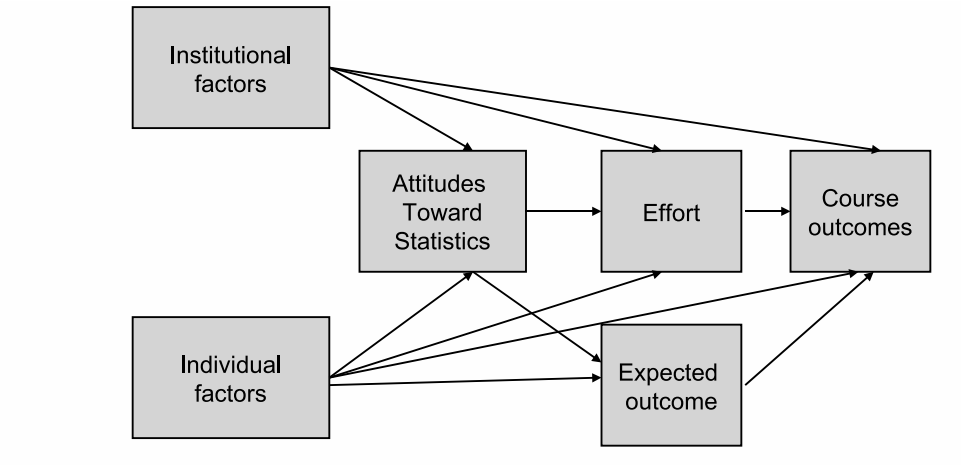

Figure 2.6: Expectancy model for this study, with Expected outcome (of students) as mediating factor

2. the more students value statistics and are interested in the topics, the more likely it is that they will use it in their future jobs (Schunk, Pintrich& Meece, 2008, 66). This will then lead to more positive attitudes and, in turn, to more positive student outcomes.

2.4.3 Gender differences in statistics achievement

Although sex differences have declined in recent years (Eccles & Wigfield, 2002), according to Greene and DeBacker (2004) male and female students still differ in the way they are motivated to take science courses. It is a well known suggestion that males usually perform better at mathematics than females and this could therefore boost their self-confidence and, hence their motivation more than that of females. At the very least, males elect mathematical topics more often than females and they engage in higher level extracurricular mathematical activities (Harris & Schau, 1999). In this study I want to test whether the gender-difference assumption holds true for Introductory Statistics courses and apply this assumption to attitudes toward statistics. According to the Expectancy Value Theory (Wigfield & Eccles, 2002; 2000) students' expectancies of success depend on past achievement outcomes and the interpretation of those outcomes (Did the students in their own perception do well in mathematics during high school). These experiences affect students' perceptions of the future, whether they think they can master the subject, how difficult they think the subject will be and what value they think the subject will hold for their future careers. In their turn, those future expectations shape the students' effort to perform a task, especially studying to pass the Statistics course (Greene & DeBacker, 2004). Higher motivation, then, is related to higher achievement.

In answer to the question whether males and females differ in statistics attitudes, and therefore in their achievement, different research results are reported. Many studies report that females have a lower outcome expectancy when it comes to mathematics, than males (Greene & DeBacker, 2004). This could explain a lower student outcome for females than for males. Males in most cases have a higher self-efficacy than females. It was also reported that women experience higher anxiety levels in statistics classes (Baloglu, 2003) than men. This indicates that it is not only the perceived ability to perform well in statistics courses that affects course outcomes, but also the anxiety for statistics. It might be added that anxiety is usually shaped by ignorance about the real difficulty, context, contents and value of the statistics course. Anxiety is usually shaped by means of rumors, negative expectations and horror stories. Cherian and Glencross showed (1997) that there is no difference in attitude toward statistics between male and female students. I want to test whether - in the Netherlands and Flanders - there is a gender difference, and to what extent.

Harris & Schau (1999) come to a different conclusion. Their focus is on student outcomes. After reviewing a large number of articles on the subject they conclude that males and females do not differ that much when it comes to student outcomes regarding statistics, especially in college education. Moreover, males' scores vary much more than females' scores. They argue that differences across gender, among other things, are due to selection effects: statistics courses are (in general) mandatory, they form a special requirement for social science majors and therefore students who traditionally score higher on mathematics (and therefore would do so on statistics), i.e. science majors, are missing from these analyses. Moreover, in most colleges females outnumber males in Social Sciences and, hence, also in statistics enrollments.

Hyde & Kling (2001) come to the same conclusion: in areas such as mathematical performance, females perform equally well compared to males. They claim it is theoretical sex stereotyping that causes these gender differences to occur. For instance, in trying to achieve their goals, women are said to take on a more intrinsic mastery goal, while men are said to use more competitive goals (extrinsic, to outperform others). If attitude tests only measure the more extrinsic goals, women are bound to have lower expectancies than men. If this is the case, these objectives cannot account for any gender differences, if they may be found. Furthermore, they argue that the Expectancy Value model uses expectancies that are related to task-specific self-confidence rather than global self-confidence, even more so by the hierarchy of task-specific self-confidence. This means that both males and females could have equal self-confidence with regard to mathematics or statistics, but if they are asked to rank it in a series of mathematics and English, women tend to rank their self-confidence higher on English whereas men rank their mathematics abilities higher. This comparison could result in a perceived gender difference.

Hence, I expect females to have a less positive attitude toward statistics than females, partly due to the internalized opinion about 'stats being a male oriented course'. Although one might expect gender differences, e.g. females starting with a less positive attitude toward statistics than males, it could well

be that female students change more. This could partly be caused by regression to the mean, and partly because I believe that stats is developed with more gender indifference than for instance math.

Task Value
Another aspect of the Expectancy Value Model is the 'task value'. Students are more motivated to perform a task if they value it higher. This 'task value' is said to be shaped by gender socialization processes, it could result in women's subjective task value to be lower than men's subjective task value.

Assumptions on gender differences
In this study I will analyze gender differences, starting from the assumption that if these differences are shown, they are not due to any differences in skills and abilities across gender, but that they are confounded by cultural, socialization and selection processes (Driessen & Dekkers, 1997).

2.5 Defining 'course outcomes'

What are 'course outcomes'? Student achievement, pur sang? Learning outcomes? Do you simply look at turnover? Number of passes relative to number of fails? Which methods of assessment and grading are used? Passing numbers? GPA? Or do you take into account assessment by students with respect to the courses, to the methods used, quality of books and readers, quality of the teacher? Hence, what is quality in teaching? What are characteristics of quality in teaching?

In this study, course outcomes are defined as the achievement a student shows as a result of taking the Introductory (Methods &) Statistics course. Course outcomes contain information on teacher- and student evaluations, turnover, but also student achievement regarding expected and final grade, as has been discussed in section 2.3.1. In this study the primary focus lies on this part of the course outcomes: student outcomes or student achievement. Hence, the stipulative definition of 'course outcomes' will be limited to 'student outcomes' or 'student achievement'.

2.5.1 Setting learning goals

What does the instructor test during and at the end of the course? What skills and knowledge should the student have mastered by the end of the course and at what level? In order to reliably assess Statistics skills and knowledge, learning goals have been developed. Learning goals constitute important information of what the student should have learned by the end of a certain week, term or semester. Learning goals are primarily set up for the benefit of the students, so that they can check whether they can already master these goals.

Additionally, the instructor develops a set of course goals, telling the students what goals the instructor wants to have obtained by the end of the course.

Course goals are, thus, more about teaching than about learning. First the instructors consider what should be learned in terms of content, abilities and values. Second, the skills students should master are established: what should students be able to do, know and understand. Then instructors will prioritize these goals and set the order in which these goals should be attained. Finally, the goals are translated into learning outcomes (Chance et al., 2004).

2.5.2 Assessing Statistical Mastery

Chance et al. (2004) define assessment as: *an on-going process of collecting and analyzing information relative to some objective or goal.* With respect to statistics education you can think of assessment as the continuous evaluation of students' progress in mastering statistical (and methodological) skills. The next question is 'how to assess whether these learning goals are attained'. A lot has been written about the numerous ways of assessing students' progress. Students' progress can be assessed individually, or in groups. It can be assessed during class meetings, by handing in homework or reports and papers, it can be assessed by presentations, final papers, mid-terms and exams (Garfield, 2003). Alternative ways of assessment are posters, group projects, portfolio's, research journals (or logs) and peer assessment.

Although it is important to formulate and assess learning goals, it is also important to constantly review these goals. The mastery of these learning goals can be seen in six executive phases, from 'knowledge recollection' to 'evaluation' in Bloom's Taxonomy (Chance et al., 2004). It is imperative that assessment can only take place when goals of this assessment have been set beforehand. In most cases, statistics teachers will set their so-called learning goals before starting a new course. These learning goals are then continuously critically reviewed and changed (if necessary), in a cyclic process.

2.5.3 Final grade

As was stated in the introduction of this section, the easiest way to look at student outcomes is to consider the grades given at the end of the term / semester. These grades form the hard data collected to analyze student outcomes. Final grade is a widely and commonly used measure of student achievement. In many studies 'final grade' was used as a dependent variable (see for instance, Breslow, 2005; Dillon & Kokkelenberg, 2002; Hall, 2008; Martins & Walker, 2006; Shachar & Neumann, 2003). 'Final grade' is considered to be an objective measurement of academic performance, compared to more 'subjective' measures such as attitudes, expectations, satisfaction and evaluation factors (Shachar & Neumann, 2003). According to Shachar & Neumann (2003), although 'final grade' is prone to 'assesor subjectivity' it is still a more objective measure than other factors.

Another reason for using 'final grade' is accessibility of this information. Furthermore comparability across institutions is an important reason. The grading is done similar across institutions, despite the use of so many different teaching

- and assessment methods. In the end, the final grade (4-point, 10-point and 20-point) is the weighted average of partial grades throughout the semester. All grading systems were transformed to a 10-point system. In general, two systems of grading are used:

1. Numerical grading system, going from 1 to 10. Usually a 5.5 out of 10 means that the student has just passed the course.

2. Letter grading system, going from F (fail) to A+ (excellent). This system has been adopted from the Anglo-Saxon system by some colleges in the Netherlands. Usually a D+ still means a provisional pass and from C- onwards the student has passed the course.

Expected grade

In what way does the final grade reflect the expectations the student had at the start of the course. And, during semester, did this expectation change? In order to answer this question, expected grade should be measured as a more subjective element of 'academic performance' (Shachar & Neumann, 2003) and compared to final grade. For what purpose? With this study I want to establish to what extent the student has realistic expectations of his own performance. Should the course become difficult at the end, expectations are assumed to go down. If the student is self confident, expected grade is assumed to be higher compared to students with a low self-confidence.

If expectations play a role in predicting student achievement, as in course outcome, it will be assumed that expectations mediate the effect of individual factors and attitudes on grade, as in fig. 2.6.

Chapter 3

Central question, subquestions and general empirical issues

3.1 Introduction

This study aims to test the effect of several educational and student determinants on the outcomes of Introductory courses in Methods & Statistics. Besides educational (i.e. institutional) indicators, I aim to investigate the effect of individual background factors, and attitudes toward statistics on course outcomes. Attitudes may change as a result of entering / following the Statistics course. Part of the analysis is meant to provide a statistical tool for analyzing attitude change and the effects of individual and institutional covariates. Furthermore, this study focuses on comparing determinants between colleges and universities throughout the Netherlands and Flanders. The study will be set up with a focus on empirical-theoretical sociology.

Finally, the results of this study can be used for the ongoing development and innovation of course material for Methods & Statistics at Universities and Colleges. After all, the aim of every university and college teacher is (or should I say 'should be' - apart from getting across statistical knowledge and skills - to motivate his / her students and to get across enthusiasm for the subjects taught. Furthermore, the knowledge, applications and examples taught during an Introductory course Methods & Statistics should make statistics memorable (Sowey, 1995).

3.2 Central question

The focus of this study lies on determining what effect educational and student factors have on the outcomes of Introductory courses in Methods & Statistics

at Universities and small-scale colleges. Hence, the central question is:

What is the effect of ***educational***[1] *(course) and* ***individual****(student) factors on course outcomes with respect to Introductory courses in Methods & Statistics at Universities and colleges in the Netherlands and Flanders?*

3.2.1 Subquestions for each research phase

Each phase of the research project contains its own subquestions. First of all I want to explore Statistics education throughout the Dutch and Flemish universities and colleges, and then use the outcomes of these explorations to develop instruments to measure the constructs under study. Secondly I want to analyze attitudes toward statistics, the effect of institutional and individual characteristics on these attitudes and, indirectly on student outcomes. Finally I want to develop a number of recommendations for future Statistics teaching. The following subquestions have been developed:

1. How are courses 'Methods & Statistics' taught at colleges and universities throughout The Netherlands and Flanders and what measurement instruments are used to measure student outcomes with respect to these courses?
   (a) What is the current situation on teaching Methods & Statistics at colleges and universities in the Netherlands and Flanders?
   (b) What didactical approaches do universities and colleges use with respect to the teaching of Methods & Statistics?
   (c) Which methods of assessment and grading are used?
   (d) Which methods of course-evaluation are used?
   (e) What educational factors can be detected in previous studies with respect to M&S Education?
   (f) What student factors can be detected in previous studies with respect to M&S Education?
   (g) What is the current situation on course- & student outcomes with respect to Methods & Statistics at colleges and universities?
2. How can student determinants with respect to courses in Methods & Statistics at universities and colleges best be measured?
3. How can student outcomes with respect to courses in Methods & Statistics at universities and colleges best be measured?
4. Which educational and student determinants affect student outcomes with respect to introductory courses in Methods & Statistics most? Are there gender and / or institutional differences?

[1] Educational factors will also be referred to as institutional factors.

5. What attitudes toward statistics contribute most to the student outcomes? Research shows that one of the individual aspects that affects student outcomes is attitude toward statistics. This attitude can be measured in six separate components, i.e. Affect, Value, Cognitive Competence, Difficulty, Interest and Effort (Schau, 2003). It is interesting to see whether these attitudes change as a result of the course being taught, over the semester. That is why I measure these attitudes twice and compare posttest to pretest results.

6. What statistical model can best be used for the analysis of the change in attitude towards statistics? This question can be divided into the following subquestions:

   (a) What model best predicts attitudinal change in statistics and their effect on expected and obtained student outcomes?

   (b) What is the added value of Latent Change Method Effect Models (LCMEM), if any, to the existing tools for analyzing change?

   (c) To what extent are attitudes and student outcomes influenced by educational and individual factors, and to what extent can this influence be combined using LCMEM and Propensity Related Method?

7. What recommendations can be made concerning the best way to teach memorable introductory courses in Methods & Statistics at universities and colleges?

3.3 Institutions under study: comparing separate systems

Eleven institutions (i.e. universities and university colleges) agreed to cooperate in this study. If I would have been able to use a much larger sample of institutions, a nested design would have been appropriate, hence a multilevel design: students within institutions. However, as only 11 institutions participate, such a design is not possible. The question then arises, what kind of comparison would be useful?

It is not my objective to make an evaluative comparison across institutions, as no competition was intended. What's more, such a comparison simply would not be not possible, as these institutions are functioning as completely separate systems, with their own policies, rules, educational-, research- and management system. Albeit all institutions have adapted to the Bachelor/Master system and they fall under either Dutch or Belgium educational law, each institution has its own emphasis, embeddedness into the educational and research community. Moreover, student populations differ widely across institutions, due to selection on the part of both institutions and the students themselves (based on what a particular University city has to offer regarding student-life.

In order to partially solve this issue on 'differences across systems' I will present some of my results within institutions, and sometimes offer a comparison on the most important institutional differences. This results in comparisons between Dutch and Flemish institutions and between universities and colleges.

3.4 Overview of this study

This study consists of a number of consecutive steps in order to answer the central question.

Qualitative set up
In a preliminary study, a series of in-depth interviews are set up, with coordinators and/or instructors Methods & Statistics at the universities and colleges under study. This qualitative part of our study has three objectives. First of all, intake interviews will be organized to discuss the participation of that particular institute in the quantitative part (i.e. the survey) of this study. Second, I want to collect conceptions and constructs that will help operationalize the questionnaire for the survey. Third, I want to collect more qualitative data (opinions, ideas and arguments) of instructors on how Methods & Statistics is taught throughout the Netherlands and Flanders. With the results of these in-depth interviews, the question on 'how Methods & Statistics is taught' will be answered. This part of the study will be described in chapter 4 of this dissertation.

Quantitative design
Next, a survey is organized among all first year (Social Science) students at the universities and colleges under study. This survey takes the form of a pretest-/posttest field-experiment with only one treatment condition, i.e. the Methods & Statistics course. During the operationalization, subquestions on how to measure constructs will be answered. The set up of the survey will be discussed in the method-section, chapter 5. Parallel to this, each department / institute fills in a questionnaire regarding the set up and duration of the course.

Quantitative analysis
During the quantitative analysis, I will answer questions what statistical model can best be used, what attitudes best predict student outcomes and what educational and institutional factors influence these attitudes. I will also look at gender and institutional differences. This will be done in two parts. First of all, uni-, bi- and multivariate analysis will be performed, using SPSS. The data will be validated, missing values analyses will tell us what the structure of missing data is in our dataset and differences across gender and institutions will be analyzed. Then, structural equation modeling (SEM) tools will be used to analyze attitudes toward statistics and their interrelations. These so-called SEM modeling tools will also be used to test the construct validity of this study, as compared to its original setup in the nineteen nineties, in the Unites States, the

details of which will be discussed in chapters 6 and 7. Secondly, more advanced tools such as latent change models and propensity score analysis will be used to determine what model best predicts student outcomes. The more technical part of the procedure will be discussed in chapter 8, the results and interpretations will be described in chapter 9.

Recommendations
Besides answering the central question in the conclusion, a number of recommendations will be given regarding the best way (to the view of the researcher) in which Methods & Statistics can be taught. The answer to this subquestion will be given in chapter 10.

Quality enhancement of a mixed method design
I specifically chose to set up this project in both a qualitative and quantitative manner, in a mixed method design. This triangulated approach is chosen because it enhances the reliability and validity of the results (for more arguments, see section 3.5). Moreover, it is chosen because I can look at the research questions from different angles (perspectives). First of all an insight is analyzed from the point of view of the teachers and / or coordinators, taking into account their perspective. Secondly, the student perspective is taken into account. As the student population is big, the logical choice was to organize a survey. As the institution population (N=11) is small, the logical choice was to organize interviews. In the next section I will take a brief look at the empirical tradition that underpins this approach.

3.4.1 Answering questions throughout this dissertation

In sum, the subquestions stated above will be answered throughout this thesis. Subquestion 1 will be answered during the preliminary phase of this project; the results can be found in chapter 4. Subquestions 2 and 3 will mainly be answered in the method section of this thesis, in chapter 5. Subquestions 4 and 5 will be answered in chapters 7 and 9, the results-section. Subquestion 6 is mainly answered in chapters 8 and 9 of this thesis. In chapter 10 I look back at all the questions answered before and and answers to the final subquestion (7) are provided.

3.5 Rationalization & Empirical Social Research

In search of an answer to the central question for this study the emphasis lies on modernization of education in terms of rationalization. A closer look will be on innovation of didactical concepts, new forms of learning, advantages and disadvantages for students and teachers. The methodological focal point lies on the performance and development of Empirical Social Research. This research tradition has been most productive in social research compared to other research programs such as functionalism and interpretative sociology. Not only produces

the Empiric Theoretic program many more results compared to other research programs. Hence, the Empirical Theoretical tradition has booked the most prominent progress over the past few centuries (Ultee, 1977, 380).

3.5.1 Signs of a research tradition

The theoretical center of the Empirical Social Research Tradition is formed by the Utilitarianism (Collins & Makowsky, 1993). The content of these theories consists of the following: people make choices in order to reach certain goals. These choices are based upon opportunities and constraints. People choose the option that best attains the goal, has fewest constraints and most opportunities (Ultee, Arts & Flap, 1992). This tradition has become known as 'Rational Choice Theory'. It concentrates on maximization of individual needs and wants. Therefore, in addition 'utilitarianistic-individualism' provides a better description for this theoretical starting point. Durkheim later added another aspect: not only did he speak of individuals, he also looked upon individuals as member of groups / societies with values and norms (Ultee, 1977, 220), a collectivist society (De Jong, 2007, 109-110).

The tradition of Empirical Social Research goes back until the 18th Century, when the first signs of development of Empirical Research were shown by Quetelet. This Belgium Statistician studied the development of crime, by the use of crime statistics. He also developed hypotheses on the prediction of crime-occurrence. Quetelet used multidimensional tables to prove his hypotheses. A century later, Durkheim - considered the founding Father of contemporary Sociology - showed his mastery by developing multiple informative hypotheses on suicidal behavior. He tried to find out why people commit suicide and what the link is to certain values in certain (religious) groups (Ultee, 1977, 228). His hypotheses proved to be of a very high standard and informative level.

Since the Second World War, Lazarsfeld further developed this tradition by analyzing the voting behavior of people in Postwar USA. One of his students, James Coleman, made an extensive contribution to the Empirical Research Tradition through his study on high school students and their social rewards and values. He also developed the 'Rational Choice Theory' as one of the anchor theories in the current Empirical Sociological Tradition.

This study wants to build on this Empirical Sociological Tradition, by analyzing attitudes of students toward statistics. These attitudes derive partially from individual backgrounds of students, partially they are formed through the setting in which the courses are given. Students form groups, with their own set of values and norms. The institutions they are embedded in, also have their sets of values and norms. The Expectancy Value Theory is the theoretical 'playground' and I want to analyze how students behave in these systems of values and norms and to what extent it affects their achievement. When educational developers know how these systems function, they can plan future teaching. One way of doing this is to use both qualitative and quantitative methods in a triangulated set up.

3.5.2 Triangulated Data Resources

Data triangulation means the use of more than two strategies of collecting information, e.g. a combination is sought of in-depth interviews, text analysis, and surveys. First of all it is considered a tool to validate the data (enhance internal validity), secondly repetition of data collection using a different strategy enhances the reliability of the results. After all, different strategies of collecting data have different advantages and disadvantages. The weakness of one method could be avoided by using another method to confirm your findings. Triangulation is seen as an important tool to enhance both validity and reliability in qualitative research. It strengthens the study by combining methods (Golafshani, 2003; see also Campbell & Fiske, 1959).

For this study a proxy of methodological triangulation is chosen, combining both qualitative and quantitative strategies of information gathering. During the first phase of this project, literature research is combined with in-depth interviews with coordinators of Introductory Methods & Statistics courses. The objective of the latter is to gain insight in experiences and opinions of the people that organize, develop and teach these courses. Furthermore, the objective is to explore the research field and develop measurable constructs from 'within' this field. During the third phase, a survey will be administered among students who take the introductory courses in Methods & Statistics in a number of colleges and universities. The main goal is to use the constructs acquired during the first phase, to operationalize them and to measure them among students. Surely, the question is to what extent the students evaluate the courses the same as the teachers do by using constructs that are familiar in the field of teaching Methods & Statistics.

3.5.3 Insider - Outsider bias

One of the main threats to this study is the so-called 'Insider-Outsider' bias. Merton (1972) describes the Insider-doctrine as the monopolistic of privileged access to knowledge by certain groups, based on biological or social grounds. Let us apply this to this study. It seems clear that analyzing one's own didactical concepts against concepts used in other Universities and university colleges holds the threat of having privileged access to inside information, and - what's more - a biased view on its usability and quality. Therefore, a teacher/researcher might favor her own institution compared to others. This of course is a disadvantage of the chosen approach. One way to overcome this is to share the results with colleagues from other institutions in a peer-reviewing situation and take good notice of each others' responses and comments. One way of doing this is to present the results to the Dutch and Flemish Statistics Education Community on Statistics Education Research Days in 2007 and 2006.

Should a researcher then act as an Outsider, a 'stranger passing by', an objective and independent observer? Merton claims that this cannot be the case. Why? First of all, both Outsiders and Insiders have their own 'ascribed status', together with their particular mindset, ideas and concepts. You can

try to look at a research object with an objective 'eye', but your opinion will be based on the ascribed status you already possess. Second, you can try to 'become' an Outsider or, as Merton claims 'have capacities to act as both Insider and Outsider'. This would not work, because it might be possible that the researcher has to deny what he always affirms in his doctrine.

Finally,...'trying to interchange Insider and Outsider information, the actual intellectual interaction is often obscured by rhetoric that commonly attends intergroup conflict.' (Merton, 1972, 36). I tend to agree with Merton, as having an insider view on my research could also be considered an advantage. As an insider, I can 'dig deeper' when conducting in-depth interviews because I will not take all answers for granted, but place them withing my own frame of reference based on many years of experience in the field. Setting up a survey could mean that I am better tuned to existing realities, pragmatics and processes.

Insiders are said to have a special ascribed status that enables them to access privileged information. However, as a University teacher and researcher analyzing your own didactical concept and practice, you have to treat this information with great care. What I really need is an intellectual controversy changing into an adoption of ideas and concepts. This can, firstly, be accomplished by setting up a reliable and valid design, replicable, measurable, intersubjective and as objective as possible. Secondly, the use of questionnaires and methods used by other institutions, proved to be both reliable and valid, is recommendable. Finally, intersubjectivity can be accomplished by interaction between scientists, intellectual exchange of information, collegial consultation, peer assessment and other evaluative instruments. The result of this can be an exchange of ideas, an understanding of each others' concepts, instead of a polarization between intellectual groups. Scientists need to understand that autonomy of science is subject to great pressure. Nevertheless, they need to pursue the truth, notwithstanding the threats to that same autonomy (Merton, 1972, 44).

There are a number of ways to minimize the threat of insider bias. First of all peer consultation and - evaluation could play a role, secondly the use of thoroughly validated questionnaires developed outside my own institution. Last but not least the triangulated approach and the addition of qualitative forms of research to this project minimizes the risk of insider bias. Especially the latter permits me to look more closely at the experiences, drives and motivation of expert teachers and allows me to keep an open mind to the opinion of teachers in other institutions.

Chapter 4

Objectives when teaching Introductory Methods & Statistics

4.1 Introduction

When setting up a project, validity is enhanced by first taking a look into the specific field of research and searching for suitable concepts and reliable operationalizations. Therefore, during a preliminary study, a qualitative approach is chosen to gain an insight into the organization of introductory courses in methods & statistics and into the opinions and experiences of their coordinators. In this chapter the results of these interviews are presented. I will start with a description of the method and a topic list.

4.2 Setting up in-depth interviews

The set up of this preliminary study is qualitative. One of the objectives for this study is to gather data and concepts that allow me to move on to the quantitative part of the data collection. One of my goals is to get to know the important constructs that can be operationalized into measurable items in the questionnaire. Not only can information and self-report data on Statistics education and attitudes be collected. It is my objective to also collect information that lies behind these data, and take the respondents' perspective. By doing in-depth interviews it becomes possible to generate hypotheses and themes that emphasize the respondents' perspective instead of the researchers' ideas and theories (Hoyle, Harris & Judd, 2002, 394). Additionally, doing in-depth interviews sheds a different light on our research question, making the answer more reliable and valid because I look at it from different angles, i.e. taking on a triangulated approach.

The set up of these interviews provides both enhanced validity and reliability as to the outcomes of the research, as the triangulated approach is looked upon as a way to combine qualitative and quantitative methods of research (Verhoeven, 2008). The advantage of such an approach is twofold:

1. hard data help the researcher to statistically test relationships and make inferences about the data that can be proved

2. besides hard data, researchers can also look at their research topic in terms of subjective perceptions and look at the arguments behind opinions (by the subjects)

The information reported here is gathered during the interviews with several departments of the universities and colleges under study. The university departments visited mostly teach Social Science students. However, at Honors colleges the population is more diverse, due to the nature of the curriculum (students take courses from all departments). In Liberal Arts & Sciences colleges, Methods & Statistics is also taught to students in the Arts & Humanities and in the Sciences department.

4.2.1 Population and sample

The population for this research consists of the faculty members of the following departments:

1. University College Utrecht - Academic Core Department

2. University College Maastricht - Academic Core Department

3. Catholic University Leuven - Sociology Department

4. Roosevelt Academy - Academic Core Department

5. University Ghent - Pedagogy Department

6. Erasmus University Rotterdam - Sociology Department, Section Methods & Statistics

7. University Maastricht - Health Department

8. Catholic University Brussels - Social Science Department

9. University of Amsterdam - Pedagogy Department

10. University of Amsterdam - Psychology Department

11. University Utrecht - Methods & Statistics Department[1]

[1] As I want to ensure confidentiality of the data, the institutions will be randomly assigned with a letter and every referral to the name of the institution is discarded.

For this part of the research project I started with a *snowball sampling method*. Starting from my own network I contacted the first coordinator from University College Utrecht and after the interview asked whether he would know someone with the same background and position at another university or college that would be willing to participate in this project. Thereby the former interviewee acted as a reference. The objective was to interview *experts* on Introductory courses in methods & statistics e.g. the coordinators of the courses or teachers and / or writers of statistical content. As the project developed the snowball sampling method was abandoned for a more *convenience* type of sampling method, where I contacted a number of coordinators from other institutions in order to obtain their permission for participation in this project.

From each department, the coordinator was approached with two questions:

1. is he[2] willing to give an interview on 'Introductory Methods & Statistics' in their college / department?

2. is he willing to let the students participate in the main survey.

Each interview lasted approximately one hour. The interview was recorded (provided the interviewee gives permission to do so) and a transcript was made. Second, the text was analyzed looking for 'sensitizing concepts'(Glaser & Strauss, 1967). Afterward the recorded interviews were erased. Before interviewing the interviewees, the curriculum of each institute / department was studied.

4.2.2 Interview topics

In order to give the interviewee the utmost opportunity to give his own insight or opinion, a topic list was used. The main topics are:

- Personal experience with teaching Statistics
- Current situation regarding Statistics courses in the interviewee's department & student population
- Should Statistics remain being mandatory and why?
- Course organization in interviewee's department
- Views on future developments with regards to Statistics Education at universities and colleges

An example of the topic list can be viewed in appendix A.

4.2.3 Objectives of the Preliminary Study

The results of the in-depth interviews are used in three ways:

[2]In this sample, no women were involved.

1. To gain insight in the opinions and experiences of the people involved in the teaching (Introductory) Methods & Statistics.

2. To acquire information on concepts / operationalizations /outcome variables / measurement instruments that can be used during the main phase of the project.

3. To acquire information on possible differences between the colleges and universities involved. If so, these differences could be controlled for during the analysis of the data.

4.2.4 Foundation for Qualitative Analysis

For the summary and analysis of the interview results, the foundations of Grounded Theory (Glaser & Strauss, 1967) have been used. This means that a number of distinctive steps are taken. During the first round of reading all relevant text is underlined, so that it will delineate the relevant information for this topic. Secondly, the text is divided into parts that deal with subtopics. Then for each subtopic, a very brief summary is developed, after which all the summaries are assessed, evaluated and ordered in order of importance. Also for each subtopic 'core concepts' are developed and for each of those labels the order of importance, the level of interpretation (person, group, institution) is determined. Finally a tentative causal ordering is applied and the text summarized for its main concepts (Verhoeven, 2008). In the next section, these results are summarized.

4.3 Interview results: How is Introductory Statistics currently taught?

In this section, an overview is presented of the way in which introductory courses in methods & statistics is taught at 11 universities and colleges participating in this study.

4.3.1 Group size, massive exams and difficult English lectures

Statistics is mandatory for all first year students in Social Sciences, as it is considered to be a basic academic skill that every student should master. According to the interviewees, most students often do not like to take statistics, as they experience Statistics anxiety.

Group size

In the honors colleges, class sizes do not exceed 25 students. However, in large university settings the group size is usually a lot bigger, in some cases even 500 students per course. Cutbacks in the departmental budgets over the past years

have resulted in fewer teachers, and group sizes became larger. However, most universities only teach Introductory Statistics once every academic year, despite the large number of enrollments.

Learning goals
In most cases, teachers operate on a longstanding teaching tradition. They have taught the course for a number of years and only make small changes every year. Therefore, most interviewees did not find it necessary to describe learning goals. Only a few departments have formulated explicit course- and learning goals for their introductory Statistics course. One or two departments even test and validate their learning objectives.

Research skills
According to most interviewees, students should acquire several types of research skills besides having knowledge of statistical methods: doing research themselves and being able to read and interpret publications about research. Introductory courses at Liberal Arts & Sciences colleges are considered a challenge because of the diversity of the student population (humanities, social science and science students). These characteristics also make it difficult to develop clear and feasible courses for all freshmen. Therefore, sometimes the aim for alpha students is that they acquire a minimum of (passive) knowledge of statistics.

Entrance level
According to most coordinators, the students' entrance level has deteriorated, partly due to the development of a 'study-center' approach in (Dutch) secondary schools where the mastering of theories and formulas has shifted towards learning by doing, leaving a lack of basic mathematical knowledge required for entrance levels of Statistics.

Statistics courses are usually organized by a separate Methods & Statistics section within the Social Science departments. Statistics instructors meet regularly to discuss the content of their courses, but in most cases these meetings are not formalized.

Teaching methods
Teaching methods include lectures, work groups, and projects. Most courses take 12 to 15 weeks, students meet once or twice a week. Furthermore, introductions to SPSS are offered, mostly by means of workshops, labs and exercises. In most cases a combination of theory and practice is sought, for instance by letting students do exercises on real-life datasets. The emphasis on methodology and statistics differs between the departments.

For large student groups, the lecture is the default teaching method. Interaction is more intense for small groups than for large groups. The use of Intranet (electronic learning environments) for course communication is not yet widely spread across universities; the Liberal Arts & Sciences colleges (or 'LAS' colleges) and social science departments have taken the first steps in that direction, for instance by introducing and testing a Workspaces-environment (alter-

natively Blackboard/Amico) for the exchange of teacher-student information, assignments and announcements. Email is still primarily processed by means of email-programs instead of via learning environments.

Student population
The student population in Dutch and Flemish universities and colleges consists of a majority of Caucasian women; in one case (Pedagogy) even 99% of the students are female. Most students come from high school or another tertiary level college (known in Dutch as HBO). Honors Colleges and Universities differ in the sense that the student population in the colleges are from an international background, and teaching takes place in English.

Statistical topics
Most universities and colleges teach the same topics: univariate and bivariate descriptions, inferential statistics, introduction to hypothesis testing, some bivariate testing, univariate regression and analysis of variance. Sometimes, a number of nonparametric tests are discussed. Mostly some methodological topics are covered: measurement level, strategies for collecting data, reliability and validity. A few departments (e.g. Pedagogy) offer combined courses in Methods & Statistics with an emphasis on quantitative research and statistics. Both LAS colleges organize 'real-life' projects for their students. They also discuss topics like 'how to write a research paper'.

Assessment and grading
Assessment and grading is very diverse: midterms, final exams, take home assignments, class exercises, individual papers, group projects and group papers. At most university departments grading is expressed in numbers (1 to 10), honors colleges use letter grading from F (Fail) to A+ (excellent). These letter grades are numerically transformed to produce the so-called 'Grade Point Average' that ranges from 0 (F) to 4 (A). The main test method used for large groups is one final exam with either open or multiple choice questions. In some colleges, attendance and active participation is graded. Most universities do not grade attendance, as it is not mandatory. Course outcomes (percentage passes) vary a lot between the universities and colleges, from about 50% to almost all students. If grading is spread over a number of elements, chances that students pass the course increase.

Course evaluations
Most courses are evaluated by means of a student survey every semester. Only when the results differ a lot from previous evaluations, teachers look into the course content for possible changes. Hence, evaluation results are largely ignored. Most interviewees anticipate small changes in the near future, although Statistics remains compulsory. They anticipate a growing emphasis on use of ICT-tools. Some instructors who now teach large groups would like to work in small groups, make use of project groups, additional assessment tools and Intranet.

4.4 Methods & Statistics at colleges and universities: a few comparisons

The Netherlands and Flanders
Dutch and Flemish universities do not differ a lot when it comes to the curriculum of the Methods & Statistics courses, background of students (mostly Caucasian), group size (I interviewed one small and two big institutions), and teaching methods. In Flanders in most cases the course starts with a series of lectures, followed by exercises (students have to make them at home or in smaller groups) and an exam. At one Flemish university in between two lecture series a 'virtual learning environment' offers an online statistics course for students.

Assessment methods in Flanders do differ from the methods used in the Netherlands, as in Flanders there are fewer assessment moments and therefore higher weights to the grades. In one case there was only one exam at the end of the course. As no 'LAS' college was interviewed in Flanders, a comparison was not possible. Moreover, the number of institutions interviewed was far too small in order to draw any conclusions as to general differences across countries.

As do their Dutch colleagues, teachers in Flanders experience a deteriorating entrance level of first year students in their Statistics classes, albeit for a different reason. In Flanders there is a polyvalent trend going on meaning that students with every kind of high school diploma can come to university. This results in rather extreme differences in entrance level across students. In the Netherlands, however, this deterioration is (in most cases) caused by the 'study-house' construction, where students learn how to apply tools to problems rather than learn to understand theories and formulas. This method has more and more become a 'toolbox' for high school pupils, but once at a Dutch university, the student lacks understanding of how those 'toolboxes' work.

Regular universities and 'Liberal Arts & Sciences' colleges
Between regular universities and so-called 'LAS'-colleges, a few systematic differences were observed:

- **Group size**. At first sight, the group sizes for 'LAS' colleges seem much smaller than for regular universities. The three colleges under study allow a maximum of 25 students in each group. These groups meet regularly, mostly twice a week for two hours. For possible projects thereafter, the group sessions are ended. Most regular universities have group sizes much bigger, up to even 500 students per lecture. However, these large audiences are usually split up into small groups for tutoring purposes, coached by a teaching assistant. In that way, the large audiences only listen to lectures from the professor, while the interaction, work on assignments and explanation takes place in much smaller groups, more comparable to group sizes in the 'LAS' colleges.

- **Teaching methods**. Much related to the first aspect is the difference

in teaching method. With small groups, giving lectures is quite different from large groups. The distance between teacher and students in small groups is smaller, hence there is more possibility for interaction.

- **Program intensity**. The intensity of the program is considered to be much higher in 'LAS' colleges than at regular universities. Students meet twice a week for two hours, they conduct student projects in small groups, they meet for assignments, exercises and lab-courses. This is partly possible because the students at those colleges live together, as the setting is residential.

- **Assessment method**. Another difference concerns the test method. For large groups of students in most cases one big multiple choice exam is organized. In small groups a variety of assessment methods is chosen, varying from exams with multiple choice and open questions, to student papers and presentations.

- **International character**. A clear distinction can be made when it comes to the international character of the institution. 'LAS' colleges really have an international student population, whereas regular universities teach in Dutch to a Dutch speaking audience.

- **Academic Core**.As research competencies and statistical skills are considered to be general academic skills, courses to teach Methods & Statistics in 'LAS'colleges have been developed in an Academic Core environment. The aim of the so-called 'core-courses' is to teach students general academic skills such as languages, argumentation and research skills. Courses within the Academic Core are often referred to as 'service teaching' (Moore, 1992).

In sum 'LAS' colleges offer small class sizes (maximum 25), 'intensive' learning program, e.g. two meetings per week of 2 hours each, a lot of homework adding up to more than 50 hours of active learning each week for 30 weeks. They organize intake interviews for every student that applies for the college and they offer a residential setting where students live and study throughout the Academic Year. Most instructors and tutors also live in the vicinity of the campus (see also Weltje-Poldervaart et al., 2001).

Chapter 5

Method

5.1 Introduction

The first aim of this study was to assemble constructs, definitions, aspects of learning and teaching statistics, historical data, insight in previous research et cetera. This was approached in two ways. A thorough review of existing literature and research results on the subject was performed. The result of this search can be found in chapters 1 and 2 of this dissertation, resulting in the central question and subquestions in chapter 3. In chapter 4 I discussed the results of the preliminary study: a number of in-depth interviews with coordinators of Methods & Statistics of the participating colleges and universities. This resulted in an overview of the current status of Teaching Methods & Statistics at universities and colleges in The Netherlands and Flanders.

5.2 Design Quantitative Methods

The next phase of this study consists of a 'field experiment', comparing individual and institutional factors from first year students from the universities and colleges under study. This design has been chosen because all educational research that is applied and of educational value should preferably be done in the field (meaning associated with a course, classroom, peer group, and instructor). Therefore I used the existing situation (the course being taught) as a set up for the data collection (Verhoeven, 2008, 109). The focus lies on students from the Social Science departments, because in this context introductory courses in Methods & Statistics are highly comparable and Methods & Statistics is mandatory in all cases. Student groups will also be compared on background, quality and experience.

In each institution data were collected on two occasions, at the start and at the end of Introductory courses Statistics. The time lag between pre- and posttest varied from 4 to 16 weeks. This variation in time lag between two measurements can cause effects to be different (part of the 'method effect'),

so this difference has to be taken into account. This is a field experiment, because data are collected in the field, while subjects (the students) undergo a certain stimulus, i.e. the Introductory Statistics Course. Baseline and follow-up measurement should reveal whether the students' attitudes toward this topic changed and to what extent it affects student achievement, i.e. the final grade. Furthermore, institutional data were collected by means of a cross sectional survey among teachers Methods & Statistics.

5.2.1 Participants

The population for this research project consists of students at universities and colleges that are required to take Introductory Statistics Courses during the first year of their college education. Mostly these students are Social Science majors, with the exception of the college students, where every freshman has to take Introductory Statistics.

Population
The total population involves all students in their first year at universities and colleges. For this study, the operational population consisted of 2667 students, i.e. all students enrolled in the Methods & Statistics courses under study (11 institutions).

Non-probability sampling
As in many educational studies (see for instance Budé (2007) and Tempelaar (2007a)), this sample has not been randomly selected. The reason for this has to do with time- and availability constraints. First of all the data collection was spread out over three consecutive semesters, i.e. almost 18 months. Secondly, not every university or college was prepared to cooperate and if they did, they did not always teach a course in the available semester. The data collection, therefore, had its limitations. It was chosen not to randomly select our respondents, with the well known consequences for both reliability and population validity. Having started with a snowball sampling method, I later added institutions that were conveniently sampled.

The most important consequence is that I will not be able to generalize the results to all first year students in the Netherlands and Flanders and this poses a threat to the external validity. Due to the volunteer bias in this study, generalization is not possible (Cook & Campbell, 1979, 74). However, generalization is not the objective of this study. Comparing a small sample of institutions consisting of a wide variety of university and administrative systems cannot result in statistical generalization to a larger population. Presentation of some results within specific institutions, as will be done in chapter 7, (as well as the comparison between universities and colleges, and between Flanders and the Netherlands) partly provides a solution to these problems.

Despite the nonprobability sample, the outcome of this study can give an indication of attitudes that students have toward statistics, how this may change after the course and what (other) individual and educational factors are impor-

tant predictors for their course grade (in the institutions under study). In any case the sample is heterogeneous enough (Cook & Campbell, 1979). In this respect, the usefulness and informativity level of the data are more important than the generalizability ('t Hart, Boeije & Hox, 2005; Verhoeven, 2008). To justify this claim, it has to be established in what sense the results of this study will be used. In my view a mix of both conceptual and instrumental use is important here. In the development of advanced models the focus is on the conceptual discussion, whereas in the implications for teaching Methods & Statistics the instrumental use is more important ('t Hart et al., 2005). Furthermore, as was stated by Rubin (1997) in large observational studies the emphasis often lies on availability of participants and data collected from natural surroundings rather than on setting up pure experiments. As will be shown later, part of the problem resulting from the nonrandomized design can be addressed by the use of propensity related covariates (e.g., Rubin, 1997).

5.2.2 Procedure for data collection in rounds

The data are collected in three rounds:

1. First round: from January to June 2006.
2. Second round: from September to December 2006[1].
3. Third round: from January to May of 2007.

The objective for this was (again) twofold. Firstly, due to teaching obligations data-collection on a continuous basis was not possible. Secondly, I had to accommodate to the teaching schedule in other universities. Sometimes the qualitative interview took place at the start of a semester where there would not be any Stats course (the requirement was an Introductory (and mandatory) Stats course for first year students). In that case I would visit the institution again at the earliest available opportunity and collect the data.

Pretest

There are two measurement-times: pretest and posttest. During the pretest measurement, I introduced the procedure to all students, present in their first week of the Statistics course. Usually this introduction was done during the first or the second lecture before the whole group. Participation was requested, questionnaires were handed out and taken in and then, the lecture continued. For this I traveled to all but one[2] universities and colleges under study and administered the pretest questionnaire.

Posttest

The posttest measurement took place in a different setting, mostly during or just

[1] With the exception of one institution where the semester ended in May of 2007.

[2] As I had to teach a class myself, one of the institutions offered to introduce the questionnaire to the students themselves.

after the final examination or during the final session of the Stats semester. The teacher in question would remind the students of the project and administer the questionnaire. In case of the examination, students would find the form together with the evaluation form besides their exam form. After having finished the exam, students had to hand in the questionnaire with the exam supervisor.

5.2.3 Dealing with institutional differences once more

Rutter et al (1979) conducted a study into the extent to which schools differ in the influence on children's progress. Their research focused on 12 London secondary schools. The strategy they used provides a thorough insight in both the measurement instruments used for this study and the research phases. First of all, Rutter et al. used both educational and student factors to study the possible effect on attainment.

What impact do differences between schools have on learning outcomes?
A problem with school variables is that in cross-sectional studies, differences between schools and the effect on student outcomes can hardly be measured. In order to achieve a thorough and reliable measure of these effects, schools should measure the starting level, then change their methods and after a period of time, measure the possible change. In other words, an extensive longitudinal experiment would be needed (Rutter et al., 1979, 5 e.v.). Our pretest- / posttest design would not be adequate. In order to overcome this problem, two things can be done. Firstly, school variables can be excluded from the setting. This is not a very satisfactory solution, since institutional differences would continue to exist and to affect the results. Secondly, starting levels can be measured and possible differences can be treated as a control condition.

Schau (2003)[3] presented a model that incorporates institutional characteristics and the (both direct and indirect) effect on student outcomes. However, literature shows that it is virtually impossible to clearly measure this construct without much white noise. Using the results of empirical models, Sorensen (2000) suggests that the optimal choice, a randomized experiment, is not feasible. Researchers would have to assign students (matched on a set of characteristics) to a random sample of schools with different teaching methods, and assess the differences in outcomes of learning after a few years. An alternative that has been used often is a field experiment, resulting in multivariate models where possible confounding factors, such as student attributes, are controlled by statistical means.

A setback, however, is the presence of a large number of unknown and unmeasured confounding external factors, such as school atmosphere towards learning, and unforeseen circumstances (such as changes in assigned class rooms, busy schedules regarding other courses, et cetera). The other problem is the definition of 'school effect'. For this study, the emphasis lies on learning outcomes of individual students, not on school effectiveness. Furthermore, it is not the

[3] see also fig. 2.2.

intention of this study to compare different universities and colleges to one another, as this comparison will be unreliable. Given the heterogeneity and the small number of institutions, statistical control using multiple (multilevel) models is not feasible. Instead, differences between institutions and their possible effects will be discussed more qualitatively in interpreting institutional effects.

5.3 Operationalization

In this section a description of the operationalization of the constructs will be given. An example of the questionnaires can be viewed in appendices B and C.

5.3.1 Questionnaires

Three questionnaires were used:

1. a PRE-TEST student questionnaire with: background (age, nationality, gender), Attitudes toward Statistics Inventory, 'global attitude' questions, statistics cognitive competence, prior stats experience, self report on math experience and achievement in high school, expected course outcomes and career value.

2. a POST-TEST student questionnaire with: background, Attitudes toward Statistics Inventory, 'global attitude' questions, statistics cognitive competence, prior stats experience, self report on math experience and achievement in high school, expected course outcomes and career value and number of hours studied.

3. a TEACHERS' questionnaire asking about:

    - the infrastructure of the course (duration, ECTS, number of enrollments)
    - teaching methods (lectures, workshops, student teams, etc.)
    - class size (or teacher/student ratio)
    - assessment methods

Part of the results from the in-depth interviews were used to operationalize the teachers' questionnaire. Operationalizations were mainly derived from the way the instructors organize their courses. One attribute has been left out of this operationalization, namely 'teaching quality'. It has not been added to my model for two reasons. Firstly ethical reasons prevented me from using teaching evaluations, simply because I did not get permission to use them. Secondly, evaluative tools differ widely across the institutions under study and a comparison would not give a clear result.

As was discussed in chapter 3 our data do have a nested structure, for the factors operate both at an individual and institutional level (Deinum, 2000) The institution will act as a control variable to see whether the analysis outcomes

differ across institutions. As was mentioned before, multilevel analysis was not possible since the institutional sample size is very small (N=11).

Pilot
Before collecting the actual data a pilot was organized to test a the measurement instruments. This pilot was conducted with a small group of first year students. The outcome of the pilot mainly resulted in minor textual changes.

5.3.2 Student questions

Instruments to measure 'Attitudes toward Statistics'
During both pre-test and post-test the SATS, Survey of Attitudes Towards Statistics is used. This inventory was originally developed to assess student attitudes towards statistics in the U.S.A.(Hilton, Schau & Olsen, 2004; Schau et al., 1995, 1999; Dauphinee, Schau & Stevens, 1997) at the beginning and at the end of Introductory courses Statistics. The first version of the SATS© holds 28 questions, later 8 questions were added. It has been revised, translated and adapted to the Dutch 'jargon' for usage in the Netherlands and Flanders. For this project the SATS36© is used in both pretest and posttest measurement. Additionally, a few demographics, such as gender, mathematics grade at secondary school have been measured.

The six components on the SATS36© are: Affect (6 items), Cognitive Competency (6 items), Value (9 items), Difficulty (7 items), Effort and Interest (both 4 items). An overview of all items per component and its definition is given in appendix D. Previous studies show a good reliability and validity (Schau, Stevens, Dauphinee & DelVecchio, 1995).

Likert scaling
The first 36 questions of both pre-test and post-test survey consist of statements about attitudes towards statistics. Students have to choose the answer that represents their view on statistics from (1) strongly disagree ... to (7) strongly agree (Likert, 1932).

Adding 'global attitude-questions' Experience with mathematics in high school, as well as self report on the results and expectations of mastery have been operationalized by the following questions:

1. First of all 'mathematics cognitive competence' is operationalized as
    - *'How well did you do in your high school mathematics courses?'*
    - *'How good at mathematics are you?'*
2. Career value was operationalized by means of the question: *'In the field in which you hope to be employed when you finish school, how much will you use statistics?'*

3. Statistics cognitive competence was measured with the question: *'How confident are you that you have mastered introductory statistics material?'*

4. Effort put in was measured by asking an *extra* question in the posttest questionnaire: On an average week, how many hours did you approximately study statistics outside class hours (Schau, 2005)? Answers ranges from 'less than 3 hours' to '15 hours or more', on an ordinal scale.

These global attitude items all have a 7-point Likert scale (Likert, 1932; Schau, 2005). Additionally, the number of years the respondents took mathematics in high school, as well as the number of statistics courses previously taken are added as (continuous) questions.

Post-test to pretest differences
In both surveys, the final questions have a slightly different set of answering categories, although 7-point-scaled. This is due to the fact that in the pretest 'expectations' are asked and in the posttest 'experiences', albeit with the same items. Additionally to 'Effort', in the posttest questionnaire the actual number of hours studied was asked.

Measuring Students' entrance level
Other important aspects while setting up the study are the conditions under which the universities and colleges take part in this research project. We have to ask ourselves whether universities and colleges in the underlying study have the same aims and ambitions, concerning statistics education. A few important conditions are:

- homogeneity of learning experiences. In other words: is the input from students at these universities and colleges the same?

- do they bring in the same qualifications, experiences across the universities and colleges under study?

This aspect needs special attention. In order to measure the starting level of students who enroll in the Introductory Methods & Statistics courses, roughly two solutions can be offered. You can either have students perform a placement test, in order to test proficiency (the entrance level) of Methods & Statistics or measure a proxy of this entrance level e.g. ask questions on previous experience in high school and self-report competencies.

It was decided to use the proxy measure of self-report competencies in high school, mathematics experience in high school and perception on 'mastery' of the topic. Students could indicate their perception of this level on a seven-point Likert scale and give the number of years they took Mathematics in high school and / or the number of college mathematics/statistics courses they already took. It is assumed that students in the social sciences, just entering the university, have a comparable entrance level. In order to control for possible selection effects, it was decided to study first year students, to collect data from

mandatory Methods & Statistics classes and to mostly collect data from Social Science students.

5.3.3 Measuring (expected) course outcomes

As was mentioned before, student outcomes are measured as part of course outcomes, by looking at students' exam- and test results. This approach causes some problems with the interpretation of the results, because the relative importance of obtaining a high grade (or an A) might differ across colleges and universities. Therefore, just looking at the final grade might not give the valid and reliable information needed. This was confirmed in the results of the in-depth interviews. Hence, an additional measure is introduced, that measures outcomes more subjectively: both pre-test and post-test scores include a self-reported measure of the expected result on a 10-point (ordinal) scale.

This use of self-reported (expected) learning outcomes is wide spread, mostly due to the need to avoid time consuming and costly procedures of objective testing (Pascarella & Terenzini, 1991). Pace and Pike (Pace, 1984; Pike, 1995, 1996) report a growing evidence to a high correlation between actual performance and self-reported results (see also section 6.2.2). Reports show that self-reported results of course results represent a reliable and valid measurement, when students are able and willing to report them. Therefore, to optimize self-reports five conditions must be met:

1. the information requested must be known to the respondent
2. the respondent faces clear questions (unambiguous)
3. the question refers to recent activities
4. the question does not in any way violate the privacy of the respondent
5. the question must merit a serious and thoughtful response by the respondent

(Bradburn & Sudman, 1988; Brandt, 1958; Converse & Presser, 1989; DeNisi & Shaw, 1977; Hansford & Hattie, 1982; Laing, Swayer & Noble, 1989; Lowman & Williams, 1987; Pace, 1985; Pike, 1995, 1996). In this study these conditions are met.

In order to deal with differences in grading systems, such as letter grades versus number grades and the Flemish grading system, a control question was added: besides asking students to report their expected (letter) grade, a small subsample of (college) students was asked to report a grade from 1 to 10 to indicate their expectancies on a 10 point (ordinal) scale. This control question was added because I expected that some student groups are not used to the letter system and this might affect the construct validity.

The dependent variable is formed by student outcomes (achievement) with respect to Methods & Statistics courses, as reflected in a continuous final grade, from 0 to 10. To avoid ethical problems, permission was asked to obtain the

final grade of the student from the University administration office. Should the student or the institution refuse, then the final grade is not used in the analysis. For the conversion of (expected) grades into the numerical system and the results of the analysis with the control variable, I refer to section 6.2.2.

Additional questions
Additionally, questions were asked about nationality of the respondent (do you have a Dutch passport, if not what passport do you hold?), nationality of the respondents' parents, degree the students are seeking (bachelor, master, phd) and the major of their choice (main field of study).

5.3.4 Teachers' questions

In the teachers' questionnaire, institutional data were collected (appendix B). First of all, in a number of open questions general information was gathered, the number of ECTS issued after passing a course, the course duration (in weeks), the major for which the course was given.

Class Size
Class size was operationalized by asking for the total number of students enrolled and the number of groups involved in that particular semester. The ratio of those two variables will be considered the factor 'group size'.

Teaching methods
Data were collected on specific teaching methods, such as lectures, work groups, project groups, individual coaching or otherwise. For each didactical method, the occurrence, frequency and length was recorded, as my aim was not to have teachers evaluate their own teaching methods (it is a well known fact that these self-report measures are subject to distortion), but simply take into account the teaching method used. If the teachers indicate that a certain method is used, they can fill in the number of times per week and the number of hours per time. By this I can distinguish between universities that emphasize certain teaching methods. Answers were coded as 'dummy's' and the number of times and duration were registered as a continuous variable.

Assessment methods
Teachers were asked to indicate what type of assessment method they use (exams: multiple choice, open questions or both; homework assignments, papers, projects, presentations or otherwise). Furthermore, they could indicate the number of assessments they have and the weighting factor for the final grade. The answers were coded as 'dummy's' and the number and weight of the methods were registered as continuous variables. Again, as evaluative self-reports could be distorted, it was chosen to only measure assessment-type and -weight as an objective measure of the methods used.

5.4 Issues of reliability and validity

5.4.1 Reliability

Much has been done in the design, to assure reliability. First, I ensured a large enough sample. In order to spot flaws in the questionnaire, a pilot was run and mistakes and mishaps were changed. For the data collection I used a triangulated design, i.e. in-depth interviews, questionnaires and literature research. Lastly, a detailed acknowledgment of the design, analysis procedure and findings is added to the report. Additionally, a reliability analysis will be performed in order to check whether the Cronbach α runs parallel to those of the original design.

5.4.2 Validity

Every study is prone to error, both systematic and random. This study also has some drawbacks regarding the validity of its design.

Internal validity
The major threat to the internal validity in this study is the lack of a control group (and hence, no random assignment). In order to reinforce my conclusions, propensity score adjustment will be used (see chapter 8). More detailed, the following threats to the internal validity are encountered:

- **Selection threat**. Students that study a specific major, usually hold a similar set of background characteristics with respect to high school profile, IQ, education of father and mother et cetera. In order to ensure comparability across institutions, almost only students with a social science major are entered in this study. An exception is the student population from the Liberal Arts & Sciences colleges, where students take a broad range of courses. In the first year they all take Introductory Statistics, irrespective of their major. However, as was mentioned in section 3.3., institutional systems differ, as do student populations that study in 11 different cities. Therefore, I expect some threat to the internal validity, due to the differences across subpopulations. A possible solution is a comparisons within institutions, as will be done in chapter 7. A multilevel approach as a solution is not feasible, because of the small institutional sample (N=11).

- **Maturation**. Since there is a difference in time lag between pretest and posttest measurements across institutions (due to differences in course organization), maturation might play a role depending on the duration of the course. In longer courses, maturation certainly plays a role. Controlling for 'duration' in some of the analyses might be a solution to this problem.

- **History**. A situation might occur in an institution between pretest and posttest, that seriously affects the internal validity, such as the hiring of

a new teacher. Although I have not been informed of such a situation occurring in one of the courses, history might still be a threat.

- **Instrumentation**. Because Dutch students in Liberal Arts & Sciences colleges still have to get used to the 'letter grades' instead of the numerical system, the construct validity for these questions is a problem. Therefore it was decided to add another question to the posttest questionnaire, therewith changing the instrument. However, I do not expect that this addition poses a big threat to the internal validity.

External validity

Mostly a convenience sampling method was used to ensure the cooperation of the universities under study, causing a threat to the external validity. However, as was stated earlier, the heterogeneity of the population is big enough (Cook & Campbell, 1979). In order to check external validity, gender and age can be used for a test of generalizability of the results. However, having statistically generalizable results is not the primary objective of this study. The results will merely be used as an indication of determinants of course outcomes. Hence, the recommendations done as a result of this study, should be evaluated thoroughly before implementing them in the school system.

- **Construct validity** Are we measuring what we are supposed to measure? As the SATS-questionnaire was validated thoroughly (Schau et al., 1995; Dauphinee, Schau & Stevens, 1997; Hilton, Schau & Olsen, 2004) and I have chosen to use the same items, I assume that the construct validity is fine regarding the Attitude items.

5.5 Analysis and results in two parts

Before turning to the procedure and results, it is important to look at the steps from now on. The analysis will be divided into two major parts. The first part consists of the results that answer the main research questions on background characteristics, attitudes toward statistics, expected and final grade. Besides, I will analyze differences across gender, time and institutions and look at the main educational and individual variables.

In the second part of the analysis I will look at the same individual and educational variables, attitudes and outcome variables, only from a different angle. I want to analyze what the added value is of the use of more advanced multivariate models such as Latent Change Models in comparison to more straightforward statistical tools, such as simple difference scores, t-tests and ANOVA's. Furthermore, I want to test whether it is possible to use the Propensity Score Method to combine distinctive variables into one combined effect variable. Lastly, I want to test a complex structural model, taking all change factors, important indicators and relations across variables into account. With these analyses I want to answer the question to what extent I can draw more precise conclusions compared to the less complex statistical tools mentioned above.

Chapter 6

Analysis Procedure I: Descriptives and Measurement Models

In this chapter the procedure for descriptive statistics will be discussed, as well as the validity study for the SATS-model. The use of parcels will be discussed as well as the measurement models that will be tested, along with fit indices and multi group comparisons. However, we will start with a description of the missing values procedure.

6.1 Missing values analysis

After the data collection it turned out that a lot of values are missing. A complete overview of the response is given in section **??** and appendix F, tables F.1 and F.2. Of all students 74% participated in the pretest, 52% of the students participated in the posttest and 32% of the students participated in both pretest and posttest. The data show some item-non-response, but very small and unsystematic($< 5\%$). Reason for these incomplete data are threefold: students start the course but do not finish it and therefore miss the posttest, students only take the exam at the end and therefore miss the pretest, and institutions did not hand in (all of) the questionnaires. The latter reason does not relate to this study and incomplete data as a result of this are Missing Completely At Random.

In order to obtain both reliable and valid results it is important to find out whether our missing data are:

- **Missing Completely At Random**. This means that the missing values are not related to any of the observed and unobserved data. The missing values occur randomly. In this case simple listwise deletion of those cases will be an acceptable solution.

- **Missing At Random**. This means that the missing data are only related to the observed data. In the case of this data-set it could mean that pretest missingness relates to a specific type of students who did not show up at the start of the course, whereas posttest missingness relates to students not showing up for the exam (second measurement instant). In this case imputation of missing data would be a solution.

- **Non-ignorable**. This means that the missingness is related to both observed and unobserved data. This situation is difficult to resolve; imputation related to external data patterns is an option (Garson, n.d.).

The missingness patterns are investigated using SPSS' (15.0 and 16.0) 'missing values analysis'. The procedures used are:

- Little's MCAR test for missingness completely at random. This test statistic has a chi-square distribution. Null hypotheses are that the grouped variables (by means of their attitude components) are Missing at Random or Completely Missing at Random, resulting in a nonsignificant test result (Garson, n.d.).

- Separate Variance t-tests. This tests if the missingness on one variable is related to the observed values of a second variable.

- For a more detailed analysis, SPSS pattern analysis is used. Here, the type of missing value is specified: either missing values, or system-missing values, extremely high values or low values. The missing cases are compared to the valid cases for the indicator variable and a t-test determines whether they have equal means. For instance if we want to spot a pattern in the item-variables for 'AFFECT', we perform a missing values analysis for the items that make up AFFECT. For any missing case we compare the means of the valid cases for the indicator variables in that group.

- Percentages of missing values will be displayed(checking whether $mv < 0.05$).

This missing values analysis will be run for each attitude-component separately, in the complete dataset with N=2,555. The results will be discussed in section **??**.

Dealing with item-nonresponse
Before the analyses start, in fact before the parceling procedure, the data will be cleaned with respect to item-nonresponse.

This is done by removing cases with ≥ 2 missings on the 36 SATS-items. If less than 5% of the items are missing, they are considered to be missing completely at random. The item parcels will be computed as shown in section 6.2.1 and appendix E.

6.2 Preparatory analyses and assumptions checks

Before analyzing measurement models, a number of preparatory analyses will be run, such as univariate descriptives and reliability analyses. Furthermore, bivariate relations will be tested using multiple statistical tests, such as t-tests and correlation analysis. Lastly, Analyses of Variance and Profile analyses will be used to test multiple group differences and prepare the analysis of change. Aim of these preparatory analyses is to check for assumptions and to check prerequisites for performing advanced analyses.

6.2.1 Parceling

In line with Schau (1999; See also Dauphinee et al., 1997; Hilton et al., 2004), Hau (2004) and Tempelaar (2007a; Tempelaar, Gijselaers, Schim van der Loeff & Nijhuis, 2007) the 36 items from the SATS36© questionnaire have been parceled. Little (Little, Cunningham, Shahar, & Widaman, 2002) defines parceling as *'an aggregate-level indicator comprised of the sum (or average) of two or more items, responses, or behaviors'*. In fact, you comprise a number of indicators into parcels that, in turn, are used as 'aggregate-level' indicators in structural equation models. These parcels will be used in the measurement model on Affect, Cognitive Competence, Value, Effort, Interest and Difficulty.

Reasons for parceling items according to Little et al. (2002) are reduction of non-normality and obtaining a more continuous scale, reduction of number of parameters for the measurement model (parsimony)or more theoretical arguments such as items 'belonging' together. The main reason for choosing parcels is that in this study I want to repeat and test the model that Schau built and I want to use parsimonious models as much as possible. According to Little et al. (2002), *"models based on parceled data (compared with item based models) are more parsimonious, residuals are less often correlated and they lead to reduction in various sources of sampling error"*.

Hau & Marsh (2004) suggest that for each component at least three parcels should be combined. This should prevent unstable results especially when population correlations (and latent factors) are uncorrelated. Tempelaar has used a similar method resulting in three parcels for each factor (2006, 2007a; Tempelaar et al., 2007). In the data for this study, however, the components are presumed to be highly correlated. Schau uses a different parceling system resulting in approximately 2 parcels per factor (Dauphinee et al., 1997). I have chosen to use the parceling system based on the procedure described by her (Dauphinee et al., 1997). This results in 2 to 3 parcels per attitude construct. Appendix E shows the procedure.

6.2.2 Conversion of different grading systems

As students at the colleges have to get used to the letter grading system, the question about expected grades (pre- and posttest) could be biased. Therefore with a small subsample of students (n=42) a control question was used by asking

students to give their expected grade both in the letter system and the numeric system. Results of a correlation analysis shows that there is sufficient overlap to recode the letter grade expectancies into the numeric system (r=0.927; $p < 0.000$). Expected grade (pre- and posttest) was recoded by converting A into 10 and A- into 9, et cetera. Because this expected grade does not precisely represent the numeric categories we are used to in the Netherlands[1] because they are an indication of expectancies, the measurement level for these variables is considered to be ordinal.

Final grade in the 'letter grade system' was converted into the numeric system, by recoding the values into the average of the numeric ranges for these letter categories. This was done because this final grade is considered to have a ratio measurement level, because (different from the expectancies) the weighting of partial grades results in a continuous value (such as a percentage), ranging from 0 to 10. Flemish grades can reach a total of 20 points. I only received the final grades for one Flemish university. They were recoded into the numeric 10-point system by transforming it into a percentage value.

6.3 Measurement Models

Performing a confirmatory factor analysis, the covariance structure of the (latent) variables will be validated (Hilton, Schau & Olsen, 2004). The objective of this part of the analysis is to test the construct validity of the SATS-model, i.e. to cross validate the model for the Dutch and Flemish universities and colleges. Additionally, I will test for institutional invariance and for gender invariance. The null-hypotheses are that the model is invariant across institutions and across gender.

After having parceled the SATS36© items into parcels according to the Schau criterion (Dauphinee et al., 1997), a first order measurement model is fitted. The main question is whether, in comparison with the results that Schau (Dauphinee et al., 1997; Hilton et al., 2004; Schau & Stevens, 1995; Schau et al., 1999) and Tempelaar (2006, 2007a; Tempelaar et al., 2007) found, this model fits our Dutch and Flemish data well. Since the model has been validated thoroughly, it is expected that this model fits the data equally well.

6.3.1 Separate pre- and posttest data

Figure 6.1[2] shows the most basic pretest measurement model as it was originally developed by Schau (1995; Dauphinee et al., 1997). In addition to that original model, I added two latent components: Effort and Interest (Hilton et al., 2004). Every component is allowed to correlate with every other component. Residual

[1] If the variables would be recoded into the exact continuous variable as final grade, the correlation between expected (letter) grade and numeric grade would still have enough (r(52)=0.999; $p < 0.000$).

[2] A=Affect; CC=Cognitive Competency; D=Difficulty; V=Value; I=Interest; E=Effort.

terms are not allowed to correlate at this (early) stage of the analysis. All residual and reference paths are fixed to unity.

Similar to the pretest model, a **posttest model** is tested (see figure H.1 in appendix H) to the posttest data following Schau (1995, 1997).

6.3.2 Combined pretest- and posttest data

The most complex measurement model in this part of the analysis is the combined pretest - posttest model shown in figure H.2 in appendix H. All components are allowed to correlate. Residuals are only allowed to correlate for identical parcels across measurement time, as it is assumed that residuals correlate across measurement moments (Hilton, Schau & Olsen, 2004).

6.3.3 Model fit indices

Besides the χ^2, I will use TLI, CFI and RMSEA in order to assess the fit of the models. Additionally, SRMR will be used occasionally. In appendix H a detailed description of these fit indices can be viewed.

Interpretation of the model parameters
In order to evaluate the relation between attitudes and the parcels that load most strongly onto the model, unstandardized and standardized coefficients will be used. Additionally, Covariance - and Correlations tables will be inspected.

Before assessing each model, multivariate normality will be checked by inspecting the Mahalanobis' distance and Mardia's coefficient (Mardia, 1970). For the latter, values under 10 will be rated as acceptable (albeit significant). Violations of multivariate normality will lead to bootstrapping procedures if possible[3].

6.3.4 Multigroup comparisons in the measurement model

Three variables will be tested for invariance: **Time**, **Gender** and **Institutions**. The 'Gender' multigroup results will be used to confirm the construct validity for this study (Hilton et al., 2004). Time invariance will also be tested to confirm the construct validity.

Institutional invariance
With institutional invariance it is tested whether the model holds equally true for all institutions. As has been discussed in section 2.3.3, this dataset could show some institutional dependency, because the data have been collected from students, nested in institutions. The bias that could result from these differences across systems can be overcome by using a multilevel approach to the analysis. In this dataset, however, a multilevel approach is not possible, because the

[3]If multivariate normality cannot be checked because of missing values being present, SPSS will be used to check for separate normality assumptions.

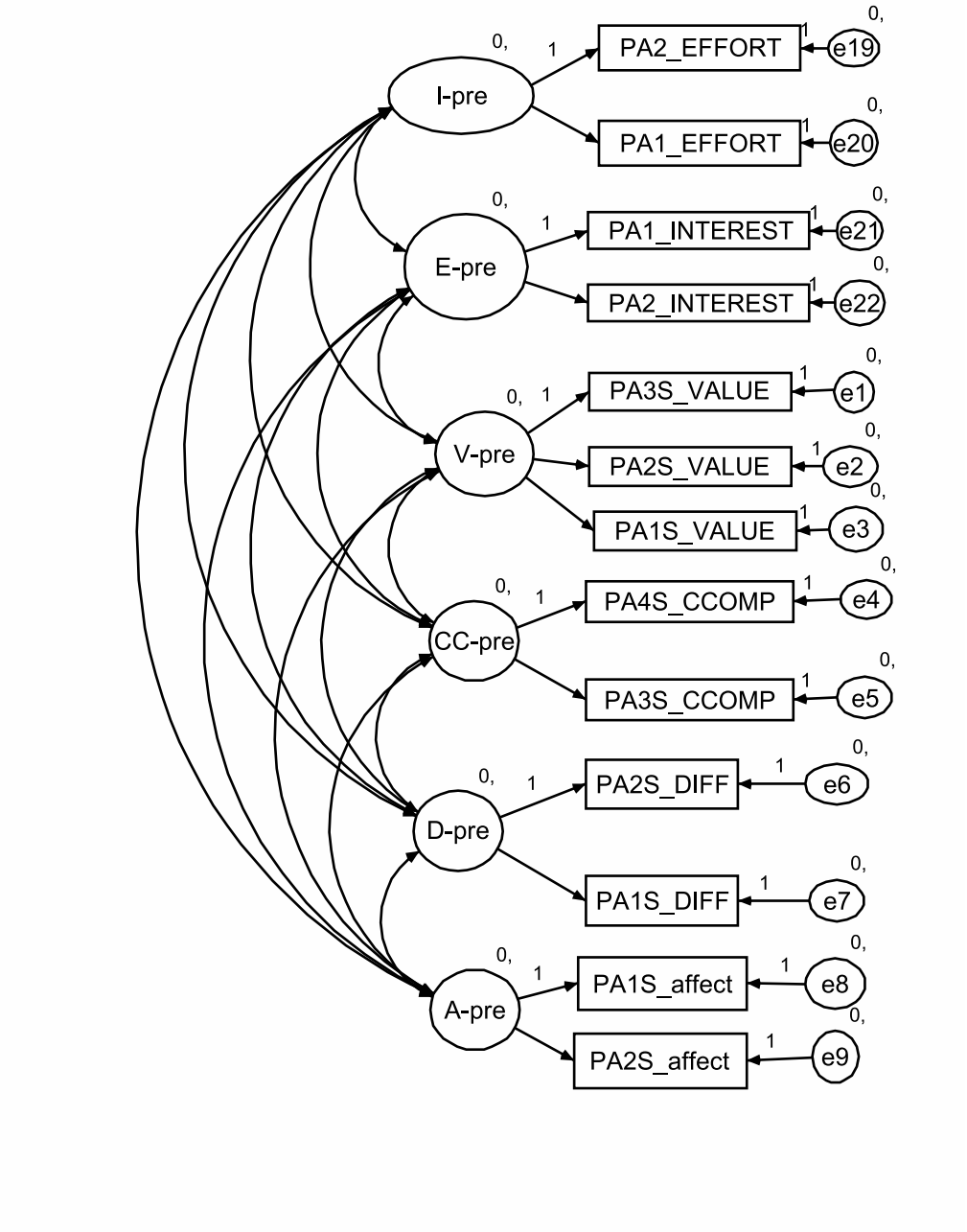

Figure 6.1: model pretest

institutional sample (N=11) is too small. Therefore, in addition to testing for institutional invariance, two solutions have been sought.

First, a few analyses will be run for separate institutions, such as the pretest-posttest attitudes. Secondly, a more dichotomous approach will be chosen and differences across 'universities and colleges' and across 'Dutch and Flemish' universities will be tested. These two comparisons represent the main differences across the institutional systems.

Time invariance

In order to test for time invariance, a different approach is needed. I have chosen two methods. The first method concerns the combined pretest- posttest model where constraints will be added to test for equality of loadings across time (Hilton et al., 2004). Another option to test for time changes (or differences) using a latent change model for each component. In part II of the analysis, results for these latent change models will be presented.

There are two main approaches to test for time differences:

1. Testing for time invariance with a complete pretest- posttest CFA model is used as a cross validation of the model tested by Hilton et al. (2004). In order to replicate this, I used the exact same method. This first approach tests factorial invariance. The results will be described in the first part of the analysis.

2. A second question of interest in this thesis is to what extent attitudes toward statistics actually change as a result of having taken a Statistics course. In order to analyze this question we are interested in true intra-individual changes and therefore we have chosen to use a method that is more appropriate to test for these differences. This second approach assumes factorial invariance and analyzes a change over time. The results of this advanced multivariate approach will be described in the second part of the results.

6.4 Analyses and sample sizes

The paired educational data under study show a lot of missing cases. In total, the grand dataset contains N=2,555 data, that will be used with structural equation modeling analyses. Three subsamples will be used: For the first-round analyses, the following datasets will be used:

- For all pretest analyses: N=1,976
- For posttest analyses: N=1,511
- For bi- and multivariate analyses with the combined data on both pretest and posttest in *SPSS*: N=861.

Both pretest and posttest subsamples contain information on one occasion of measurement. They will be used when analyzing only pretest or only posttest

measurements. Replicability of the American analyses is the main objective of these analyses. Prior to estimating advanced models in SPSS, a number of comparisons across measurement times will be done in SPSS, using the paired dataset with N=861 cases. All subsamples have been cleaned and therefore contain no missing cases on the 36 items.

The fitting of measurement models is accomplished in a number of stages, thereby following the data collection described in chapter 5:

1. complete pretest data (N=1,976, part of the students will also participate in the posttest).

2. complete posttest data (N=1,511, part of the students also participated in the pretest).

3. complete pretest - posttest dataset, cleaned and including missing values. This dataset includes all combined pretest- and posttest data, the part that only participated in the pretest and the part that only participated in the posttest (N=2,555).

Chapter 7

Analysis Procedure II: Advanced models

The purpose of the second part of the analysis (presented in this chapter) is to assess the added value of the use of advanced models compared to more conventional statistical tools for the analysis of change. In section 7.1 methods for analyzing incomplete data will be introduced, followed in section 7.2 with an introduction of the latent change models. Section 7.3 discusses the application of method effects. After introducing an application of the propensity score method (resulting in propensity related covariates) in section 7.4, I describe the procedure for a stepwise interpretation of Latent Change Method Effect Models with covariates and a dependent variable (sections 7.5 and 7.5.1). The main question for the second part of the analysis is: *what is the added value of the use of these advanced models compared to more conventional ways of looking at 'change' and the factors that influence it?*

7.1 Analyzing incomplete data

The combined pretest-posttest analyses so far have only used subjects that had complete data for the pretest and the posttest. This is equivalent to listwise deletion, the default procedure in most statistical procedures in SPSS. Although deleting incomplete cases from the analysis appears simple and effective, it does assume that the missing data are Missing Completely At Random (MCAR), which is a strong assumption. The preliminary analyses of the missingness mechanism (see sections 6.1 and **??**) indicates that the missingness is probably mostly MCAR, but it also shows some relations between observed data and missingness patterns, pointing towards data that are Missing At Random (MAR; the more complicated mechanism of data being Not Missing At Random is beyond the scope of this study). In this chapter, analysis methods will be used that analyze the incomplete data without deleting cases, effectively assuming MAR for the missingness mechanism. This is done using two different approaches.

One central approach is to employ structural equation modeling using the raw data likelihood approach, as implemented in Amos. The second approach is to use propensity score methods to adjust for possible differences between pretest and posttest due to nonrandom selection. This section introduces the Amos approach to incomplete data, section 7.4 describes the propensity score method.

When the data are MAR, two Likelihood based procedures are generally available to estimate a model directly on incomplete data: the EM-method and the factored likelihood approach. Amos uses the factored likelihood method which it denotes as the raw data likelihood method. Factored likelihood is based on the principle of separating the likelihood function into different parts for different groups. In structural equation modeling, each distinct missingness pattern is represented by a separate data group. In this formulation, all groups have complete data for a subset of the variables. Since the total log-Likelihood is an additive function of the log-likelihood for each group, Amos can use standard methods to maximize the Likelihood for all groups combined, using the same model. Factored likelihood can be used with the multigroup option in structural equation software, with each missingness pattern defining a group (Muthén, Kaplan & Hollis, 1987). If there are many missingness patterns, multigroup SEM is unwieldy, but modern SEM programs (such as Amos, Mplus, or Mx) allow raw data ML estimation directly on the observed part of the raw data. This is identical to the multigroup approach to incomplete data, with each individual case defining a different group, only such a model would not run in the multigroup option of classical SEM software. The procedure used by Amos is described in some detail by Arbuckle (1996). Simulations by Wothke (2000) show that the raw data likelihood method is unbiased when missing data are MAR, and that it is more efficient than classical methods as listwise deletion when data are MCAR.

7.2 Introducing Latent Change Models

So far the analysis of attitudes in statistics education has been conducted using conventional statistical tools such as bi- and multivariate methods and measurement models. If there are no missing values and a randomized design is used, an analysis could take place by means of simple difference-scores that would show an Average Causal Effect (ACE). However, in this study this is not the case.

Additionally, I want to analyze to what extent attitude changes can be attributed to true individual change and how much 'disturbance' from factors outside our control is encountered, how much individual and institutional factors influence this change (if any) and to what extent they influence course outcomes. This research question requires more advanced models, including latent variables, than have been used so far. The main objective is to find a model that provides a robust way of analyzing these simultaneous relations, taking into account method effects.

It is possible to analyze true intra-individual change on statistics attitudes, using an application of a latent change model originally developed by Steyer

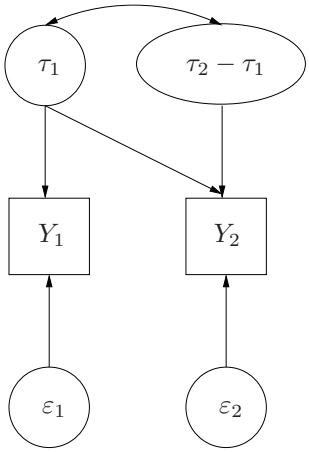

Figure 7.1: Steyers'(1997) model for Intraindividual change

(1990; 2005; Pohl, Steyer & Kraus, 2008; Steyer, Eid & Schwenkmezger, 1997). In this model true intra-individual change between two occasions of measurement can be made visible in the value of a latent 'change' variable (Steyer et al., 1997). The original theoretical model is depicted in fig. 7.1; factor loadings are constrained to '1'. Raykov states that the true score of the observed variables can be explained by a linear function of the true score of the measurement on occasion one, and the difference between the true score on occasion two minus one (1992). The value of the second latent variable represents the true intraindividual change (it will be referred to as 'change factor').

$$\mathrm{Y_{i2}} = 1 \cdot \tau_1 + 1 \cdot (\tau_2 - \tau_1) + \epsilon_{i2},\ where\, \epsilon_{i2} = Y_{i2} - \tau_2, i = 1, ..., m, m \geq 2, \quad (7.1)$$

and uncorrelated measurement errors,

$$Cov(\epsilon_{ik}, \epsilon_{jl}) = 0, i \neq j, k, l = 1..., n. \quad (7.2)$$

Source[1]: Steyer et al., 1997.

Statistics as a 'treatment'

Let us assume that the Introductory Course Statistics is a 'treatment', then the same latent construct (each of the six attitude constructs) is measured twice, once at the start of the course and once at the end. Every student undergoes this 'treatment' so there is no control condition: a paired design.

Let us depict the average causal effect as the difference between the posttest and the pretest measurement. If the Average Causal Effect (ACE) equals the

[1] In the equation *'m'* indicates the number of different measures considered (e.g. tests) to measure the same latent variable, *'n'* refers to the total number of measurement occasions, *'k'* and *'l'* are two measurement occasions; *'i'* and *'j'* regard two different measures of the same (latent) variable.

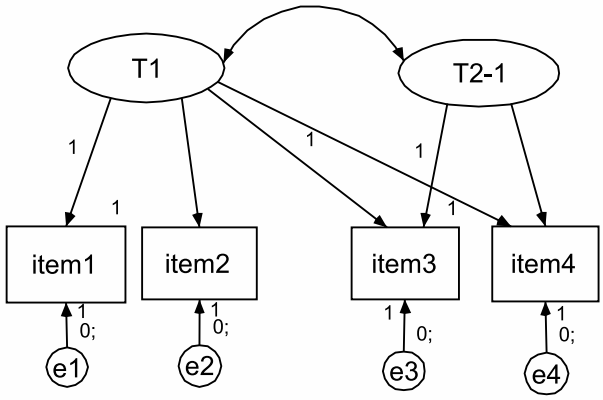

Figure 7.2: Basic Latent Change Model[2]

Individual Causal Effect (ICE) then every individual responds at the treatment in the same way. This of course is not true in most cases. Then there would be no 'prima facie effect' (effect at first sight; see also: Steyer, Partchev, Krohne, Nagengast & Fiege, 2007). The model that evolves is considered to be a latent **change** model, because it not only analyzes the path diagrams for separate measurement moments, but it also analyzes the change from the pretest on the posttest manifest variables, by looking at the means of and correlation between those two measurement moments.

Now, let us apply the information from fig. 7.1 to a given sample in fig. 7.2, with two measured items (in this case 'parcels') per latent factor, parallel to most statistics-attitude models. In this application, the notation changes from τ to 'T', because we are now looking at a sample instead of a population. I want to analyze true intra-individual change in statistics attitudes, i.e. (T_{2-1}). First, the true pretest score should equal the expected outcome score at T_1, second the true posttest score should equal the expected outcome score at T_{2-1}. If this assumption is true, according to Steyer (2005) the values of $(T_{2-1})-(T_1)$ can be interpreted as the individual causal effect, the expected value of $(T_{2-1})-(T_1)$ can be interpreted as the average causal effect of the treatment and the variance of $(T_{2-1})-(T_1)$ can be seen as the deviation of the individual effect from the average effect.

[2]As this figure forms the first of a number of applications to this study and the previous figure is an example from a theoretical population, different symbols are used compared to fig. 7.1. T_1 represents the latent trait factor in a given sample at t_1 and T_{2-1} represents the latent change factor at t_2. As the second latent factor is regarded as the '*change* factor' this is shown in the symbol by '2-1'. Because fig. 7.2 is an example of the general application to this study, the parcels have not been specified, but named 'item'. Note, that items 3 and 4 are equal to items 1 and 2 at measurement T_1. In the models to come, these parcels will be specified according to the attitude component that is modeled.

7.3 Latent Change Method Effect Models

Method effects play a role when in pretest-posttest designs differences in methodological circumstances occur, such as different settings and raters, and differences in time span (between pre- and posttest). This is especially true when the measurements take place using the same measurement instrument (Steyer, 2005). Thus, when modeling changes in attitudes it is important to distinguish between true change and change caused by a method effect, whatever this effect is. In this study for instance, the pretest was conducted by the researcher and the posttest by the own teacher. Additionally, the pretest was done during a lecture and the posttest was done at the end of the (final) exam. In one case the tests were done on the computer, in all other cases they were done by means of paper-and-pencil questionnaires. Albeit accidental, this could be considered as a mixed-mode set up. A method effect could be present, hence, it needs to be modeled. It is assumed that the method effect accounts for that part of the difference between the posttest and pretest measurement (Vautier et al., 2007) that cannot be attributed to intra-individual change.

LCMEM

The Latent Change Method Effect Model (LCMEM) is a true intra-individual change model that distinguishes individual method effects. Determining a method effect means I can isolate changes caused by something else than the true change, such as the design and the passing of time (Steyer, 1997). However, the aim of this analysis is not so much as to measure the effect of time or setting separately, but to isolate all alternative explanations from true change.

This model has been developed in accordance with the latent state-trait theory and multi-trait-multi-method (MTMM) analysis. According to Eid, Lischetzke, Nussbeck & Trierweiler (2003) method effects can be determined in different ways. A simple option would be to use correlated error terms, another option would be to use unique (uncorrelated) method factors to load onto the items in the model. However, with correlated error terms we would never know how big any method effects are and with unique method factors the latent method variables would be undefined.

The solution for this is to construct a LCMEM, with a latent method factor that is allowed to correlate with the two latent (change) factors (Pohl, Steyer & Kraus, 2007, p.11). With this method factor, consistent method effects, their mean and variance can be identified from the trait effect. So true intra-individual change can be decomposed into a trait effect, a method effect and an error term. In this case, the trait factor and the method factor are well defined as (conditional) expectations given a person variable. The individual causal effect in my model is the difference between the true score during the posttest and the true score during the pretest measurement. The imperfect correlation between the true score variables T_{2-1} and T_1 can account for the individual method effect (Pohl & Steyer, 2005, 2006; Pohl, Steyer & Kraus, 2008; Vautier et al., 2007). The conditional expectations are the true score variable, with the trait and method factor as a special condition of this true score variable.

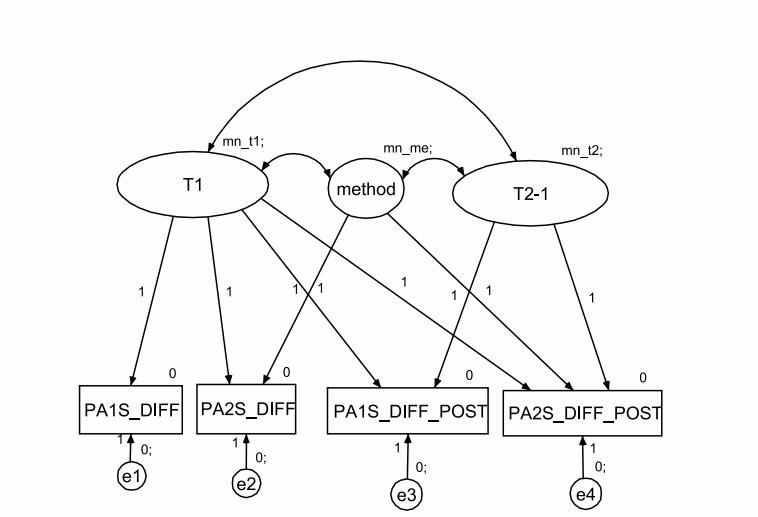

Figure 7.3: Latent Change Method Effect Model; example Difficulty[3]

Fig. 7.3 shows an example of an LCMEM applied to Difficulty for this dataset. For each of the attitude components, an LCMEM is tested (see appendix J for identification). According to Vautier, Steyer & Boomsma (2007) this model allows for two random effects:

1. The change of attitude over time, i.e. between T_{2-1} and T_1.

2. A **cross-sectional** method effect, the mean difference (and significance test) between the change factor (T_{2-1}) and the trait factor (T_1), i.e. $(mn-t2)-(mn-t1)$. Vautier et al. (2007) claim that this could be due to usage of different methods across measurement times. In this study the circumstances under which the questionnaires were administered differed a lot and that could have caused a method effect. Another cause could be that the questionnaire was administered by different people, i.e. by the researcher at the start of the semester and by their own teacher at the end of the semester. Lastly, the differences in duration between the measurements could have played a role; bivariate analyses in section **??** already showed that 'duration' did not disturb the correlation between pretest- and posttest attitudes.

[3]PA1S-DIFF and PA2S-DIFF refer to the pretest parcels; PA1S-DIFF-POST and PA2S-DIFF-POST refer to the posttest parcels. The labels have been named as follows: 'PA1' refers to 'parcel' and number, 'S' refers to the method used (Schau) and posttest parcels have an indication called 'post'. Lastly, 'DIFF' refers to 'Difficulty'. Labels such as 'mn-t1', 'mn-me' and 'mn-t2' refer to the means of the latent factors, see also appendix J.

7.4 Adding propensity related covariates

Whenever you are doing a research project, the ideal situation is that you can obtain complete data in a randomized design. Unfortunately that is not always possible, such as in this setting. In this study a pretest-posttest design is used. However, because of missing values a large number of students are absent on the pretest or the posttest. Since this is not a randomized assignment, the possibility of selection bias exists. Although the MVA analyses regarded in chapter 7 indicate that missingness may be largely completely at random, using a statistical control for the bias is desirable. Propensity scores can be used in order to control for any group (pretest/posttest) differences and to produce unbiased estimates of the treatment effect (D'agostino & Rubin, 2000; Rosenbaum & Rubin, 1983; Rubin, 1997; Rubin & Waterman, 2006). According to Rosenbaum & Rubin (1983; see also Imai & Van Dijk, 2004), a propensity score is *'the conditional probability of assignment to a particular treatment given a vector of observed covariates'*. Treatment groups, especially in non-randomized designs, often differ on a number of important characteristics, causing results to be biased. If we want to compare equal groups, we would have to control for these differences. One way of doing this is using the Propensity Score Method (Rubin, 2001; Shadish, Clark & Steiner, 2008).

In our design, the propensity score is used to predict group membership in the pretest and/or posttest group, thus offering some degree of control of the potential selection bias. The propensity related method has another, unforeseen advantage: the resulting covariate can be used as an simultaneous prediction score of a number of independent variables. The result is an unbiased estimator of the Prima Facie Effects (PFE), or in other words a conditional PFA given a combined covariate (in this case a propensity related covariate) as is argued by Steyer et al. (2007, p. 183; see also Kang & Schafer, 2007). Hence, in this study the propensity score can be interpreted as the probability a student changes his/her attitudes from pre- to posttest, conditional to institutional and/or individual background variables (Deheija & Wahba, 1997, 2002). The combined effect of one propensity related covariate yields a much leaner explanatory model as they reduce the number of variables (Steyer et al., 2007) resulting in a better fit. Therefore the use of propensity related covariates makes the models more robust (Kang & Schafer, 2007). The research question for this partial analysis is: to what extent do propensity related institutional and individual covariates affect the change in attitudes and, therefore the course outcomes in Introductory Statistics?

Comparable use of the propensity score method has been illustrated in several fields of study. A wide range of applications can be found in medical research (see for instance, Hirano & Imbens, 2001; Huang, 2003; Weitzen, Lapane, Toledano, Hume, & Mor, 2003). Recently, a growing number of Propensity Related methods have been published in Social Sciences, but also in educational research (Grunwald & Mayhew, 2008; Leow, Marcus, Zanutto, & Boruch, 2004; Pruzek, 2004; Rubin, 2001; Rubin & Waterman, 2006; Shadish, Clark & Steiner, 2008; Titus, 2006). Additional studies have proved the use of propensity scores

in complex survey data (Zanutto, 2006). An application similar to ours is the control for non coverage and nonprobability sampling schemes in volunteer panel surveys (Lee, 2006).

Because the propensity method is applied in order to obtain weighted indicator covariates additionally to just controlling for any existing group differences, the resulting covariates are referred to as '*propensity related covariates*' or *'PRC's'*.

What groups to use?
The Propensity Score Method (PSM) has predominantly been used for experiments with an experimental and a control group, performing either a Logistic Regression or Discriminant Analysis in order to obtain propensity scores. In the basic procedure the Propensity Score (PS) represents the estimated conditional probability that a subject will be assigned to a particular treatment (Pasta, n.d.).

In this study all respondents received the same treatment (i.e. the Introductory Statistics Course) and measurements were taken pairwise using a pretest-posttest design. Hence, there is no clear treatment or no-treatment group. In this study the choice of groups is therefore based on missing data patterns. In a complete pairwise setting, all pretest participants would also take the posttest. In this study, however, a large number of students did not take part in both measurements, such as students who started the course and then quit, or students who only participated in the exam. So, to some extent groups taking part in the surveys were different and could there could be baseline differences with respect to background factors. I decided to make the distinction between students taking only the pretest measurement and students taking **at least** the posttest. Students in the last group, at some point in time, went through a possible change during the semester because they took the Introductory course. Then, as Rubin states (1997; Rosenbaum & Rubin, 1983) differences in the observed covariates could lead to biased results, hence it can be determined which covariates influences the model the most. The propensity related covariate summarizes all the information from the covariates in one single value: the probability of being assigned to one of the groups.

Propensity score analysis (PSA) in this analysis will be used for three goals:

1. Baseline PSA results in balancing the scores for our two groups of respondents: students who only took part in the pretest, or students who at least took part in the posttest. The assumption is that the weight of the propensity related covariate results in an unbiased estimation of the (individual or institutional) effect.

2. Propensity scores are used for **'parsimony reasons'**. The combined propensity scores are built from a number of important group characteristics that influence the model. This combined score results in a less complex model compared to models where all influential factors would be included separately. In this study there are two main groups of assumed influential characteristics: individual (student) factors and institutional

(educational or course) factors. It simplifies the model extensively (Steyer et al., 2007). The assumption is that differences in model fit after adding the PRC are indications that there is an effect of the combined PRC, and that the probability that a student changes his or her attitudes throughout the course is conditional to institutional or individual factors.

3. The (interpretation of the) Standardized Canonical Discriminant Function used to construct the propensity score, can serve to assess the relative importance of the contribution of each included variable to the variates. In this way we can evaluate what variable has the biggest influence.

What variables to include?
Starting point for the selection of variables is the Expectancy Value model (Schunk, Pintrich, & Meece, 2008) where individual and institutional factors affect motivation and, in return, motivation affects outcomes such as Effort, expectancies and course outcomes (see chapter 3). The following variables will be included in the Propensity Score analysis:

- Institutional variables: (class) size, course duration and ECTS, didactical approach (teaching methods) and assessment methods (exams, papers and presentations)
- Individual variables: age of the students, major, nationality, self-confidence, math and stats experience and expected grade.

In sum, the institutional assumption is that particularly group size, the number of ECTS, didactical approach, assessment methods and course length influence the change, and therefore the course outcome. The individual assumption is that gender, age, self-confidence, stats and math experience influence the change in attitudes, and therefore the outcome of Introductory Stats course. These assumptions have been thoroughly documented in chapter 2.

Additional expectations are that institutions differ in the way 'attitude changes' are modeled. This is probably caused by the diversity of the course organization, the circumstances in which the courses are given and the characteristics of the student groups (selection threat) that enroll in statistics courses. However,I expect that the model will become too complex, when institutional invariance (for 11 institutions) will be tested. If this is the case a more dichotomous approach will be taken, testing differences between universities and colleges and between Dutch and Flemish colleges. Additionally, institutional differences will be looked at in a more qualitative manner.

7.4.1 Results of the Discriminant Analysis

In order to obtain propensity related covariates, a Discriminant Analysis with two groups is run in SPSS. The exact procedure is described in appendix K. Tables 7.1 and 7.2 show the relative standing of the variables with respect to group membership, as they will be used to interpret the relative influence on the latent change model. The individual variables that are most influential are

Table 7.1: Standardized Canonical Discriminant Function Coefficients - individual PS

variable name	**discriminant function 1**
self-confidence	-.873
age	.252
gender	.409
statistics experience	.401
how good were you in math during high school	.655

Table 7.2: Standardized Canonical Discriminant Function Coefficients - institutional PS

variable name	**discriminant function 1**
small work groups per week	1.273
small work groups duration	-.483
work groups duration	.300
course duration	.819
exam weight mpc + open	-.694
exam weigh mpc	.198
papers weight	.614
active participation weight	.361

self-confidence, gender, age, statistics experience and mathematics experience during high school, albeit the loadings are not too strong. As 'number of hours' confounded the loadings of other variables onto the Discriminant Function, it was decided to use 'number of hours' as a separate covariate instead of adding it to the propensity related score. Another reason for doing this is that the individual variables mentioned so far represent a cross-sectional influence on the model, whereas 'number of hours studied' represents a situational variable.

The most important findings here are the contribution of 'self-confidence' to group membership, followed by the perception of mathematics skills in high school. The fact that the two relative weights hold an opposite sign (positive and negative) indicates that group differences are explained by the difference between the two variables. Gender and statistics experience show a medium contribution to the variate, followed by Age.

The institutional propensity related covariate (table 7.2) consists of course duration, duration and occurrence of (small) work groups, weighting of exams (mpc and open) and active participation. The number of small work groups contributes most to group membership, exam weight and duration show a negative contribution. These propensity related covariates (PRC) are then used in the

LCMEM[4]. In sum, the loadings onto the discriminant function point towards an expectation that the adjustment for group differences, hence the effect, will not be substantial.

7.5 Summarizing the levels in the LCMEM on 'Statistics Achievement'

Fig. 7.4 shows the conceptual model consisting of four levels of modeling and interpretation:

- **Level I**: ***Covariates***.
  This is the level where the propensity related covariates, the combined individual / institutional covariates are located. They are the starting point, the indicators.

- **Level II**: ***Engine***.
  The main part of model building for this LCMEM takes place in the background at level 2, the **'technical'** level of the model building. Here all the different techniques can be found that were put in in order to identify the model and interpret the results. The propensity related analyses results in a combined latent covariate, pointing at the parcels that were developed for the 6 attitude components, and they build up into a combined pretest and posttest latent factor, the first one being the 'trait' factor, followed by the method effect and, in the end followed by the 'change' factor.

- **Level III**: ***Change***.
  This is where the interpretation of the changes takes place, along with the decomposition into true score, change score and method effect. This is applied to all six components. The covariates, indirectly, influence the model.

- **Level IV**: ***Achievement***.
  Finally, the Dependent variable is added to the model, and a complete model at all four levels can be fitted and interpreted simultaneously: attitude change, the effect of individual and institutional covariates and the effect on student achievement, i.e. final course grade. Additionally, expected grade is added as I assume the effect of attitudes on final grade is mediated by expected grade.

[4]Prior to the advanced modeling, a logistic analysis was run with two predictors i.e. individual and institutional PRC and 'pretest/posttest' grouping variable as a dependent variable. The results show that institutional PS are more likely to predict attitudes at the pretest measurement moment, whereas the individual PRC predicts posttest attitude (and therefore change) better. Additional method effects were tested with LCMEM.

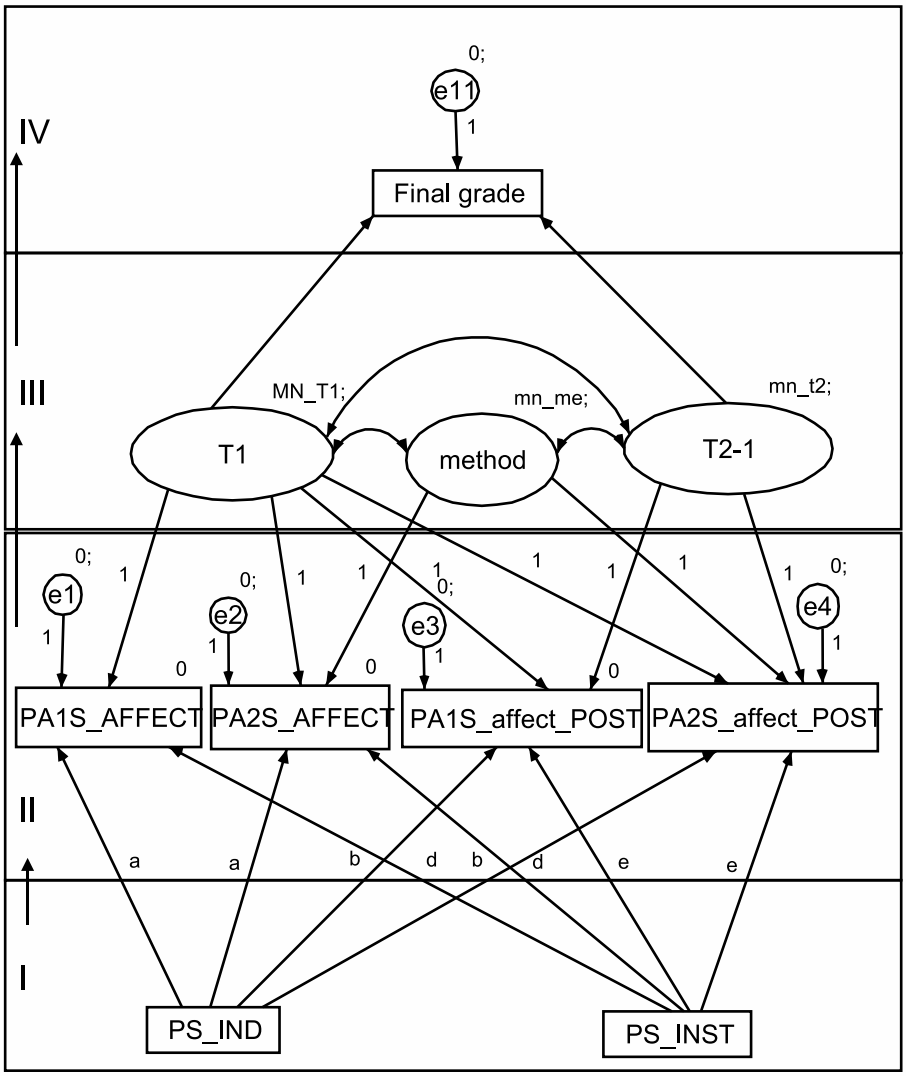

Figure 7.4: Levels of modeling in LCMEM

7.5.1 Application of the four-level approach to this study

The next step is to apply the four-level approach to this analysis:

1. First, the basic LCMEM, as depicted in fig. 7.3 is tested for all attitude components. In the 4-level approach, this is the 'Engine' and the 'Change' level in fig. 7.4. The purpose of this basic test is to model the existence of a method effect and look at basic change of attitudes. The results will be presented and interpreted in section 8.1.

2. In section 8.2 LCMEM's with propensity related covariates are discussed. The purpose is to assess the influence of these PRC's on the attitude-change. Addition of each covariate will be discussed separately. This analysis mostly concerns the Covariate level of modeling in fig. 7.4. With this result the added value of the advanced analysis will be shown. In subphases, covariates are added at the covariate level of the model:

   (a) Institutional and/or individual PRC's are added directly underneath the model (see fig. K.1 (app. K).), as an alternative to the so-called 'latent covariate' (mentioned by Steyer et al., 2007). Analyses showed that the model with a latent covariate does not fit better than the model with directly fitting PRC's. Moreover, interpreting a latent covariate is very complex and it does not enhance the clarity of the results.

   (b) Number of hours is added as a separate covariate..

3. Final Grade as dependent variable is added to the model with institutional and individual covariates. This represents the 'Achievement level' depicted in fig. 7.4.

   (a) Additionally, institutional and individual PRC's are added to the model, as well as 'number of hours'.

   (b) Expected grade pretest[5] is added as mediating variable. See fig. 7.5.

   (c) A complete LCMEM with covariates and the dependent variable will be interpreted in section 8.2.3. The purpose is to assess possible differences in outcome compared to more conventional results.

4. The leanest most explanatory model is presented in section 8.3. The objective is to look at the strongest effects on course outcomes, in a parsimonious model.

The example models have been developed for the attitude construct of 'Affect'. Each mode is constructed six separate times, for six attitude components. Identification procedures are described in appendix K.

[5]The pretest expected grade is added because it is assumed that this expectation better predicts the influence of 'final grade' as the students are still somewhat ignorant as to what to expect. Coming closer to the end of the course, the 'posttest expected grade' much better reflects the actual grade because students usually know part of their grade already.

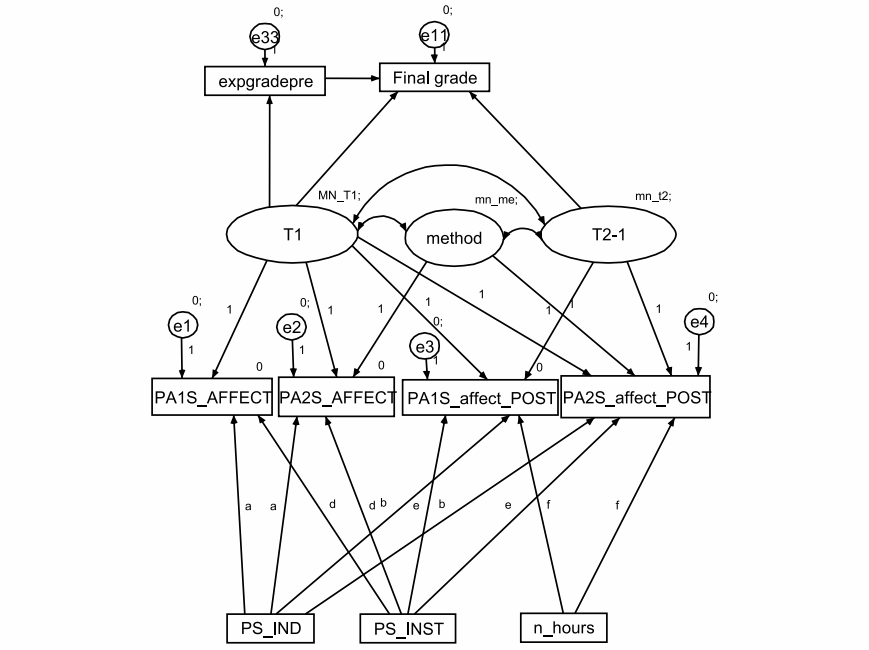

Figure 7.5: Model for Affect with covariates and Grade[6]

The main question for this part of the analysis is *to what extent the change in attitude varies with respect to the covariates and what effect does it have on 'course outcomes'?* In this part of the analysis, I will look at attitude change for the 6 attitude-components separately. In chapter 9, these separate findings will lead to the fitting of one hybrid model that enables me to interpret the results in terms of the theoretical model presented in chapter 2, fig. 2.6.

7.5.2 The 5th step: a hybrid approach

Lastly, a structural regression model is tested that contains the effects of separate indicators onto the change-factors of all six attitude components (in one model), the effects on 'number of hours studied, on expected and final grade. Residual values and correlations are tested simultaneously. The structure of and relations in the model are based on the results of the LCMEM with PRC's that will be presented in sections 8.1 and 8.2. The main question to be answered with this structural model is: *What are the contributions to 'course outcomes' from separate individual and institutional variables?*

[6]PA1S-AFFECT and PA2S-AFFECT refer to the pretest parcels; PA1S-AFFECT-POST and PA2S-AFFECT-POST refer to the posttest parcels. 'EXPGRADEPRE' refers to 'expected grade pretest'. PS-IND represents the individual propensity score, PS-INST the institutional propensity score. N-hours is the variable that measured the number of hours studied. All labels starting with 'mn' refer to the estimation of the means of latent variables.

Chapter 8

Results II: Advanced models

In this chapter the main results of the second part of the analysis on advanced models will be shown. As was described in chapter 8 (fig. 7.4), four levels of interpretation will be used. Part of these results can be viewed in appendices J, K and L.

8.1 Basic change and method effect

Table J.1 (app. J) shows the results of the model fit for the basic LCMEM at the Change level, as shown in fig. 7.3. For interpretation both fit indices and chi-square values are observed, although the chi-square values tend to increase with sample size. The model only fits well for two out of six components, i.e. Difficulty and Interest. The models for Difficulty and Interest show a good fit, followed by the model with Value and Effect. The models for Cognitive Competence and Effort need considerable improvement. The models do not adequately describe the data yet, as covariates and dependent variable still need to be added.

Table 8.1 shows the correlations between T_1 and T_{2-1} for the 6 basic component-models (LCM) and for models with a method factor (LCMEM). Correlations between the trait and change factors are usually negative, due to the technique of the model (Steyer et al., 1997), so only the strength is observed. In all but one component (Cognitive Competency) the correlation between latent factors increases as a result of adding a latent method factor, bus effect sizes in all cases are low (judging by the R^2, ranging from 2 to 10% for the basic LCM and from 1 to 11% for the LCMEM). Together with the significant variances in table J.2 (app. J) the existence of a method effect can be established. This means that the correlation between T_1 and T_{2-1} is suppressed by the existence of a method effect, so circumstances of the field experiment could have played a role in the change of attitudes over time. The mean changes in table J.2 correspond to

a low to (high) medium effect size (judged by Cohen's d ranging from 0.08 to 0.75).

Table 8.1: Correlations between T_1 and T_{2-1} for 6 LCM and LCMEM.

Component	**Basic LCM**	**LCMEM**
Affect	-0.162	-0.232
Cognitive Competence	-0.279	-0.227
Difficulty	-0.235	-0.276
Value	-0.143	-0.258
Interest	-0.313	-0.336
Effort	-0.083	-0.111

One possible explanation could be that the measurements were taken by two different people in pre- and posttest, differences in settings (pretest during a lecture, posttest after an exam), duration of the course so that a case of maturation could occur. The extent to which the method effect can partly be accounted for by other factors will have to be tested by adding covariates and, later, by adding a dependent variable.

In two out of six models the change across measurement periods is positive. This indicates that students have become more positive in their attitude with regards to Affect (the extent to which they have positive feelings towards statistics) and Cognitive Competency (self report knowledge and skills). For Difficulty, the change is negative, indicating that students did find it more difficult at the end than when they started. Attitudes toward Value (usefulness and worth in personal and professional lives) also decrease over the semester, as does Interest. The latter indicates that students became less interested in statistics compared to the start. The mean for Effort also decreases, but this indicates that students report of having put in less Effort than they anticipated.

8.2 Adding covariates to the models

The next phase of this results concerns the 'covariate-level' and the 'achievement-level' of the 4-step model. The main research questions answered are questions 6a, 6b and 6c from section 3.2.1.

8.2.1 Adding institutional and individual PRC's

Firstly, models with individual and institutional PRC's are fitted, both in separate models and combined. The PRC's are added directly onto the parcels. All models fit well (even slightly better than the basic LCMEM) with one exception: the model with institutional PRC does not fit for the data on 'Effort'. Apparently it does not depend upon the institutional setting whether students' attitudes toward Effort is set; it is more likely that this attitude is shaped under

the influence of individual covariates. This is confirmed by Astin (2003) who stated that institutional differences rather reflect upon individual characteristics of students than upon real institutional effects[1]. Additionally, both individual and institutional PRC's are added to the model simultaneously. Table K.1 shows the same results, again with a good fit for most components except for Effort.

8.2.2 Adding 'number of hours' as a covariate

Table K.2 shows that adding 'number of hours studied' to the posttest part of this LCMEM deteriorates the fit. The effect of 'hours studied' is significant in the model with Difficulty ($p < 0.01$), Interest ($p < 0.000$) and Effort ($p < 0.000$). For Difficulty the effect of 'number of hours' is understandable: the more difficult a student finds statistics, the more hours he will study. The (strong) effect of 'Effort' indicates that the more Effort the student reports, the more hours he or she studies. The effect of Interest shows that the more interested students are in Statistics, the more hours they put in. This is a sign of a more 'deep' learning approach.

Mean change and the influence of attitudes

In the appendix table K.3 shows the mean change across measurement period and the method effect for all six models that include both institutional and individual PRC's and 'number of hours studied'. The effect size (represented by Cohen's d) ranges from small to large (0.027 for Affect to 1.625 for Effort).

For Affect, the mean change is insignificant ($p = 0.629$); this is also the case for Difficulty($p = 0.093$) and Value ($p = 0.254$). Cognitive Competency shows a positive and significant change ($\Delta\ T_{2-1} = 0.261; C.R. = 6.017; p < 0.000$). Effort and Interest show a negative change ($respectively\ \Delta\ T_{2-1} = -1.606; C.R. = -28.602; p < 0.000; \Delta\ T_{2-1} = -0.495; C.R. = -9.964; p < 0.000$), but also significant. The negative change for Effort indicates that students report having put in less Effort after the course compared to the start of the course and the decline in Interest simply means that the Attitude toward Interest became less positive. All attitudes show a significant Method Effect (shown in the variances).

Comparing table table J.2 to table K.3 shows that in table J.2, all mean changes are significant, but in table K.3, the change factors for Affect, Difficulty and Value are not significant and two signs change. This indicates that adding covariates influences both the nature and the magnitude of the changes. Comparing the results of table K.3 against results of the bivariate results in table **??** two things are noticed. First of all under the influence of a method effect, the change in Affect turns insignificant. Secondly, determining method effects causes three out of six changes to become insignificant. Hence, method effects do play a role.

Standardized effects

Individual PRC's in all cases show a significant effect on attitude-change albeit

[1]Detailed results from these intermediate stages can be requested from n.verhoeven@roac.nl.

not very strong ($0.052 \leq \beta \leq -0.163$). Institutional PRC's show an even weaker relation: most effects are either non significant or very small ($-0.121 \leq \beta \leq 0.013$). The same result is shown for 'number of hours studied', only small effects, and only in two cases significant ($-0.061 \leq \beta \leq 0.035$). This clearly shows that the models need further adjustment.

8.2.3 Adding the dependent variable: Grade

The next step is to add the outcome variable to the models with the propensity related covariates, the 'achievement' level of the 4-step model.

Grade, institutional and individual covariates
In general the models with institutional and individual PRC's and Grade fit reasonably except for EFFORT and (to some extent) for VALUE. Yet another indication that Effort has a special position among the STATS attitudes. The fit becomes worse when 'number of hours' is added to the models ($251.67 < \chi^2 < 505.92(15)$; TLI for each attitude component model ranges from 0.567 to 0.828; CFI for each attitude component model ranges from 0.820 to 0.928; RMSEA ranges from 0.074 to 0.113). Attitudes do not mediate the effect of 'number of hours studied', especially when institutional PRC's have been added. Institutional variables do not seem to improve the fit of the model. Models without institutional propensity related covariates fit better in all cases (ranges reported for each attitude model are: $12.67 < \chi^2 < 273.08(5); 0.615 < TLI < 0.874; 0.839 < CFI < 0.944; 0.062 < RMSEA < 0.101$)[2].

Fitting a complete model
As a final step, a complete model is tested with institutional and individual PRC's, 'number of hours studied' and 'pretest expected grade'. Table L.1 (app. L) shows that the fit again becomes worse.

Institutional PRC's
Throughout this analysis, institutional PRC's have worsened the fit when added as a combined factor onto the models. First of all, as has been mentioned before, the nature of the institutional data is different from the nature of the individual data. They have been measured on a different level. Furthermore, the educational variables differ a lot across institutions, indicating that no general 'institutional influence' can be modeled. Thereupon it has been decided to (also) interpret the institutional differences in a more qualitative manner. This will be done in chapter 10.

[2]Detailed results from these intermediate stages can be requested from n.verhoeven@roac.nl.

8.3 The *leanest* most explanatory LCMEM

The final step in this part of the analysis is to analyze a model that fits best, with all variables at the same level, i.e. the student level.

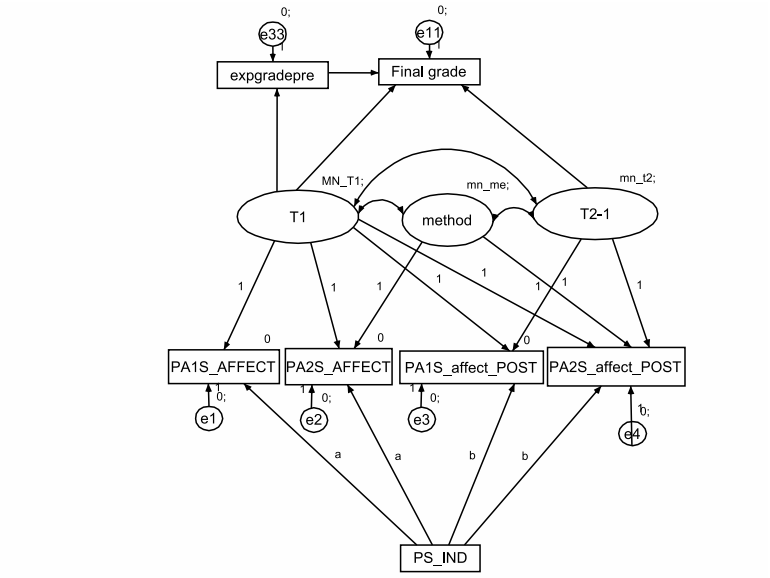

Figure 8.1: Example of best fitting hybrid model on Affect

'Individual covariates' explain attitudes and their outcome best (see also Verhoeven, 2008). Institutional covariates have been taken out, as the previous results indicate that institutional variables do not improve the model fit at all and as they have been measured on a different level, they should be analyzed as such. An adjusted institutional invariance test will be interpreted later in this chapter.

Additionally, I decided to leave out 'number of hours studied' because the mediating effect of attitudes with regard to 'number of hours studied' was not established. Hence, the leanest model chosen is shown in fig. 8.1. Fit indices in table 8.2 indicate a good fit for all components, including EFFORT.

These results indicate that changes in attitudes are conditional to individual background factors. However, the existence of an effect should be concluded with caution, seeing as the relative importance of the individual variables that formed the PRC were not that strong. Next, attitude changes, individual and method effects in this 'best fitting model' will be discussed.

Trait, change and method effects

Table 8.3 shows the mean values (and variances) for the Trait and Change factors, and the method effects. Effect sizes (Cohen's d) range from 0.002 for Affect to 0.968 for Effort. I compared these results to table J.2, the basic

Table 8.2: Leanest LCMEM with individual PRC and (expected) 'Grade'.

	Affect	**Cogn. Comp.**	**Difficulty**	**Value**	**Interest**	**Effort**
TLI	.947	.963	.944	.919	.951	.930
CFI	.983	.988	.982	.955	.984	.978
RMSEA	.053	.042	.044	.055	.046	.046
χ^2	72.88	48.67	54.14	218.06	58.52	57.33
P-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Df	9	9	9	25	9	9

LCMEM model and table K.3, the LCMEM model with individual, institutional covariates and number of hours, but without (expected) grade. Looking at the three tables, most values do not differ much. With the exception of Difficulty, the average is located in the upper part of the scale. Changes for 2 out of 6 factors are slightly positive, the other four are negative. This means that although PRC's and other variables significantly affect the model, the changes inflicted are in most cases not very big. Method effects are present in every model. A few differences are observed in the change-factors:

- The Change factor for Affect is only significant in the basic LCMEM.
- The mean change of Cognitive Competency decreases as more covariates and the dependent variables are added.
- The change factor for Difficulty is mildly negative in the basic and the complete LCMEM, but mildly positive in the LCMEM with covariates and without (expected) grade. The latter is insignificant.
- The change factor for Value not significant in the LCMEM without the outcome variable.
- The change factor for Effort decreases in all cases, but most strongly in the model without outcome variable (table K.3).

All in all the mean changes are negative (indicating a slightly more negative attitude after the course) and small, and only 'Cognitive Competency' shows a positive change indicating that students perceive themselves as more competent compared to the start of the Statistics course. A negative score on Effort indicates that students put in less Effort compared to their expectations at the start of the course. Effect sizes range in similar directions, Affect showing the smallest effect size (Cohen's d=0.002) and Effort showing the largest effect size (Cohen's d=0.986).

As in table 8.1, table 8.4 shows that correlations between T_1 and T_{2-1} in Cognitive Competency and Effort are not affected by individual PRC's, but the other components are (effect size reflected in R^2 range from only 1.4% for Effort

Table 8.3: Means and Variances of latent attitudes - best fitting model

Construct	Means	S.E.	P	Variance	S.E.	P
AFFECT						
T1	4.175	0.023	0.000	0.730	0.036	0.000
T2-1	0.002	0.032	0.957	0.675	0.046	0.000
Method	-0.416	0.020	0.000	0.178	0.024	0.000
COGN.COMP.						
T1	4.345	0.024	0.000	0.922	0.040	0.000
T2-1	0.170	0.029	0.000	0.574	0.039	0.000
Method	-0.110	0.020	0.000	0.204	0.024	0.000
DIFFICULTY						
T1	3.202	0.016	0.000	0.356	0.020	0.000
T2-1	-0.075	0.023	0.000	0.355	0.027	0.000
Method	0.178	0.016	0.000	0.104	0.018	0.000
VALUE						
T1	4.530	0.020	0.000	0.567	0.031	0.000
T2-1	-0.115	0.024	0.000	0.429	0.031	0.000
Method	0.414	0.017	0.000	0.208	0.022	0.000
EFFORT						
T1	5.952	0.020	0.000	0.584	0.039	0.000
T2-1	-0.999	0.039	0.000	1.065	0.067	0.000
Method	0.004	0.020	0.807	0.185	0.031	0.000
INTEREST						
T1	4.471	0.025	0.000	0.993	0.047	0.000
T2-1	-0.377	0.032	0.000	0.799	0.053	0.000
Method	0.042	0.021	0.043	0.176	0.053	0.000

to 13.2% for Interest). The correlations for Interest and Value become larger, for Affect and Difficulty they become smaller. This indicates that part of the attitude-changes are related to individual factors.

Table 8.4: Correlations between T_1 and T_{2-1} for the 6 best-fitting models.

Component	**LCMEM Best fitting**
Affect	-0.184
Cognitive Competence	-0.227
Difficulty	-0.225
Value	-0.286
Interest	-0.364
Effort	-0.119

Relative contribution of covariates to the model
Tables 8.5 and 8.6 show the results of the hypotheses tests for the covariates in the model according to fig. 8.1.

Table 8.5: Standardized estimators leanest LCMEM[3]

Construct	**Individual PS**	**Expected grade**	**Grade**
Affect	$-0.113 < \beta < -0.153$ (***)	0.450 (***)	$0.152 < \beta < 0.292$ (***)
Cogn.Comp.	$-0.112 < \beta < -0.130$ (***)	0.513 (***)	$0.106 < \beta < 0.369$ (***)
Difficulty	$-0.097 < \beta < -0.147$ (***)	0.379 (***)	$0.194 < \beta < 0.260$ (***)
Value	$-0.099 < \beta < -0.141$ (***)	0.185 (***)	$0.114 < \beta < 0.265$ (***)
Effort	$0.062 < \beta < 0.069$ (**)	-0.110 (***)	$0.019 < \beta < 0.270$ (n.s., ***)
Interest	$-0.044 < \beta < -0.064$ (**)	0.140 (***)	$0.024 < \beta < 0.276$ (n.s., ***)

All contributions are significant and positive, indicating that the higher the expected grade is, the higher the final grade will be. In the model with 'Interest' this effect is the strongest ($\beta = 0.276$) compared to the other relative weights. Thus, being interested in Statistics enhances the expectations and the final grade $\Delta_{effectongrade} R^2 = 0.093$.

[3]Note that in parenthesis the level of significance is shown: $* : p < 0.05$, $** : p < 0.01$, $*** : p < 0.00$. Furthermore, the β represents the ranges for all parcel loadings.

Table 8.6: Best fitting model, the effect of 'Expected grade' on 'grade'.

Component	**b-coefficient**	**S.E.**	**C.R.**	**P-value**	β
Affect	0.295	0.061	4.797	0.000	0.152
Cognitive Competence	0.205	0.064	3.193	0.001	0.106
Difficulty	0.375	0.060	6.288	0.000	0.194
Value	0.494	0.056	8.872	0.000	0.255
Interest	0.536	0.055	9.723	0.000	0.276
Effort	0.525	0.055	9.574	0.000	0.270

Additionally, I tested whether the effects of the individual propensity related covariates were equal across parcels. In all cases the fit of the models showed that the hypotheses that all parcel loadings are equal (within a measurement time) are retained. Furthermore, the models showed a significant contribution of the individual propensity scores on all parcels in all six models. In all cases but one (Effort) that contribution is negative. An explanation for this could be that changes in the model can be accounted for by group differences. The propensity score is in principle a balancing score and differences between groups merely indicate an unbiased estimate of the treatment effect (Rosenbaum & Rubin, 1983). Lastly, table 8.7 shows that the squared multiple correlations in the 6 best fitting LCMEM models are not very high. Cognitive Competency best predicts expected grade ($R^2 = 0.263$) and grade ($R^2 = 0.197$), followed by Affect and Difficulty. Explained variance of expected grade is minimal for Value, Interest and Effort, although for those components the explained variance in Grade is higher.

The effect of a PRC as a combined indicator is difficult to interpret. There are signs of a suppressor effect due to the combined nature of this individual PRC, indicating that the attitude change is influenced by individual variables. Additional analyses will be run in order to test the relative contribution of the separate individual variables to the model (see section 8.4).

Table 8.7: Best fitting model, squared multiple correlations.

Component	R^2 **Expected grade**	R^2 **Grade**
Affect	0.203	0.161
Cognitive Competence	0.263	0.197
Difficulty	0.143	0.165
Value	0.034	0.135
Interest	0.020	0.113
Effort	0.012	0.122

8.3.1 Comparisons across institutions

Adding institutional variables does not improve the fit for a number of reasons mentioned before. In order to obtain at least some statistical information on institutional differences, a dichotomous approach is taken:

1. **Nationality or 'region'**: a division into Dutch and Flemish universities (differences across regions of education).
2. **Type of institution**: a division into (more traditional) Universities and (newly founded) 'Liberal Arts & Sciences' colleges.

A first comparison (App. L, table L.2) shows that the model structure for all attitude components differs more often across 'type' than across 'region'. Table L.3 shows the differences in multiple squared correlations with and without baseline institutional comparisons. In general, only a small part of the variance in (expected) grade is explained by the factors for the 'institutional difference' models. Notwithstanding the differences across institutions, Expected grade is best explained in the models with Affect and Cognitive Competency across university type. Final grade has the highest explained variance in the model with Effort for Flemish universities (23.3%), indicating that for Flemish students the Effort put in and expected grade explain a lot of the variance in Final grade.

8.4 The effect of covariates - a hybrid approach

Lastly, *separate contributions* of the most important factors are tested in a hybrid structural model. The previously described 'four-level' model is not chosen here. Instead a hybrid regression model is fitted that comes close to the theoretical model presented in fig. 2.6, where all influential factors, a combined measure of 'attitudes' and the effect of Effort and Expected outcome are tested simultaneously. The change factor is represented by a latent variable extracted from every attitude-component. In this way, the method effect and the weighted change scores for every attitude-component (resulting from the LCMEM with covariates) are taken into account. The detailed procedure can be found in appendix L. The fit of the model is good ($\chi^2 = 86.47(31); p < 0.000; TLI = .954; CFI = .982; RMSEA = .026$). Fig. 8.2 shows the standardized estimates.

Standardized estimates

Looking at the standardized estimates in fig. 8.2 the loadings onto the attitude factor range from small to medium ($0.068 \leq \beta \leq 0.487$). There is a positive effect ($\beta = 0.487$) from self-confidence, AGE ($\beta = 0.155$) and good-math ($\beta = 0.068$) onto 'attitude'. The strongest predictor of 'final grade' is 'attitude' ($\beta = 0.253$) followed by 'expected grade'($\beta = 0.184$) and 'good-math' ($\beta = 0.161$). The strongest predictor of 'expected grade' is 'self-confidence'($\beta = 0.318$), followed by good-math ($\beta = 0.271$). Furthermore, 20.6% of the variance in 'final grade' is explained by the predictors and 26.0% of the variance in 'expected grade' is.

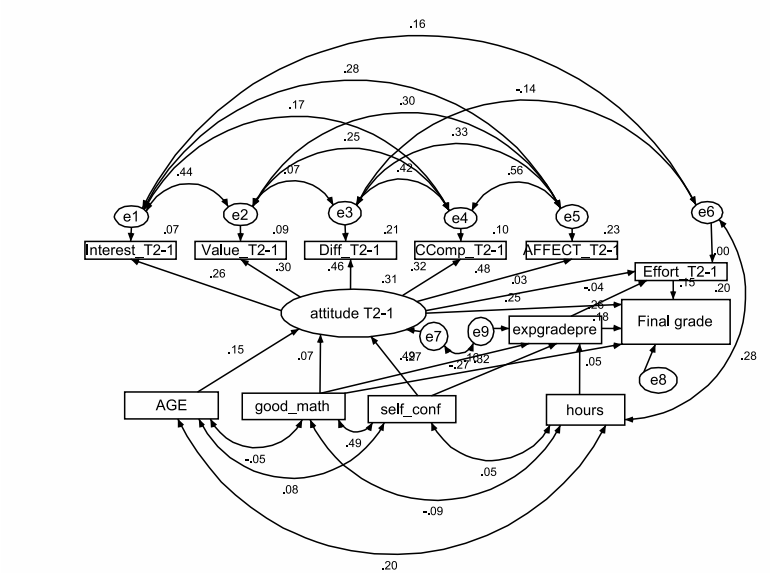

Figure 8.2: Standardized estimates regression model

A medium correlation is found between good-math and self-confidence ($r = 0.485; p < 0.000$), indicating that positive math experience enhances ones self-confidence regarding statistics. A small correlation ($r = 0.082$ is found between Age and self-confidence, an indication than when growing older, ones self-confidence with regard to passing increases. A negative correlation ($r = -0.048; p = 0.017$) is found between 'good-math' and age: the older one gets, the less positive a student thinks of his math competencies during high school. All predicted correlations with number of hours are significant, except between self-confidence and hours ($p = 0.061$). The negative correlation between good-math and hours ($r = -0.087$) indicates that the better a student perceives his math skills, the less hours he reports. The positive correlation between Age and hours ($r = 0.198$) indicates that the older one gets, the less hours one spends. The correlation between Effort and Hours is positive and significant ($r = 0.278; p < 0.000$).

Mathematics experience
According to the results in chapter 7, mathematics experience in high school relates to the expected grade of students, and to some extent to the final grade and to Effort. Having had more mathematics experience during high school influences course outcomes. In the final regression model, this finding is confirmed. How good a student perceives his or her mathematics achievement in high school has a small effect on attitudes, on student outcomes (grade) and a medium effect on expected grade ($\beta = 0.271$).

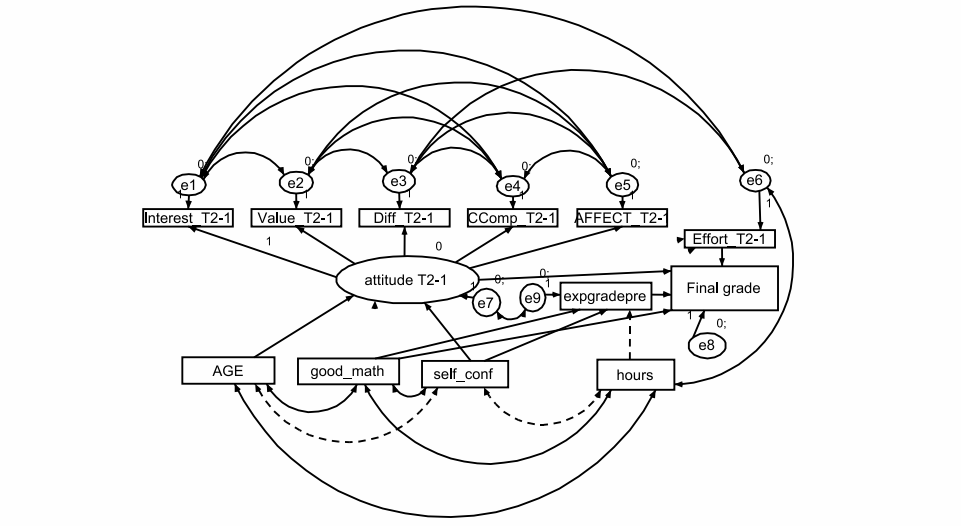

Figure 8.3: Final hybrid model with hypotheses tests[4].

Hypotheses testing
Fig. 8.3 shows the significance tests for the model. The following effects are not significant:

- Good-math on the latent 'attitude change factor'($b = 0.010; C.R. = 1.070; p = 0.284$)
- the Latent Attitude Change factor on 'Effort'($b = 0.127; C.R. = 0.200; p = 0.631$)
- Hours on 'expected grade'($b = 0.014; C.R. = 1.869; p = 0.062$)

Gender invariance testing
Instead of adding 'gender' as an indicator to the regression model, it is used here to test for multigroup invariance, just like in chapter 7. Comparing the baseline model to the metric invariance model for males and females a significant result is shown ($\Delta\ \chi^2\ = 23.76(\Delta df = 6); p = 0.001$), indicating that the model structure is not the same for males and females. Table 8.8 shows the differences, presented for the baseline model. Hence, the models are partially invariant across gender. In 8 cases, factor loadings are significant for females, whereas for males significant results only occur in 4 cases[5].

A short overview of the differences:

[4]The black lines indicate $p < 0.05$, the dashed lines indicate $0.05 < p < 0.10$, and the absence of a line indicates $p \geq 0.10$.

[5]After constraining factor loadings, the model for males changes more than the model for females. In the constrained model male factor loadings change to significant whereas female loadings remain the same. More detailed results can be requested from n.verhoeven@roac.nl.

Table 8.8: Regression model - hypotheses testing for males and females

	males		**females**	
Relation	β	r	β	r
Age → Attitude 2-1	.195		.132	
p-value	(0.382)		(0.019)	
Self-confidence → Attitude 2-1	.466		.488	
p-value	(0.358)		(0.000)	
Good-math → Attitude 2-1	.114		.071	
p-value	(0.472)		(0.292)	
Self-confidence → Expgrade	.405		.266	
p-value	(0.000)		(0.000)	
Good-math → Expgrade	.279		.285	
p-value	(0.000)		(0.000)	
Effort → Grade	.190		.108	
p-value	(0.018)		(0.013)	
Attitude 2-1 → Grade	.381		.322	
p-value	(0.390)		(0.000)	
Good-math → Grade	.217		.133	
p-value	(0.008)		(0.002)	
Hours → Expected grade	.101		.052	
p-value	(0.083)		(0.106)	
Expgrade → Grade	.018		.201	
p-value	(0.887)		(0.000)	
Age ↔ Good-math		-.103		-.029
p-value		(0.009)		(0.215)
Good-math ↔ Self-confidence		.498		.484
p-value		(0.000)		(0.000)
Self-confidence ↔ Age		.001		.110
p-value		(0.976)		(0.000)
Self-confidence ↔ Hours		.082		.036
p-value		(0.125)		(0.224)
Age ↔ Hours		.149		.216
p-value		(0.005)		(0.000)
E7 ↔ E9		.423		-.313
p-value		(0.365)		(0.000)
E6 ↔ Hours		.349		.224
p-value		(0.000)		(0.000)

- For females the following (causal) relations are significant, for males they are not:
  - A positive effect of Age on 'attitude' for females.
  - A positive effect of self-confidence on 'attitude' for females.
  - A positive effect of attitudes on final grade for females.
  - A positive effect of expected grade on final grade for females.
  - A positive correlation between age and self-confidence for females.
  - Attitudes (E7) and expected grade (E9) correlate negatively for females.
- For males the following (causal) relations are stronger than for males:
  - Being self-confident has a bigger effect on expected grade for males than females.
  - Age and 'being good at math' correlates more negatively for males than for females (n.s.).

Squared Multiple Correlations
The explained variance for males' Attitude is .317. Furthermore, the R^2 for Expected grade in males is .308. Lastly the R^2 in Grade for males is .310. The female model shows a smaller explained variances compared to the male model. Firstly, the R^2 for Attitude is .308 (baseline). For Expected grade the R^2 is .226 and the R^2 for Grade is .220.

Means for males and females
The means for males and females in the regression model differ to some extent. It is primarily the mean of 'number of hours' that differs (6.4 hours for females, 5.1 hours for males) and self-confidence (4.2 for females, 4.5 for males). These results support the findings from chapter 7. Differences across gender are consistent and they add to the previous results because with the more advanced models, more detailed explanations for the gender differences have surfaced.

Chapter 9

Conclusion and Discussion

9.1 Introduction

The primary focus of this study was to determine what effect institutional and individual factors have on the outcomes of Introductory courses in Methods & Statistics at Universities and small-scale LAS-colleges. The following central question was formulated:

What is the effect of educational (course) and individual (student) factors on the course outcomes with respect to Introductory courses in Methods & Statistics at Universities and colleges in the Netherlands and Flanders?

Theoretically speaking the best possible answer to this question is given by the goal oriented Expectancy Value Theory. Behavior can be seen as a result of the expectancies, norms and values a student experiences and the goal that the student formulates in a given course. It was expected that students choose the option that has the best possible combination of expected success and value. In terms of Statistics education, most students are expected to choose to put in so much effort that it results in passing the course. Highly motivated students put in more effort and strive for a higher grade. Students that are more motivated usually see a future use of statistics in their jobs, they want to pursue a career in research, they have had a great deal of positive mathematics experience in high school. In this conclusion I will also answer the question to what extent the Expectancy Value Theory holds true with regard to Statistics Education in my study.

Methodologically speaking the answer consists of two parts. First, straightforward comparisons will be evaluated by looking back at part I of the analysis, starting with conclusions on attitudes toward statistics, on individual and educational factors and on gender- and institutional differences. Additionally, I will evaluate to what extent Latent Change Method Effect Models provide good statistical tools to analyze changes in attitude toward statistics and to what extent

propensity related covariates can reliably predict these changes and (indirectly) affect final grade. Aspects of the methodological quality of this study will be reviewed, followed by recommendations for the development and organization of statistics courses.

9.2 Conclusion SATS©

In this section I will answer subquestions 4, 5 and 6 from section 3.2.1. They concern attitudes toward statistics that contribute most to course outcomes, the model that best fits the pretest and posttest attitudes and the model that best predicts attitudinal changes with regards to attitudes toward statistics. The latter will be done with reference to part II of the analysis, i.e. the 4-level method.

9.2.1 Attitude changes regarding Statistics courses

Except for 'Difficulty', attitudes are on average located in the upper half of the attitude scale. They do change over the course, although the change is not always in the positive direction. In most cases attitudes change for the worse. This was the case for Value and Interest, i.e. students do not grow to value statistics just because they took the course. One of the possible explanations could be the mandatory nature of Statistics courses. Students do not take Statistics out of their own free will, but they simply have to take statistics in order to fulfill major requirements. However this is by no means a plea for making statistics a course 'free of choice', because it remains an important part of academic skills. Part of the origins of these attitudes come from prior mathematics and statistics experience and 'notions' of what statistics is all about (Gal & Garfield, 1997).

Scores on attitudes toward Effort decreased. This means that students put in less work than they anticipated. In general students find Statistics difficult. Difficulty only increases slightly across measurement periods, indicating that students, in hindsight, perceive Statistics as slightly more difficult compared to when they started. According to bivariate comparisons, students show a little more Affect after the course than before, indicating that they report more positive feelings about statistics at the end of the course. However, results of Latent Change models shows a more diffuse picture. The change in Affect usually is very small and not significant, but in all but one case slightly positive. In general, students feel they have become more competent during the Statistics course.

What attitude best predicts student outcomes regarding Statistics?
Cognitive competency is the most important contributor to attitudes with regards to statistics education (has the highest factor loading), closely followed by Difficulty, Effort, Interest and then Affect and Value. Cognitive Competency improves according to students. With regard to Effort, students report

having put in less Effort than they expected. All other attitudes show some degree of decrease. Not only is cognitive competency the strongest predictor, the explained variance in expected grade and final grade are the largest in the 6 models that were tested. Moreover it has proved to be the most consistent factor throughout the analyses, whether it were bi- or multivariate tests or advanced models. The nature of this factor is individual, as individual factors play a more prominent role in course outcomes than external, institutional factors. This is confirmed by Wisenbaker, Scott & Nasser (2000). The SATS© can be applied to measure attitudes as an individual predictor of expected and final grade. If teachers want to make an effort in changing negative attitudes with students, monitoring these attitudes is important. Gal & Garfield (1997) state in this respect that attitudes should not only shape individual student learning but also the teaching approaches.

A special position for EFFORT

Effort has a special position when it comes to predicting student achievement. It shows a certain type of duality, possibly reflected in two different approaches to student learning, 'deep' versus 'surface' learning. It was shown in the results that one group of students reports a higher Effort score in the posttest measurement compared to the pretest and another group reports a lower score. This reflects having put in more or less Effort than expected. The clear distinction between two residual groups regarding Effort (see chapter 7, fig. **??**) points in the direction of a distinction between deep and surface learning approaches as mentioned by Tempelaar (2007a). This duality indicates that students may put in a lot of Effort for different (opposite) reasons. Either, they are seriously interested in Statistics and they put in a lot of Effort, or they think it is a very difficult topic that requires a lot of Effort to pass. On the other hand students may not put in much Effort, because they do not want to put in more Effort than is necessary to pass the course, or they perceive Statistics as being 'easy' and it does not take a lot of Effort to pass. Furthermore, no significant relation was found between 'expected grade' and 'Effort', yet another indication that expected grade does not lead to a certain Effort. Lastly, Effort does not fit models where institutional covariates have been added. It indicates that Effort truly is an individual factor.

The special position Effort holds in the Expectancy Model makes it difficult to interpret. Schau (2003) uses Effort as one of 6 components but the position of Effort in the model is not clear. Using Effort as a mediating variable would do more justice to the Expectancy Value Model. From this study it has become clear (see section 8.4) that Effort is situated relative to the other 'attitudes toward statistics' and that it actually mediates the effect of other variables on 'student outcomes'. This special position of Effort is confirmed by Tempelaar (2007b)[1].

There is something else. Effort has proved hard to operationalize. It is of-

[1]Tempelaar uses a different setup and operationalization for Effort, but he does show that the original position between the other attitude constructs as indicated by Schau (2003) is different. Effort partially mediates the effect of attitudes on student outcomes.

ten operationalized as 'number of hours' (in this study 'number of hours' and Effort were measured separately) but that operationalization is known for its unreliability with regard to the time during which Effort is measured, the inclusion of weekends, weekdays et cetera. Furthermore, there are reports about a problematic construct validity because it has not become clear how students define 'Effort' (Rau & Durand, 2000). I would like to define 'Effort' as some form of 'Academic Ethic', or the extent to which students take a deep learning approach. Rau & Durand (2003) state that students who possess such an Academic Work Ethic are intrinsically motivated to study and they put in more hours irrespective of their perception of statistics skills or GPA. They have, as they put it, *'an academic locus of control'*. In future research the factor 'Effort' could be better defined and simply split up in 'number of hours studied' and 'Academic Work Ethic'.

What model best predicts course outcomes?
At the 'achievement level', the model that best predicts course outcomes for students includes attitudes toward statistics and their change, expected grade as a result from those attitudes, and *individual factors* such as 'perception of mathematics skills in high school', previous statistics experience and self confidence. Additionally gender and age play a role. In this study, institutional factors only play a marginal role. In the next section, individual factors will be discussed (the so-called 'covariate level'), followed by another - more qualitative - look at institutional factors.

9.3 Interpreting the main individual effects

At the 'covariate level' I have analyzed the influence of several student determinants on attitudes (subquestion 4 in 3.2.1), and as a result on student outcomes. Additionally the (mediating) effect of 'expected grade' has been studied. The factors that influence course outcomes most are 'experience with statistics', 'mathematics results in high school', and self confidence. The effect on student outcomes is both direct and indirect, through 'attitude changes'. Besides this, gender and age play an important role in predicting student outcomes.

Prior statistics / mathematics experience
Prior statistics experience and the perception of mathematics skills in high school[2] have shown an effect on attitudes toward statistics, on expected grade and on final grade. This confirms earlier results by Wisenbaker et al. (2000) and Gal & Garfield (1997). The more experienced students are, the higher their expected grade is. Also their work motivation is higher, their attitudes are more positive, and as a result their grade is better.

This result is confirmed by the teachers in the universities under study, however in a different way. Dutch teachers complain that the entrance level

[2]Self report has proved to be a better predictor than 'number of years of mathematics during high school', as most students report the same number of mandatory years of math.

of students nowadays is deteriorating rapidly as a result of developments in the 'study center' in high schools, where pupils more and more learn 'tips and tricks' instead of proper mathematical and statistical methods, procedures and formulas. Flemish teachers also worry about the entrance level, but that has a different background. The polyvalent entrance policy in Belgium allows students of diverse high school background to go to university. A different reason, but with the same implication: it makes it difficult to teach statistics and keep the students motivated. German students (in this study) show an advantage with regard to this background. Having had more mathematics, more statistics, their expectations, motivation and final grade improve as a result.

Self-confidence
Self-confidence is a moderate predictor of expectations and of student achievement. It is related to previous experience with mathematics, it also positively contributes to attitudes toward statistics. Bivariate analyses substantiate these findings. Students with low self-confidence have lower expectations than students with high self-confidence. Moreover they also show a bigger decrease in expectations than students with high self-confidence. The latter group holds on to their self-confidence and the difference is only minimal. In sum there is both a direct and an indirect positive effect of self-confidence on expectations and on student achievement. Gender differences for self-confidence will be discussed in the next subsection.

Age
In this study the age distribution was not very wide. This is largely due to the fact that only students participated in this study besides a few part time students that were a bit older. Yet, there is a small but positive effect of age on attitudes, and a relation with the number of hours one studies. As a student gets older, he (she) matures (a student becomes an 'experienced' learner), this affects study-behavior and, in turn, student achievement. Furthermore, his or her self-confidence grows gradually, in return this affects the expectations and student achievement.

Number of hours studied
Number of hours has not proved to be a good predictor of student achievement: if you are not good at statistics, no matter how many hours you study, the grade will still be low. Effort and 'number of hours studied' are obviously positively related. As has been said before, in some studies Effort and 'number of hours' are operationalized in one measure, but I chose to measure it separately. Moreover it was shown that Effort directly affects 'final grade' and the effect on the student achievement of 'hours' goes (indirectly) through 'expected grade'. Apparently the two constructs do measure separate things. As you study longer, your expectations of the final grade rise (as if you want a reward for all your hard work), and through that (indirectly) final grade. As you put in more Effort, this more directly affects achievement although part of the effect is directed through expectations. Effort, in this explanation refers more to intrinsic motivation and

(as has been said before) to 'academic ethic'.

Future use
'Expectations of future usage of Statistics' is a weak predictor of student outcomes. Although there is a positive relation between future use and attitudes (the more positive the attitude is, the more likely the student thinks he/she will use statistics in his future career and vice versa), the variable was not very powerful and it was removed from the final model, as the effect of 'future use' on 'attitudes' is confounded by other predictors.

Expected grade
Expected grade is a significant but mild positive predictor of student outcomes. The higher expectations are, the higher the outcome is. Furthermore expected grade correlates positively with attitudes: the more positive attitude-changes are, the higher the expected grade is. 'Pretest' expected grade was preferred over 'posttest' expected grade, because this better reflects expected grade based on personal background and attitudes. Posttest expectations are closer to the final outcome, partially because many students already know their partial grade. The findings are in line with the expectations and they confirm previous research results (Svanum & Bigatti, 2006) that states that pretest measurements most accurately represent the students' expectations and that they are the most optimistic ones. Measures of expected grade taken at a later point during the semester (after some degree of assessment) show a decrease in expectations.

Previous experience also positively affects expected grade. Good grades in high school and a good self-confidence raise expectations and therefore outcomes. If students were good at mathematics in high school, expectations for statistics are usually good, they feel they have to put in less Effort and their final grade reflects their expectations well. The fact that self-confidence and previous achievement positively correlate with expectations and outcomes might partially be explained by a comparison between students at a higher and lower level of achievement. Usually higher previous achievement levels relate to higher self-confidence and therefore higher expectations. Turning this argument upside down, it has been shown that students with high self-confidence expect higher grades and although the expectations drop over the course of the semester, the drop for students with high self-confidence is still smaller than for students with low self-confidence. In turn, higher expected grades lead to higher final grades. Ideally the students' expected grade would contain some element of true self-assessment and an idea of the expectations by the teacher (by means of a good course manual for instance), also known as 'informed' expectations (Svanum & Bigatti, 2006; see also section 9.7.1).

9.3.1 Gender differences

As was stated in chapter 2, reports on gender differences with regards to student achievement show various results. Many research results report that females have lower expectancies and self-confidence, and higher anxiety levels than males

(Greene & DeBacker, 2004; Baloglu, 2003). To what extent do the models in this study differ across gender? The findings are in line with previous literature (Hilton et al., 2004; Harris & Schau, 1999), meaning that males and females partly differ across the effects of separate indicators. For instance:

- Females show a more negative attitude toward statistics than males in all aspects but one: Effort. On Effort, females report that they will put in more Effort than males and, after the course, although both averages decrease, females still reported to have put in more Effort than males.
- Female students put in more hours to study Statistics than males.
- Self-confidence with respect to statistics is lower for females than for males.
- The relation between age, self-confidence and the 'attitude change' is positive (and significant) for females but much weaker and not significant for males.
- The effect of attitude change and of expected grade on final grade is positive (and significant) for females and not significant for males.
- Males have a more optimistic expectation of their grade than females, irrespective of their level of self-confidence.

In general the baseline models fit the female sample much better than the male sample. A possible explanation for this is the fact that in Social Science 2/3 of the population is female, and therefore a female 'model' is more prominent. However there are some 'typical' gender-related findings. For males the perception of math skills and self-confidence does not predict attitudes but it does positively affect expectations. For females, attitudes and expected grade predict course outcome better. This to some extent supports the assumption that differences in socialization and selection instead of skills underlie these findings (Driessen & Dekkers, 1997).

9.4 Interpreting the main institutional effects

In this section the effect of institutional covariates on student outcomes is discussed, as well as the extent to which latent change models are influenced by institutional covariates. Additionally the mediating effect of attitudes will be considered. The subquestions from section 3.2.1 dealt with are 4 and 6c. Those questions have proved to be difficult to answer.

Institutions differ in many respects as to the organization and development of courses. In this study, this complexity on the institutional level has surfaced both in conventional (bivariate and multivariate) statistical tools as in more advanced models. Since the number of potential institutional variables exceeds the number of institutes, no reliable conclusions can be drawn from their results. That does not mean that institutional factors do not play a role in predicting

course outcomes at all. They do, but at a different level. There simply is too much difference in teaching methods across universities to conclude reliably that there is one main effect of institutional factors on 'achievement', as was tested in the latent change approach. In fact, institutional propensity related covariates deteriorated the latent change models. One possible explanation for this is of a methodological nature, i.e. the variables have been entered at an institutional level with no missing values, nor any imbalance. Hence, group membership based on institutional indicators is not biased. However, the different level at which these constructs were measured, makes the interpretation inadequate. A possible alternative approach was to perform multilevel analyses, but due to the small institutional sample size, this solution was abandoned.

I have sought two alternative approaches to the aforementioned problem. The first one was to only indicatively show institutional differences in the quantitative analysis, for instance by testing the differences between universities and colleges (as a group) or between Dutch and Flemish universities. Alternatively it is possible to sketch the results in a more qualitative manner as described in chapter 4 of this thesis. The conclusions in this section reflect upon both quantitative and qualitative results.

9.4.1 Differences across institutions

Course duration

The longer a course lasts, the higher student achievement becomes. The relation is positive. The 'course duration' also proved to be an important factor in the combined propensity related scores. Although it was decided not to interpret the institutional propensity related covariates for reasons mentioned earlier, results from separate analyses show that course duration positively affects students' attitude and achievement. This finding is confirmed by Budé (2007).

Learning Goals

According to most teachers learning goals have not entered the university curricula yet. On the other hand in some departments learning goals already are validated. Because of this diversity (and the small institutional sample) it is difficult to determine their influence on attitudes and student achievement. Furthermore, if there are explicit sets of learning goals, they differ a great deal across institutions. Although learning goals are not assumed to contribute a great deal to student achievement, they are believed to add to the clarity of the course. Knowing what is expected of you as a student, will help you prepare for an exam and will therefore affect the grade. This refers to the 'informed' expectations (Svanum & Bigatti, 2006) that a student can develop. It is therefore recommended that as in other types of (higher) education learning objectives become a common part of the course manual; future studies then can determine their influence on course outcomes.

Teaching methods

Contrary to my expectations, teaching methods show a diffuse result when it

comes to the effect on attitudes and student outcomes. The most commonly used method still is the lecture, mainly due to the fact that large student groups are not split up into smaller groups because of budgetary reasons. The results show that lectures do not influence the models on course outcomes. Moreover, institutions who organize lectures did not show lower attitudes at all. A possible explanation could be the 'teaching quality'. This variable was not measured in this study.

The only method that does seem to affect course outcomes is the frequency and duration of small work groups. This factor contributed to the propensity related covariates (PRC). However, the institutional PRC, as we saw, did not improve the attitude model at all, due to the diversity across institutions. A small number of students indicated having done applied student projects. They report very positively about these projects, however the number of indicated projects was too small to yield any (statistical) effects on the model. Some colleges prepare students for these projects by teaching a mix between methodology and statistics (as other institutions only teach statistics). In sum, to answer the question whether teaching methods really affect student achievement, additional research is needed. There is not enough evidence to support this claim in this study.

Assessment methods

Again a wide range of assessment methods was reported, both by the interviewed teachers and in the teachers' questionnaire. The most commonly used method still is the multiple choice exam. It becomes clear from the qualitative report that the multiple choice exam keeps the teachers' workload down. Another influential factor is the assessment of 'active participation'. Although only a small student group reported on this, the influence on the PRC is significant. Both influential factors reported negative loadings, and again this paints a diffuse picture. Differences across institutions and the apparent low construct validity added to the confusion, making it difficult to draw any reliable conclusions. However the qualitative reports clearly show a preference (by the teachers) for the multiple choice exam, although this does not mean to say that this reflects students' preferences. Final grade is a diverse mix of weighting and adding, across institutions. It is used as a dependent variable, because it comes closest to measuring objective learning outcomes. In this study a few institutions refused to reveal the students' final grades although students themselves gave permission. The main reason for not doing this is of an ethical nature. Additionally, a more subjective approach was taken, by measuring student expectations.

Class size, does it really matter?

Group size is considered an instrumental factor when it comes to attitudes toward statistics and course outcomes. However, it is important in a different setting than I had expected. First of all interaction plays a role. The larger the group, the less interaction between teacher and students is reported. However, this does not mean that attitudes become more negative or outcomes become

lower. As was concluded earlier, attitudes might even be more positive, as was found true for Affect and Cognitive Competency. It was also found that the larger the class size, the lower self-report Effort is. In this study (in line with findings from Gilbert (1995)), class size did not turn out to be very influential, but few aspects have to be kept in mind:

- class size only starts to play an influential role when the groups are smaller than 15.
- previous results are not really consistent (only meta analyses were performed; emphasis on outlying results only; mostly done in K-12 settings (Hattie, 2005).
- one university in this study starts with lecturing to large group of 400 but then splits up into 40 groups of 10.
- the effect of class size could be suppressed by teaching quality. I.e. if the teacher is good, class size really does not matter.

In this study I found indications for these findings, but no solid proof. In sum, institutional factors are too diverse in this study to show a consistent and reliable contribution to students' attitudes, their motivation, expectancies and therefore their final grades. Part of this result is due to the nature of this field experiment and the instruments used (questionnaires). It would be better to set up a longitudinal experiment for the testing of teaching- and assessment methods and compare the results. There is another explanation for the lack of consistency in institutional results: institutional differences not so much contribute to student achievement, but the background of the student is reflected in the choice for a particular institution, hence student level and a student's own input rather than teaching and 'assessment methods (Astin, 2003). This contextual approach would mean that institutional differences rather reflect upon students' characteristics than to differences in institutional settings. This has already been discussed in the section on 'individual factors'.

9.4.2 Looking back at the comparison across systems

To what extent do the models differ across institutions? Is there a difference between Flemish and Dutch colleges and universities, between universities and colleges per se? The results tend to show that differences across institutions influence the course outcomes but they do not directly affect the students' attitudes. Moreover, institutional differences are more prominent between universities and colleges than between Flanders and the Netherlands.

Colleges and universities - differences and similarities
It has not been proved consistently in this study that institutional factors affect student outcomes. As Astin (2003) explained, a possible reason for the absence of a clear result could be that individual differences between students are already

reflected in institutions, because of the existence of a 'selection effect'. It is as if every university or college has it's own group of students with a special set of characteristics. At certain colleges this is a result of 'selection at the gate' but usually certain students choose certain universities. So context does matter, but more indirectly.

One result that did become apparent is that university and college students differ with respect to Cognitive Competency and Effort. College students show a higher Cognitive Competency than university students at the start of the semester, but university students experience a larger positive change. On the other hand, college students report (in hindsight) to have put in more Effort than university students. Most pertinent differences between colleges and universities are reflected in group size, type of interaction and assessment methods. Another aspect of the difference is the 'student life' as some colleges have a residential setting. Although group size differs, this is not the main difference. According to the coordinators, it is the 'student orientation', interaction and personal approach at colleges that makes the difference. Although the most commonly used teaching method is still the lecture, the approach at colleges is more interactive. Assessment methods differ with respect to weighting, number and method as most colleges use some kind of continuous assessment and they use various methods of assessment. Some universities still only have one exam at the end.

In sum, much like the example of the Anglo-Saxon colleges, the main characteristics that makes colleges different from university settings are the residential setting, the size of the college, the selectivity and the student orientation (by means of didactical approaches, but also by means of the tutoring system). Future research should establish to what extent these contextual variables are the key to success i.e. student achievement. Results of similar studies in the USA (Astin, 2003) showed that these variables play a strong and consistent role in that success. This explanation however needs to be handled with care. After all, US and Dutch colleges differ much more than they are alike (Bruinsma, 2003).

The Netherlands and Flanders - different systems or both in transition?
Both Dutch and Flemish universities have to get used to the bachelor/master structure, as indicated by the vast number of pilots, trial courses, new set ups et cetera. Although it seems Europe wants to adapt to the Anglo-Saxon system, there are still more differences than similarities (Bruinsma, 2003). The closest to a comparison with the new Bama-structure we get is with the three Liberal Arts and Sciences colleges in this study, for they hold similar characteristics. However no solid proof as to benefits for student outcomes has been found college-wide.

The Netherlands and Flanders encounter the same type of problems with respect to the transition to the new system, although differences in academic culture in the two countries require different solutions. However, Flemish and Dutch universities also have been collaborating for many years; many projects have been jointly set up (accreditation, educational programs, exchange). Better accommodation to research- and educational objectives will strengthen the

position of the Dutch and Flemish institutions.

One clear difference was shown with regard to Effort, where Flemish students really relate this to expectations. Flemish students report higher Effort scores than Dutch students, whereas Dutch students report higher attitudes on all other components except for Interest. During the qualitative interviews and observations at the start of the semester, it was also noticed that in Flanders the teacher-student interaction is much more formal than in the Netherlands. To what extent this affects course outcomes has not been tested. Another clear difference can be seen from assessment: at Flemish universities often only one final exam is organized.

9.5 To what extent does the Expectancy Value Theory hold?

The application of the EV Model for this study can explain achievement related choices by means of individual factors, expectancies, and effort; for the most part it holds for Dutch and Flemish data. In short this study contributes to the following findings:

- Individual factors contribute to Attitudes.
- Individual factors affect final grade directly and indirectly through expected grade.
- Institutional factors are too diverse to give a clear picture of the effect on student achievement. There are some indications that the level of interaction, course duration and class size make a difference, but further research is needed.
- Effort mediates the effect of Attitudes on achieved grade. It is also linked to 'number of hours studied'.
- In general, attitudes contribute to student achievement, but the effects are not very big.
- Number of hours indirectly affects final grade through of expected grade.

Is Effort an indicator or a mediator?
The special position of Effort in this study has been made visible in several analyses. In conclusion Effort mediates the effect of the other attitudes on student achievement and therefore maintains this special position. In the separate analysis the special position holds, since Effort refers more to 'active learning *behavior*' rather than the other 5 attitude components that refer to *beliefs* rather than behavior.

9.6 Methodological quality of this study

Due to some limitations in the setup of this study, the results cannot be generalized to the entire Dutch / Flemish student population. However, this was not the objective of this study. In this section aspects of validity and reliability are discussed. I will start with a discussion on the added value of LCMEM and the use of Propensity Related Methods.

9.6.1 Added value of LCMEM

Is the LCMEM a good statistical tool to analyze changes and predict course outcomes? In answer to subquestion 6c from section 3.2.1, it can be concluded that Latent Change Method Effect Models have proved to be powerful statistical tools to model the change of attitudes across a semester, the effect on student achievement and the influence of individual factors. Adding institutional factors did not have the expected result. On the contrary, institutional differences are not easily incorporated in latent change models.

The set up of this study only provided data for two measurement periods, one at the start and one at the end of an Introductory Statistics course. This test-retest design is suitable for testing true-intraindividual change and individual method effects (Vautier, Steyer & Boomsma, 2007; see also Steyer, Eid & Schwenkmezger, 1997; Pohl & Steyer, 2006b; Pohl et al., 2008)). As indicated below, the true change is the change across the posttest (T_{2-1}) and the pretest (T_1) measurement, reduced with the method effect:

$$Change = E(T_{2-1}) - E(T_1) - E(m) \tag{9.1}$$

For the purpose of this study, the method for testing latent attitude change, the effect on outcomes and the influence of individual factors, was split up into four levels of analysis: *engine*, *change*, *covariates* and *achievement*. This set up has added to the clarity of the interpretation. Not only does it become possible to split up the more technical set up of the model from the true changes, also the fit of models at different levels can be compared. This has led to results that were split up in a 'technical fit' of the model and a 'substantive fit'.

Method effects

Besides analyzing true-intraindividual change it is important to detect any changes that are caused by other than the treatment effects. Changes might be caused by differences in the set up of any field experiment instead of true-intraindividual change. Students might also have reacted differently to the questions at the two different measurement times, especially since the same questionnaire was used (Vautier, Steyer & Boomsma, 2007). When modeling latent attitude changes a method effect was determined: after controlling for

this method effect (filtering it out), the correlations across measurement periods increased showing true intraindividual change. This points out that possible differences in the set up of this field experiment influence the attitude change rather than the intra-individual change itself. A number of situations could have played a role here. Firstly, the pretest was administered by the researcher and the posttest by the 'own' teacher. Secondly the pretest was administered at the first lecture, the posttest however right after the exam. A learning moment here is that students are usually not that motivated to fill in questionnaires after their exam and this might have biased the results. Thirdly unforeseen circumstances beyond our control might have played a role, such as set up of the rooms, time of day, noise, et cetera.

LCMEM's added value to more conventional statistical tools
It has been shown that the use of 'advanced models' over more conventional tools, such as bi- and multivariate comparisons, certainly has an added value. The significance of the MF-variance in all cases showed the presence of such a method factor, and as a result of filtering out this method effect true change by means of increased pretest-posttest correlations became visible. Adding a latent method factor has shown that a more stable interpretation of individual causal effects was possible (Steyer et al., 1997).

A strong argument to choose for latent change models over bi- and multivariate comparisons is the nature of the simultaneous testing of several types of parameters, both manifest and latent factors. There is another important reason for preferring advanced models over the aforementioned conventional tools. When analyzing data from complete datasets using a randomized design, reliable and stable information about changes in attitudes can be given. However, in this study, as in many others, the design was not randomized, let alone complete. This justifies the use of more advanced models, the advantage being that true intra-individual changes can be determined from average changes. Most importantly, most changes in attitude hold, notwithstanding the method effect. This confirms once more that the SATS©tool is a robust method to analyze attitude changes and the mediating effect of those changes from individual factors onto course outcomes. This robustness is determined both in a statistical and a substantial way.

9.6.2 Added value of the Propensity Related Method

Another question to be answered is what the added value is of using a propensity related method in order to analyze the effect of combined individual and institutional covariates on course outcomes. After all, interpreting these combined factors has proved to be difficult and diffuse. For this study the method has been extended in two ways. First, the propensity related method is used to combine a large number of individual (and institutional) indicators into one nicely 'bundled' effect on the LCMEM. This has a clear parsimonious advantage. By adding the PRC's it was shown that the combined individual factors have a larger effect on the attitude change than institutional factors. Addi-

tional multivariate analyses confirmed these findings, albeit without revealing a method effect (that can only be done with LCMEM). The use of propensity related analysis has the additional advantage that it balances out possible group differences in a nonrandomized setting with missing values. That means that the attitudinal change that is observed can be truly attributed to intraindividual change, and that these changes are not the side effect of group differences (Pasta, n.d.; Rosenbaum & Rubin, 1983; Rubin, 1997).

In this study no treatment and control group were available and a pretest-posttest design was used. Since some students participate in only the pretest or the posttest, differences between the pretest and the posttest may be the result of (self) selection, which was treated by analyzing the partially overlapping pretest and posttest student groups. This choice of groups is of special interest for the propensity score method. The original method implies that the propensity score is the conditional probability of being assigned to a particular independent treatment group (Rosenbaum & Rubin, 1983). In this study the choice was based on the differences between pretest and posttest groups, due to students only taking the pretest or the posttest. There was enough overlap but also enough difference between pretest and posttest groups to justify their use. Expectations are that replication of this method will lead to further improvement of this application and a reliable result. Until this is the case, the results are to be treated with some caution.

Interpreting the relative contribution of the canonical structure coefficients and in return their contribution to the propensity score has proved to be diffuse and complex. In order to partly solve this problem, separate contributions of individual factors to course outcomes were tested using a structural regression model. In order to establish true representation of the LCMEM in PRC's, the factor score weights from these LCMEM's were used to calculate the so-called 'change indicators' in the advanced regression model. In that sense, the structural regression model was not so much an alternative for the LCMEM, but an addition to the LCMEM, where the latent change models served as a starting point.

The LCMEM results have shown that the **individual** propensity related covariate most contributes to the fit of the models (Verhoeven, 2008), but that mean changes do not differ substantially as a result of adding those covariates. This is probably caused by the fact that the loadings of the Discriminant Function were not very strong and therefore the influence is rather weak. Moreover, separate contributions of indicators, if any, are also only small. These separate effects of the individual background variables were found in age, gender, self-confidence, and the perception of math-skills. They are in line with my expectations. The biggest contribution to the model comes from self-confidence on attitude change and (expected) grade. However, mathematics experience contributes more to expectations than to attitude change.

Simple bivariate or advanced tools?

At the elementary level, it will always be important to report straightforward difference-scores and correlations, as we are interested in attitude differences

across gender, institution and time. However, if the development and success of teaching statistics depends on these results it is equally important to know what part of the attitudes can really be attributed to true score, and what part to the effect of a design, what part is caused by a selection effect from a certain background of students and what part can be influenced by changing the organization and didactics in the course. Then, advanced methods will give more stable and robust results.

9.6.3 Reliability

The sample size was large enough. The use of high quality and thoroughly validated instruments increases the replicability of this study and therefore enhances the reliability of its outcomes. Usage of research logs, meticulously recording the steps in the analysis, having good information / introduction to students also enhanced the quality of this study. Lastly, triangulation was used in order to answer the central question; this is considered to further increase the reliability of the outcomes of this study.

9.6.4 Validity aspects of this study

External Validity

The external validity, especially the population validity is not high for this study, because nonprobability samples such as snowball and later convenience sampling were used to select the institutions and participants. Moreover, I included all students in the sample once I received permission to collect data. However, generalizing the results to a student population was not the objective of this study. The construct validity of the SATS©-instrument is good, indicated by the comparison with test results by Tempelaar (2007a), Schau (2003, 2005; see also Dauphinee et al., 1997). The measure of attitudes was consistent with the measures taken in the United States. It can therefore be concluded that this tool is suitable for usage as an evaluation instrument at colleges and universities in the Netherlands and Flanders. The construct validity of the teachers' questions regarding teaching methods is low. Teachers gave diverse answers with regard to teaching methods, although I know that their course setup is similar to other institutions under study. Apparently the definition of 'small work groups' has not been made very clear.

Internal Validity

There are some threats to the internal validity. First of all, there is a selection threat. The data have not been sampled randomly, participation was voluntary. This might have caused a special group of students to participate in this study. Students could have been especially interested in this type of experiment or, on the other hand, they could have chosen to complain about the mandatory and complex character of this course. Students who are good at statistics might have more often chosen to participate in the posttest as opposed to students who do not want to be reminded of their low grade. Furthermore, institutional

systems may have been incorporated in student characteristics. One of the remedies sought was to control for possible differences by use of propensity related covariates. Secondly a history threat might be present, due to changes in the curriculum during the semester. I was not made aware of any possible changes. Instrumentation seems under control, as in both occasions, the same questionnaire was used. As was pointed out earlier, the internal validity is under threat because a method effect was detected.

9.7 Discussion: Predicting student outcomes regarding Statistics

Most changes in attitudes toward statistics are small and negative. Why? Does taking introductory statistics result in a more negative attitude toward statistics? And if this is the case, what can teachers, developers, but also students do to make it better? If we look beyond attitudes and also take into account students' characteristics, we see a specific group of students, social science majors that do a mandatory course. The mandatory nature of the course in itself is a reason for negative attitudes. However, becoming acquainted with the statistical topics should at least help. Well, the picture is not that pessimistic. Attitudes on average are not that negative and students do believe they become more skilled over the semester. Furthermore, the effort they put in actually seems less than expected. Part of their anxiety at the start of the course seems to result from being uninformed. Students still find statistics difficult, even more than when they started. The fact that they do not value statistics highly and that their interest goes down, does not surprise me. A follow up study, asking the same questions at the start of a student's career or when he is writing his thesis could show a different picture. Hence, attitudes towards statistics, its usefulness and future scope could be perceived more positively when a career is within reach. Meanwhile, can a teacher/developer make a course better suitable, feasible, to the needs and wants of the students?

Statistical literacy, thinking or reasoning

In the introduction of this thesis I expected that 'statistical reasoning' would be the main focus of the Dutch and Flemish Statistics teachers. Given the teaching methods that are currently used, statistical reasoning skills seem to be the main objective in most institutions, as Statistics teachers do not only want students to recognize, understand and discuss statistics (Snee, 1990), they also want students to develop a 'helicopter view', to critically discuss topics of a statistical nature and to be able to apply the knowledge to other fields of study (Garfield, 2002; Garfield & Gal, 1999). One way to do this is to organize statistics projects in small groups and have students experience what it is to actually set up and conduct a research project and report about the results to real clients. Not only the real-life experience but also working in small groups serves the purpose of learning to apply statistical reasoning.

Additionally, the focus of teachers should not only be on students' development of statistical reasoning, but in a broader sense on statistical competencies. This means that on top of making sense of statistical information in the way Garfield (2002) puts it, the actual possibility and competency of applying statistical knowledge in a broad field of topics adds to this objective. This changes statistics education from a teacher-focus to a student-focus, i.e. a student-oriented approach (Tempelaar & Nijhuis, 2007c), even more than currently is the case.

9.7.1 What can teachers and institutions do?

It was not consistently shown that institutional factors play the same role, because they differ so much across universities. However institutions can still benefit from a number of changes in teaching methods, assessment and other institutional factors, as there are indications that individual characteristics are to some extent reflected in institutional factors. After all, students partly select institutions based on educational characteristics.

More statistics in high school curricula
There are signs that more mathematics education in high school (and more statistics in the curriculum) might help students overcome their stats anxiety. They have a more open mind when they start a statistics course because they know what to expect, they feel more competent and therefore they obtain a better grade.

Class size
As a teacher I would like to teach a class with only a handful of motivated students that are eager to learn and work in collaboration to complete this course (see also Garfield, 1993). However, reality is different. Class size would only be beneficial to student achievement if it were really small, smaller than 15 students. Maintaining group size as small as possible is still recommended for better interaction, motivation and communication. The quality of teaching should also be taken into account, in spite of group size. I therefore emphasize that keeping track of teaching quality is important. In this respect the SATS© questionnaire can be used as an evaluation tool. As sure as I am that class size is beneficial to both students and teachers, Illig (1996) points out that there is a financial aspect to take into account that withholds schools from the necessary reduction.

How instructors should interact with students.
Interaction and motivation are the keywords in Statistics teaching. The teacher and the student should interact as much as possible on several levels, by means of lectures, discussions, group work, individual appointments, class exercises, presentations, et cetera. Methods like this could improve stats education, although a lot is already undertaken to improve the quality of teaching stats. Use interactive teaching methods, questions and answers during class meetings, let students discuss topics, ask and answer each others questions, make and dis-

cuss class exercises. As Lee (1996) puts it: *P*rojects, *A*ctivities, *C*lass-exercises, and *E*xercises (PACE, see Lee, 1996). Projects in this sense are self-developed projects by student groups. Activities are small in-class group activities and team-work. Hence, a student-oriented approach is highly recommended.

Equally important is motivation. How can a teacher motivate the students? According to Sowey (1995) teachers can do a lot by providing insight in the statistical structure of the subject, show them in what perspective students can place the statistical subject, motivate students by encouraging them to uncover statistical problems, work with real-life data and striking examples, demonstrate the usefulness of statistical tools. Additionally, acknowledgment of different approaches to teaching and learning (Prosser & Trigwell, 1999) could play an important role. Different teaching contexts evoke different approaches to learning. In order to enhance the quality of teaching and thereby learning, the teacher might adapt the appropriate approach for students to effectively learn statistics. Lastly, teachers should inform students at the start of the semester of the expectations, by means of a good course manual. Being informed about the expectations will take away some of the uncertainty (and hence 'anxiety') and can be the basis for a *'well-informed expectation'* of course outcomes and, as a result, of a well-aimed work ethic (Svanum & Bigatti, 2006).

How can we change their attitudes?
'Stats anxiety' is experienced almost without exception, so this question is difficult to answer. *Continuous assessment* makes students aware of the fact that not only the final exam counts, but it also includes class exercises, homework, presentations, active participation, (group) paper. This will ensure that the students' anxiety does not merely concentrate on the final exam, but is eased and spread out (see also Garfield, 1994). An example is the use of '*real-life projects*' that teach students how to conduct real-life research and make them acquainted with all possibilities and constraints of the research process. It helps them understand this process and the technique behind it. The material becomes alive. If classes are fairly large, usage of cooperative learning techniques (such as group work during class meetings) also enhance the attitudes of students (Magel, 1998).

What can students do themselves?
First of all, being uninformed does not help much. Reading course manuals and asking the teacher about the course-expectations would already help a great deal to develop *well-informed expectations*. Furthermore, students should help each other. *Peer assessment methods* actively involve students in their own assessment, it is a form of collaborative learning. It helps students get involved into each others but foremost their own learning (Joosten-ten Brinke, Verhoeven & Van Buuren, 2003; Dochy, Segers & Sluijsmans, 1999). For example, if they make mistakes they would rather hear that from each other than from the teacher. During student projects, the group work should be evaluated separately for this purpose. They should not take all the information for granted and be critical and, moreover, not be afraid to give the wrong answer for once! Lastly,

with the help of their instructor, students should develop statistical skills that are based on critical insight, not on recollection, and thus a helicopter view.

9.7.2 Recommendations for future research

The vast number of research questions that result from this project cannot be left unnoticed and a number of recommendations are made. After all, good empirical research raises more questions than it answers.

'Effort'
A first question that derives from these research results is into the nature and position of Effort in the model of student outcomes. Not only should this position be analyzed, but a closer look at the operationalization of Effort should be taken. Effort could be split up into a 'deep' and a 'surface' learning approach, as has been done by Tempelaar (2007b), but also an operationalization in terms of 'academic work ethic' might be considered. Lastly the special place Effort takes in the SATS© can be a topic of future research.

Mathematics experience
A cross national study into the effects of high school mathematics on student learning at universities may contribute to the understanding of statistics achievement at universities. It is my opinion that the analytical approach that usually constitutes math education helps students develop the 'helicopter view' so needed to plan and conduct research and draw the right conclusions.

Assessing teaching quality
One important factor has been left out of this study: the evaluation of teaching quality. Teaching methods, group size, course duration all have a diffuse effect on students' achievement, but the teaching of the instructor is assumed to determine attitudes, student motivation and therefore student achievement. After all, teaching is considered to be a complex, contextual and individual process (Nijveldt, Brekelmans, Beijaard, Wubbels & Verloop, 2008). Unfortunately no information about the evaluations of teaching quality was available for this study, mostly for ethical reasons. It reflects the insider-outsider bias mentioned in chapter 3. As a teacher and phd-graduate you research your own topic and that of your colleagues throughout the country and it is sometimes considered threatening to show one's evaluations. However, it is assumed that the evaluation of the teacher is taken into consideration when a student perceives his own attitudes, motivation, expectations. When a teacher really is good, this subordinates all effects of group size, methods, assessment and objectives. As a student, I have seen full lecture halls with 600 students silent and focused, listening to one of our most popular professors. I have also witnessed small groups that experienced unstructured and confusing lectures resulting in bad motivation and, hence, bad grades, both for students and teacher. Future research needs to look at the effect of teaching quality (see for instance Nijveldt et al., 2008), as it is expected that teaching quality plays an important role in student

motivation to learn and (in)directly in student achievement. Additionally, my interest lies in the question to what extent a student-oriented approach affects student achievement.

Future design - methodological aspects
A future study about the effects on student achievement should have a *longitudinal* set-up with multiple measurements in time, so as to make sure the development over time is evaluated well. Expectations are that when questioned at a later point in time, a students' attitude will reflect much more the true value, interest and affect for statistics. Second, the use of Propensity Score Methodology to correct for differences due to (self-) selection should be further studied, as the use of the PRC's as a 'bundled' effect still is not widely spread and tested. The same recommendation can be made for the choice of groups in a paired setting. As this new application is still handled with caution, future tests should improve the reliability and validity of this tool. Furthermore, *random assignment* should more often be part of the set up, with less 'noise' to keep down possible method effects. Besides, an experiment as to the long term effects of class size on learning and student achievement could shed some light on the effects on student achievement. Lastly, more institutions could be added to the sample, to enable a *multilevel* approach to the analysis. By using a multilevel approach the nested nature of this type of data can be truly acknowledged.

9.8 Closing remarks

A mild case of 'insider-outsider bias' and its remedy
As a researcher/teacher the experience of an 'insider bias' forces itself upon me. This type of bias as Merton (1972) describes it, is caused by the fact that if one's own institution is part of the project, the researcher holds his own 'ascribed status', together with a mindset, ideas and concepts (see chapter 3). I could not take a more objective (outsider) view, just because I already possess this 'ascribed status'. It would not be possible to deny based on the results of what I teach. A teacher has to 'practice what he preaches'. However, I do find Merton's standpoint a bit too strict. I my view it is possible to keep an open mind to the information you receive from other universities and perhaps use that in your own teaching practice. After all we are 'life long learners' besides 'teachers' and 'researchers'.

During this phd-project, I tried to minimize this type of bias by means of a number of measures. First of all I chose a triangulated approach, and looked at the research question from different angles. Secondly, for this project I did not use my own instruments, but a validated and reliable questionnaire that has been widely used at universities in an international setting for many years already. Thirdly, interaction across researchers was achieved by the exchange of information about the results of this study at conferences, research days, papers and by personal consultation. In this way a high degree of intersubjectivity can be accomplished. Lastly, I did learn a great deal from the experience of

other researchers/teachers throughout the Netherlands and Belgium. For one, I learned about the great diversity of teaching methods. The main threat to this type of bias is, in my view, the fact that data could not be collected randomly, therefore violating the assumption of 'objectivity' a great deal. In future research, a similar set up might be accomplished by a multilevel sampling method or a clustered sample.

This phd-thesis contains a vast amount of results and conclusions. Let me summarize this concisely and then add a personal touch. Average attitudes toward statistics are located in the upper half of the attitude scale, and they do change over the course of the semester, albeit not only for the better. Perceived cognitive competencies change positively, whereas changes on values, interest, difficulty and affect do not. Merely taking an Introductory statistics course apparently does not help improve attitudes. Reports on Effort decrease, indicating that students, in hindsight, spent less effort than they expected. These changes are affected by individual characteristics, such as self-confidence, age, mathematics experience and the effect these characteristics have on attitudes differs across gender. Attitudes also relate to the number of hours a student wants to study (Effort) and actually studies and, as a result, expectations and final student outcomes are influenced. Effort both has an indirect and a direct effect on student outcomes. Lastly institutions play a role in the prediction of student achievement, but in a different way than I expected. Institutional characteristics are too diverse to generalize to one main conclusion, but certainly didactical and assessment methods, class size and duration play a role in modeling student achievement. They are reflected to some extent in student population, as some selection effect is present. Generally speaking teachers should motivate and stimulate their students so that they become more self confident and positive about their capabilities and the level of information from the institutions should give students clarity as to what is expected of them.

I have often wondered what suits me better, being a teacher or being a researcher. Actually the two are interchangeable as far as I am concerned. However, when I reread the results from this study, the teacher in me responds. The main aspect of *teaching* statistics irrespective of time and place is that you are highly motivated and that you can motivate your students. If you can accomplish that, student achievement will be high, even (or should I say *especially*) in statistics. I wish you all the best in trying!

Pieternel Verhoeven
Middelburg, January 2009

Reference list

Ahlgren, A. (2001). *A Coherent Fabric of Statistics Literacy*. ICOTS6, Cape Town.

Aliage, M., Cobb, G., Cuff C., Garfield, J., Gould R., Lock, R., Moore T., Rossman A., Stephenson B., Utts J., Velleman P., & Witmer J. (2005). *GAISE College Report*, American Statistical Association.

AmStat News (2001). *Curriculum Guidelines for Undergraduate Programs in Statistical Science*. Retrieved 28th June, 2005, from www.amstat.org/education.

Arbuckle, J. L. (1996). Full information estimation in the presence of missing data. In G.A. Marcoulides & R.E. Schumacker (eds). *Advanced structural equation modeling*. Mahwah, NJ: Erlbaum.

Arbuckle, J. L. (2007). *AMOS 16.0 User's Guide*. Chicago: SPSS.

ARTIST (2002a) *Assessment Resource Tools for Improving Statistical Thinking*. Retrieved 12th June, 2005, from www.data.gen.umn.edu/artist/about.html.

ARTIST (2002b) *Evaluating the Impact of Educational Reform in Statistics: A Survey of Introductory Statistics Courses*.

Astin, A. (2003). Studying how college affects students. *About Campus*, 8(3).

Bakker, A. (2004). *Design research in statistics education; on symbolizing and computertools*. Utrecht: Freudenthal Institute. Phd-thesis.

Baloglu, M. (2003). Individual differences in statistics anxiety among college students. *Personality and Individual Differences*, 34, 855-865.

Bloom, B.S. (Ed.)(1956). Taxonomy of educational objectives: The classifi cation of educational goals. In *Handbook I, cognitive domain*. New York, Toronto: Longmans, Green.

Bradburn, N.M., & Sudman, S. (1988). *Polls and surveys: Understanding what they tell us*. San Francisco: Jossey-Bass.

Brandt, R. M. (1958). The accuracy of self estimates. *Genetic Psychology Monographs*, 58, 55-99.

Breslow, R.M. (2005). A Comparison of Academic Performance of Off-Campus Nontraditional PharmD Students With Campus-Based PharmD Students. *American journal of Pharmaceutical Education*.

Bressoux, P., Kramarz, F., & Prost, C. (n.d.). *Teachers' Training, Class Size and Students' Outcomes: Learning from Administrative Forecasting Mistakes*. Universit Grenoble.

Brophy, J. E. (1988). Educating teachers about managing classrooms and students. *Teaching and Teacher Education*, 4 (1).

Bruinsma M. (2003). *Effectiveness of higher education. Factors that determine outcomes of university education.* Rijksuniversiteit Groningen/ Pedagogie. Phd-thesis.

Bryce, G. (2005). Developing Tomorrow's Statisticians. *Journal of Statistics Education* 13(1).

Budé, L. (2007). *On the improvement of students' conceptual understanding in statistics education.* University Maastricht (dissertation), Maastricht.

Campbell, D.T. & Fiske, D.W. (1959), Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56 (2): 81 - 105.

Carter, R.L. (2006). Solutions for MIssing Data in Structural Equation Modeling. *Research & Practice in Assessment*, 1 (1).

Chance, B. (1997). Experiences with Authentic Assessment Techniques in an Introductory Statistics Course. *Journal of Statistics Education*, 5(3).

Chance, B., delMas, R., & Rossman, A. (2004). *Designing and Evaluating Assessments*
for Introductory Statistics: ARTIST, University of Minnesota.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation. Design & Analysis Issues for Field Settings.*. Boston: Houghton Mifflin Company.

Cherian, V. I. & Glencross, M.J. (1997). Sex, Socioeconomic status, and attitude toward applied statistics among postgraduate education students. *Psychological reports.*

Cobb, G. (1992). Teaching statistics. In: Lynn, A. Steen (Ed.) *Heeding the call for change: Suggestions for curricular action* (Notes no. 22), 3-43.

Cobb, G. W. & Moore, D.S. (1997). Mathematics, Statistics and Teaching. *The American Mathematical Monthly* 104(9): 801-823.

Coleman, J.S. (1961). *The adolescent society.* New York: The Free Press.

Coleman, J.S. (1966). *Equality of Educational Opportunity.* U.S. Government Printing Office.

Collins, R. & Makowsky, M. (1993) *The discovery of society.* McGraw-Hill Inc.

Converse, J. M., & Presser, S. (1989). *Survey questions: Handcrafting the standardized questionnaire.* Newbury Park, CA: Sage.

Cribbie, R. A., & Jamieson, J. (2004). Decreases in Posttest Variance and The Measurement of Change. *Methods of Psychological Research Online*, 9(1), 37-55.

D'Agostino, R. B., & Rubin, D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Organization*, 95(451), 749-759.

Dauphinee, T., Schau, C., & Stevens, J. (1997). Survey of Attitudes Toward Statistics: Factor structure and factorial invariance for females and males. *Structural Equation Modeling*, 4, 129-141.

Dehejia, R., & Wahba, S. (1997) Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs. In R. Dehejia, (1997) *Econometric Methods for Program Evaluation.* Ph.D. Dissertation, Harvard University, Chapter 1.

Dehejia, R., & Wahba, S. (2002) Propensity Score-matching methods for non-experimental causal studies. The Review of Economics and Statistics, 84 (1), 151-161.

Deinum, J. F. (2000). *Schoolbeleid, instructie en leerresultaten.* University Groningen.

Den Brok, P., Brekelmans, M. & Wubbels, Th. (2004). Interpersonal Teacher Behaviour and Student Outcomes. *School Effectiveness and School Improvement*, 15, 3-4, 407-442.

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology*, 62, 641-644.

Dillon, M., & Kokkelenberg, E. C. (2002). *The Effects of Class Size on Student Achievement in Higher Education: Applying an Earnings Function*, 42nd AIR Forum. Toronto, Canada.

Diamond, N. T. & Sztendur, E.M. (2002). *Simplifiying consulting problems for use in introduction statistics lectures.* Cape Town, ICOTS6.

Driessen, G., Dekkers, H. (1997). Educational opportunities in the Netherlands: Policy, Students' performance and issues. *International Review of Education*, 43(4), 299 - 315.

Dochy, F., Segers, M., & Sluijsmans, D.M.A. (1999). The use of self-, peer-, and co-assessment in higher education: a review. *Studies in Higher Education*, 24(3), 331- 350.

Duchesne, I & Nonneman, W. (1998). The Demand for Higher Education in Belgium. *Economics of Education Review*, 17(2) 211 - 218.

Eid, M. (1997). Happiness and Satisfaction: An Application of a Latent State-Trait Model for Ordinal Variables. In J. Rost & R. Longeheine (Eds.), *Applications of latent Trait and Latent Class Models in the Social Sciences* (148 - 154): Waxman.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating Trait Effects From Trait-Specific Method Effects in Multitrait-Multimethod Models: A Multiple-Indicator CT-C(M-1) Model. *Psychological Methods*, 8(1), 38 - 60.

Gal, I. & Garfield, J.B. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal, & J.B. Garfield (Eds.), *The Assessment Challenge in Statistics Education*(pp. 1 - 13). Amsterdam, Berlin, Oxford, Tokyo, Washington DC, The International Statistical Institute, The Netherlands.

Garfield, J. (1993). Teaching Statistics Using Small-Group Cooperative Learning. *Journal of Statistics Education*, 1(1).

Garfield, J. (1994). Beyond Testing and Grading: Using Assessment To Improve Student Learning. *Journal of Statistics Education*, 2(1).

Garfield, J. (2002). The Challenge of Developing Statistical Reasoning. *Journal of Statistics Education*, 10(3).

Garfield, J. (2003). Assessing Statistical Reasoning. *Statistics Education Research Journal*, 2(1), 22-38.

Garfield, J. & I. Gal (1999). Teaching and Assessing Statistical Reasoning. In L. S., Reston (eds.), Developing Mathematical Reasoning in Grades K-12

(pp. 207-219). VA: National Council Teachers of Mathematics.
Garfield, J., B. Hogg, C. Schau & D. Wittinghill. (2002). First courses in Statistical Science: The Status of Education Reform Efforts. *Journal of Statistics Education*, 10(2).
Garson, G. D. (n.d.). *Data Imputation for Missing Values. Quantitative Methods in Public Administration*, 2007, from http://www2.chass.ncsu.edu/garson/pa765.
Gilbert, S. (1995). Quality Education: Does Class Size Matter? *Research File*, 1(1).
Glaser, B. J. & Strauss, A.L. (1967).*The discovery of Grounded Theory.* Aldine: Chicago.
Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, 8(4), 597 - 607.
Greene, B.A., & DeBacker, T.K. (2004). Gender and Orientations Toward the Future: Links to Motivation. *Educational Psychology Review*, 16(2), 91 - 120.
Grunwald, H.E., & Mayhew, M.J. (2008). Using Propensity Scores for Estimating Causal Effects: A Study in the Development of Moral Reasoning. *Research in Higher Education*, 49, 758-775.
Hall, M. (2008). Predicting Student Performance in Web-Based Distance Education Courses Based on Survey Instruments Measuring Personality Traits and Technical Skills. *Online Journal of Distance Learning Administration*, Volume XI, Number III, Fall 2008.
Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement /performance measures. *Review of Educational Research*, 52, 123-142.
Harris, M.B., & Schau C. (1999). Successful strategies for teaching statistics. In S.N. Davis, M. Crawford, & J. Sebrechts (eds.), *Coming on her own: Educational Success in Girls and Women* (pp. 193 - 220). San Fransisco: Jossey-Bass.
Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research,* 43, 387-425.
Hau, K., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical Statistical Psychology*, 57, 327-351.
Hilton, S.C., Schau, C.S. & Olsen, J.A. (2004). Survey of Attitudes Toward Statistics: Factor Structure Invariance by Gender and by Administration Time. *Structural Equation Modeling*, 11(1), 92-109.
Hirano, K., & Imbens, G.W. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services & Outcomes Research Methodology*, 2, 259 - 278.
Hoogveld, A. M. W. (2003). *The teacher as a designer of Competency-Based Education.* Heerlen: Open University.
Howley, C. B., & Bickel, R. (1999). *The Matthew project: National report.* Randolph, VT: Rural Challenge Policy Program.

Hoyle, R.H., Harris, M.J. & Judd, C.M. (2002). *Research Methods in Social Relations*. London: Wadsworth, Thomsonlearning.

Huang, I.(2003). *Application of a propensity score approach for risk adjustment in profiling mulitple physician groups on asthma care*. Paper presented at the Joint Statistical Meeting.

Imai, K. & Van Dyk, D.A. (2004). Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association*, September 2004.

Jong, M.J de (2007). *Icons of Sociology*. Amsterdam, Meppel: Boomonderwijs.

Joosten-ten Brinke, D., Verhoeven P.S. & Van Buuren, J.A. (2003). Peerassessment en selfassessment in een competentiegericht methodologiecurriculum. *Tijdschrift voor Hoger Onderwijs*, 21 (4), 273-286.

Kang, J.D.Y., & Schafer, J.L. (2007). Rejoinder: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population mean from Incomplete Data. *Statistical Science*, 22(4), 574 - 580.

Karagiorgi, Y. & Symeou, L. (2005) Translating Constructivism into Instructional Design: Potential and Limitations. *Educational Technology & Society*, 8(1) 17-27.

Kenny, D.A. & McCoach D. B., (2003) Effect of the Number of Variables on Measures of Fit in Structural Equation Modeling. *Structural Equation Modeling*, 10(3), 333-351.

Kline, R.B. (2005). *Principles and Practice of Structural Equation Modeling*. New York: The Guildford Press.

Laing, J., Swayer, R, & Noble, J. (1989). Accuracy of self-reported activities and accomplishments of college-bound seniors. *Journal of College Student Development*, 29, 362-368.

Lee, C. (1996). *Promoting Active Learning in Introductory Statistics Course Using the PACE Strategy*. Michigan, Central Michigan University.

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics*, 22(2), 329 - 349.

Lee, V. E. & Smith, J. B. (1997). High School Size: Which Works Best and for Whom? *Educational Evaluation and Policy Analysis*, 19(3), 205-28.

Leow, C., Marcus, S., Zanutto, E., & Boruch, R. (2004). Effects of Advanced Course-taking on Math and Science Achievement: Addressing Selection Bias Using Propensity Scores. *American Journal of Evaluation*, 25(4), 461-478.

Likert, R.A. (1932). A technique for the measurement of attitudes. In: *Archives of Psychology*, 22, 140.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not to Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling*, 9(2), 151-173.

Lowman, R.L., & Williams, R.E. (1987). Validity of self-ratings of abilities and competencies. Journal of Vocational Behavior, 31, 1-13.

Magel, R. C. (1998). Using Cooperative Learning in a Large Introductory Statistics Class. *Journal of Statistics Education* 6(3).

Mardia, K.V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika*, 57(3), 519 - 530.

Martins, P. & Walker, I. (2006). *Student Achievement and University Classes: Effects of Attendance, Size, Peers, and Teachers*. Discussion paper 2490. Forschungsinstitut zur Zukunft der Arbeit Institute for the Study of Labor.

Merrill, M. D. (1991). Constructivism and instructional design. *Educational Technology*, May, 45-53.

Merton, R.K. (1972). Insiders and Outsiders: A Chapter in the Sociology of Knowledge. *The American Journal of Sociology*, 78(1), 9 - 47.

Ministerie van Onderwijs en Vorming (2008). *Curriculum 'Wat heb je vandaag op school geleerd'? Secundair onderwijs: studieprofiel Wiskunde: beschrijving van de profielcomponenten*. Retrieved July 23, 2008, from http://www.ond.vlaanderen.be/dvo/secundair/studieprofielenaso.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. Gordon, *Statistics for the Twenty-first century.*(pp. 14-25). Washington DC: The Mathematical Association of America.

Moore, D. S. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistics Review* 65, 123-165.

Moore, D. S. and G. W. Cobb (2000). Statistics and mathematics: tension and cooperation.*The American Mathematical Monthly*, 615-630.

Motivation. (2006). In: *Encyclopdia Britannica*. Retrieved June 19, 2006, from Encyclopdia Britannica Premium Service: http://www.britannica.com.

Murphy, J. (1985). Does the Difference Schools Make, Make a Difference? *The British Journal of Sociology*, 36(1), 106-116

Musch, J. & A. Broder (1999). Test anxiety versus academic skills: A comparison of two alternative models for predicting performance in a statistics exam. *British Journal of Educational Psychology*, 69, 105-116.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431-462.

Nijveldt, M., Brekelmans, M., Beijaard, D., Wubbels, Th. & Verloop, N. (2008). Validiteit in paarsgewijze beoordelingen van docentcompetenties. *Pedagogische Studiën*, 85(4), 261-279.

Pace, C. (1984). *Measuring the Quality of College Student Experiences*. Los Angeles: University of California, Higher Education Research Institute.

Palardy, G. J., & Rumberger, R. W. (2002). *The Effects of School Size on Student Learning and Dropout and Transfer Rates*. Santa Barbara: Gevritz Graduate School of Education; University of California.

Pascarella, E.T. & Terenzini, P.T. (1991). *How college affects students*. San Fransisco: Jossey-Bass.

Pasta, D.J. (n.d.) *Using Propensity Scores to Adjust for Group Differences: Examples Comparing Alternative Surgical Methods*. Paper 261/25. San Fransisco, CA: The Lewin Group.

Pike, G.R. (1995). The relationships between self-reports of college experiences and achievement test scores. *Research in Higher Education*, 36, 1-22.

Pike, G.R. (1996). Limitations of using students self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education*, 37, 89-114.

Pohl, S., & Steyer, R. (2005). *Analyzing the effect of negatively formulated items.* Paper presented at the 14th International Meeting of the Psychometric Society, Tilburg, the Netherlands.

Pohl, S., & Steyer, R. (2006, July). *Modeling method effects as individual causal effects.* Paper presented at the SMABS-EAM Conference, Budapest, Hungary.

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling Method Effects as Individual Causal Effects.*Journal of the Royal Statistical Society, Series A.*

Prosser, M. & Trigwell, K. (1999). *Understanding Learning and Teaching: The experience in higher education.* Buckingham: Open University Press.

Prosser, M., Trigwell, K., Hazel, E. and Gallagher, P. (1994). Students' experience of teaching and learning at the topic level. *Research and Development in Higher Education*, 16, 305-310.

Pruzek, R.M. (2004). *Applications and graphics for propensity score analysis.* State University of New York at Albany.

Raykov, T. (1992). Structural models for studying correlates and predictors of change. *Australian Journal of Psychology*, 44, 101-112.

Rau, W., & Durand, A. (2000). The Academic Ethic and College Grades: Does hard Work Help Students to 'Make the Grade'? *Sociology of Education*, 73(1), 19-38.

Roiter, K. and P. Petocz (1996). Introductory Statistics Courses – A New Way of Thinking. *Journal of Statistics Education*, 4(2).

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika,* 70(1), 41-55.

Rubin, D. B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scores. *Annals of Internal Medicine*, 127(8), 757-763.

Rubin, D. B. (2001). Using Propensity Scores to Help Desigh Observational Studies: Application to the Tobacco Litigation. *Health Services and Outcomes Research Methodology*, 2, 169-188.

Rubin, D. B., & Waterman, R. P. (2006). Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology. *Statistical Science*, 21(2), 206-222.

Rutter, M., Maughan, B., Mortimore, P., Oustton, J., & Smith, A. (1979). *Fifteen Thousand Hours.* Cambridge Mass.: Harvard University Press.

Schau, C. (2003). *Students' attitudes: The 'other' important outcome in statistics education.* Paper presented at the Joint Statistical meetings.

Schau, C. (2005). *Evaluation and Statistics.* Retrieved in 2007, from www.evaluationandstatistics.com.

Schau, C., T. Dauphinee, et al. (1999). *Survey of Attitudes toward Statistics.* Albuquerque, Simpson Hall, College of Education, University of New Mexico.

Schau, C. G., Dauphinee, T., & Del Vecchio, A. (1992, April). *The development of the Survey of Attitudes Toward Statistics.* American Educational Research Association, San Francisco, CA.

Schau, C., Stevens, J., Dauphinee, T., & Del Vecchio, A. (1995). The development and validation of the survey or attitudes toward statistics. *Educational & Psychological Measurement*, 55(5), 868-875.

Schield, M. (2002). *Three kinds of statistical literacy: what should we teach?*, ICOTS6. Durban, South Africa.

Schield, M. (2004). *Statistical literacy is critical thinking about statistics in arguments*. Retrieved in November 2007 from www.augsburg.edu/ppage/schield/.

Schunk, D.H., Pintrich, P.R. & Meece, J.L. (2008). *Motivation in Education. Theory, Research and Applications.*. New Jersey: Pearson Education.

Shachar, M, & Neumann, Y. (2003). Differences between traditional and distance education academic performances: A meta-analytic approach. *The International Review of Research in Open and Distance Learning*, 4(2).

Shadish, W. R., Clark, M.H., Steiner, P. M. (forthcoming). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment, *Journal of the American Statistical Association*.

Simons, R. J., Van der Linden, J., & Duffy, T. (2000). New learning: three ways to learn in a new balance. In R. J. Simons, J. van der Linden, & T. Duffy (Eds.), *New Learning*(pp. 1-20) (chap 1). Dordrecht: Kluwer Academic Publishers.

Snee, R. (1990). Statistical thinking and its contribution to quality.*The American Statistician*, 44, 116-121.

Sorensen, A. B. (1996). Educational Opportunities and School Effects. In J. Clark (Ed.), *James S. Coleman*: Routledge.

Sorensen, A. B., & Morgan, S. L. (2000). School Effects: Theoretical and Methodological Issues. In M. T. Hallinan (Ed.), *Handbook of the Sociology of Education*. New York: Kluwer Academic/Plenum Publishers.

Sorge, C., & Schau, C. (2002, April). *Impact of engineering students' attitudes on achievement in statistics*. American Educational Research Association, New Orleans, AERA 2002.

Sowey, E. R. (1995). Teaching Statistics: making It Memorable. *Journal of Statistics Education*, 3(2).

Steinhorst, R. K. & C. M. Keeler (1995). Developing Material for Introductory Statistics Courses from a Conceptual, Active Learning Viewpoint. *Journal of Statistics Education*, 3(3).

Steyer, R. (2005). Analyzing Individual and Average Causal Effects via Structural Equation Models. *Methodology*, 1(1), 39-54.

Steyer, R. (2007). *Freeing Intercepts in LCMEM*. Personal communication.

Steyer, R., Eid, M., & Schenkmezger, P. (1997). Modeling True Intraindividual Change: True Change as a Latent Variable. *Methods of Psychological Research Online*, 2(1).

Steyer, R., Partchev, A., Krohne, U., Nagengast, B., & Fiege, C. (2007). *Causal effects in between-groups experiments and quasi-experiments: theory. Jena: University*, Department of Methodology and Evaluation Research.

Steyer, R., & Schmidt, M. J. (1990). Latent state-trait models in attitude research. *Quality & Quantity*, 24, 427-445.

Steyer, R., Schwenkmezger, P., & Auer, A. (1990). The Emotional and Cognitive Components of Trait Anxiety: A Latent State-Trait model. *Personality and individual differences*, 11(2), 125- 134.

Svanum, S. & Bigatti, S. (2006). Grade Expectations: Informed or Uninformed Optimism, or Both? *Teaching of Psychology*, 33(1), 14-18.

't Hart, H., Boeije, H. & Hox, J. (2005). *Onderzoeksmethoden*. Amsterdam: Boomonderwijs.

Tempelaar, D.T. (2007a). *Expectancy-value based achievement motivations and their role in student learning*. Universiteit Maastricht, Maastricht.

Tempelaar, D.T. (2007b). *The role of self-perceived metacognitive knowledge, skills, and attitudes, in learning mathematics*. Paper presented at EARLI2007.

Tempelaar, D. T. & Nijhuis, J. F. H. (2007). Commonalities in attitudes and beliefs toward different academic subjects. In M. K. McCuddy, H. van den Bosch, J. W. B. Martz, A. V. Matveev & K. O. Morse (Eds.), *Educational innovation in economics and business X:, The challenges of educating people to lead in a challenging world*(pp. 225-250). Berlin: Springer.

Tempelaar, D. T., Gijselaers, W. H. & Schim van der Loeff, S. (2006). Puzzles in Statistical Reasoning. *Journal of Statistics Education*, 14(1).

Tempelaar, D. T., Gijselaers, W. H., Schim van der Loeff, S. & Nijhuis, J. F. H. (2007). A structural equation model analyzing the relationship of student achievement motivations and personality factors in a range of academic subject-matter areas. *Contemporary Educational Psychology*, 32(1), 105-131.

Tempelaar, D. T., Schim van der Loeff, S., & Gijselaers, W. H. (2007). A structural equation model analyzing the relationship of students attitudes toward statistics, prior reasoning abilities and course performance. *Statistics Education Research Journal*, 6(2), 78-102.

Titus, M. (2007). Detecting selection bias, using propensity score matching, and estimating treatment effects: an application to the private returns to a masters degree. *Research in Higher Education*, 48(4), 487-521.

Ultee, W. (1977). *Groei van kennis en stagnatie in de sociologie*. Universiteit Utrecht: Faculteit Sociale Wetenschappen.

Ultee, W., Arts, W., & Flap, H. (1992). *Sociologie. Vragen, uitspraken, bevindingen*. Groningen: Wolters-Noordhoff.

Van Merrienboer, J.J.G. (1997). *Training Complex Cognitive Skills: a Four-Component Instructional Design Model for Technical Training*. New Jersy: Englewood Cliffs.

Vautier, S., Steyer, R., & Boomsma, A. (2007). The true-change model with individual method effects: Reliability issues. *British Journal of Mathematical and Statistical Psychology*.

Verhoeven, N. (2008). *Doing research. The hows and whys of applied research*. Amsterdam, Meppel: Boom Education.

Verhoeven, N., Brand-Gruwel, S., & Joosten-ten Brinke, D. (2004, paper). *Innovation of a Methodology Curriculum for Psychology students in*

distance education: evaluation of a competence-based course in an electronic learning environment. Heerlen: Open University.

Verhoeven, P. S. (2008). *Modeling attitudes toward statistics using LCME Models and propensity scores.* Paper presented at the III European Congres of Methodology.

Vlaams Ministerie van Onderwijs en Vorming (2008). *Curriculum. Wat heb je vandaag op school geleerd? Secundair onderwijs, tweede graad ASO: vakgebonden eindtermen wiskunde.* Retrieved June 19th 2008 from http://www.ond.vlaanderen.be.

Weitzen, S., Lapane, K.L., Toledano, A. Y., Hume, A.L., Mor, V. (2003). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety*, 13(12), 841-853.

Weltje-Poldervaart, M., Cade, M., & Holleman, W. (2001). *Ontwikkeling van algemene academische vaardigheden in de bachelorfase: ervaringen uit twee proeftuinen binnen de Universiteit Utrecht.* Utrecht: IVLOS.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68-81.

Wigfield, A., & Eccles, J.S. (2002). The development of competency beliefs expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J.S. Eccles (eds.), *Development of Achievement Motivation*(pp. 92 - 122). Orlando, Florida: Academic Press.

Wigfield, A., Tonks, S., & Eccles, J. S. (2004). Expectancy Value Theory in Cross- Cultural perspective. In: *Research on sociocultural influences on motivation and learning*, 165-198.

Wisenbacker, J.M., Scott, J.S. & Nasser, F. (2000). *Structural Equation Models Relating Attitudes About and Achievement in Introductory statistics Courses: A Comparison of Results from the U.S. and Israel.* Paper presented at the 9th ICME, Tokyo, Japan.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T.D. Little, K.U. Schnabel, & J. Baumert [Eds.] *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples*(Pp. 219-240). Mahwah, NJ: Lawrence Erlbaum Associates.

Yilmaz, M.R. (1996). The challenge of Teaching Statistics to non-specialists. *Journal of Statistics Education*, 4(1).

Zanutto, E.L. (2006). A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal of Data Science*, 4, 67-91.

Samenvatting - (Dutch)

Aanleiding, doel- en vraagstelling

Het vak statistiek vormt een belangrijk onderdeel van opleidingen Sociale Wetenschappen aan universiteiten in Nederland en Vlaanderen, zo begint hoofdstuk 1. Statistiek is een 'academische vaardigheid'. Dit houdt in dat studenten een onderzoek kunnen opzetten en uitvoeren en dat zij de verzamelde gegevens kunnen analyseren met behulp van statistische methoden. Het is in de meeste gevallen dan ook een verplicht vak voor bachelorstudenten. Dat neemt niet weg dat studenten statistiek in veel gevallen niet leuk vinden. Ze zien er tegenop om het vak te volgen, ze denken dat het moeilijk is en ze betwijfelen soms of ze wel zoveel statistiek in hun toekomstige loopbaan nodig hebben. Deze houding ten aanzien van statistiek bepaalt in veel gevallen de motivatie, het leergedrag en verwachtingen ten aanzien van de studieresultaten en het eindcijfer wordt erdoor beïnvloed. Althans, dat is wat de Expectancy Value Theorie daarover zegt.

In deze studie is nagegaan in hoeverre de Expectancy Value Theorie toegepast kan worden op statistiekonderwijs aan Nederlandse en Vlaamse instellingen van universitair onderwijs. Kan worden vastgesteld welke factoren invloed hebben op de houding ten aanzien van statistiek en heeft deze houding invloed op (de verwachtingen ten aanzien van) het eindcijfer, dan kunnen ook aanbevelingen gedaan worden voor de inrichting en verdere ontwikkeling van statistiekonderwijs aan universiteiten. De centrale vraag luidt dan ook wat het effect is van individuele en institutionele factoren op cursusuitkomsten ten aanzien van Introductiecursussen Methodenleer & Statistiek, aan Nederlandse en Vlaamse universiteiten en (university) colleges.

De houding ten aanzien van statistiek is onderverdeeld in zes componenten (ook wel 'constructen' genoemd): Affect (positieve en negatieve gevoelens over statistiek), Cognitive Competency (hoeveel kennis en vaardigheden heb je ten aanzien van statistiek), Value (wat is de relevantie, waarde en praktische bruikbaarheid van statistiek), Difficulty (hoe moeilijk is statistiek als vak), Interest (hoe geïnteresseerd ben je in statistische onderwerpen) en Effort (hoeveel moeite doe je om statistiek onder de knie te krijgen, ook wel 'werkhouding' genoemd).

Expectancy Value Model

In hoofdstuk 2 bespreek ik het Expectancy Value Model en de vereenvoudigde

toepassing voor deze studie. Het model verklaart prestatiegerelateerde keuzes van mensen, in dit geval van studenten statistiek. Om bijvoorbeeld de cursus statistiek met goed gevolg af te leggen, maken studenten keuzes omtrent hun leergedrag. Dit gedrag levert bepaalde verwachtingen op in de toekomst (een inschatting van de slaagkans) en vervolgens laten de cursusuitkomsten (het eindcijfer) zien of deze verwachtingen uitkomen. Hoe studenten keuzes maken is afhankelijk van een aantal factoren. In de eerste plaats spelen cognitieve processen een rol zoals ervaring met statistiek (en wiskunde) en verwachtingen omtrent de slaagkans van peers (dat zijn personen in een overeenkomstige fase, in dit geval studenten van dezelfde leeftijd). Volgens het model hangen deze verwachtingen weer af van individuele kenmerken zoals zelfvertrouwen, geslacht, leeftijd en nationaliteit. Daarnaast spelen omgevingsfactoren een rol zoals de inrichting, organisatie en duur van de cursus, de frequentie van bijeenkomsten, didactische methoden, wijze van tentamineren en groepsgrootte.

Samengevat luidt de verwachting dat individuele- en omgevingsfactoren de houding ten aanzien van statistiek bepalen, op haar beurt beïnvloedt deze houding de motivatie, het leergedrag en de verwachtingen van de student, wat ten slotte effect heeft op het studieresultaat. Ook wordt verwacht dat de kwaliteit van de docent (door middel van evaluaties door studenten) een rol speelt. Om ethische redenen zijn deze evaluaties echter niet in deze studie meegenomen.

Methodologische opzet

In hoofdstuk 3 geef ik een overzicht van de onderzoeksvragen en de organisatie van het onderzoek in drie delen. Voor de dataverzameling is gebruik gemaakt van een getrianguleerde opzet: meer dan twee methoden van dataverzameling. Zo wordt de probleemstelling vanuit verschillende invalshoeken bekeken en kan een zo valide en betrouwbaar mogelijk resultaat worden verkregen. Het resultaat van de literatuurstudie is reeds beschreven: een overzicht van de ontwikkelingen op het gebied van statistiekonderwijs en de gebruikte theoretische modellen.

Vervolgens heb ik elf diepte-interviews afgenomen, met als doel om de achter liggende mening en ervaring van experts in het veld (docenten, schrijvers van statistiekboeken en onderzoekers) te peilen en informatie over concepten en operationalisaties te verzamelen. Nevendoel was de geïnterviewden toestemming vragen om een vragenlijst bij hun studenten te mogen afnemen, wat in alle gevallen lukte.

Het derde deel van de methode beslaat een veldexperiment waarbij op twee momenten vragenlijsten onder studenten werden afgenomen. Ik bespreek deze methode in hoofdstuk 5. De vragenlijst bestaat uit een standaard deel met 36 vragen die worden samengevat in 6 eerder genoemde constructen (de SATS©36). Verder is naast achtergrondkenmerken een aantal algemene attitude vragen gesteld over zelfvertrouwen en ervaring met statistiek- en wiskundeonderwijs. De vragenlijst uitgetest (in een pilot) en waar nodig aangepast. In totaal hebben 2.555 studenten aan de voormeting, de nameting of aan beiden meegedaan. Het grote aandeel missende waarden heeft een duidelijke oorzaak: afhakers deden alleen de voormeting, studenten die alleen een hertentamen kwamen doen, deden alleen de nameting. Er zijn 935 gepaarde gegevens verzameld (opgeschoond

is dat 861). Item-nonresponse was er nauwelijks.

Naast een survey onder studenten, is een korte docentenvragenlijst (N=11) afgenomen, met daarin algemene vragen over de organisatie van de cursus. Deze numerieke gegevens worden per instituut in z'n geheel ingevoerd (geen missende waarden dus) en ze vormen de institutionele factoren voor de analyse.

Resultaat diepte-interviews

Met de resultaten van de diepte-interviews (in hoofdstuk 4) wordt een antwoord gegeven op de vraag hoe Methodenleer & Statistiek wordt gegeven aan Nederlandse en Vlaamse universiteiten. De verschillen zijn groot, zowel in lengte van de cursus (tussen 4 en 26 weken) als in groepsgrootte (tussen 14 en 475 studenten), zowel in didactische methode (alleen colleges of alleen interactieve werkgroepen) als in toetsing (papers en projecten of alleen een multiple choice examen). Er wordt nog steeds opvallend veel college gegeven voor grote groepen. Dit is ingegeven door bezuinigingsrondes: aan universiteiten worden kleinere groepen niet altijd ingesteld. Projectonderwijs wordt als arbeidsintensief gezien, het is over het algemeen alleen op de kleinere colleges te vinden.

Op één aspect is meer overeenstemming gevonden: het curriculum van de cursus statistiek is voor studenten uit deze steekproef ongeveer gelijk, zij het met enkele kleine verschillen. Meest in het oog loopt het feit dat sommige cursussen alleen statistiek in hun curriculum hebben, terwijl bij andere cursussen een stuk methodenleer en projectontwikkeling wordt toegevoegd, met name bij de kleinere University Colleges. Verder blijkt dat leerdoelen over het algemeen nog niet toegepast worden in de cursussen. Ten slotte spelen volgens de docentenevaluaties door studenten een belangrijke rol in de voorspelling van cursusuitkomsten. Een uitstekende docent voor een zaal vol studenten zal nog steeds een hoge waardering krijgen, vergeleken met de wat minder aansprekende collega voor een kleine groep. Men verwacht dat het effect van groepsgrootte hierdoor (gedeeltelijk) teniet gedaan wordt.

Procedure kwantitatieve analyse

Hoofdstuk 6 bevat de procedure voor de eerste fase van de analyses. Allereerst wordt naast een 'missing values analysis' een aantal univariate en bivariate analyses uitgevoerd, waarbij de bivariate resultaten inzicht moeten verschaffen in de verandering van houding tussen de voor- en de nameting (als gevolg van het volgen van de cursus statistiek) en de verschillen daarin tussen mannen en vrouwen en tussen instituten. De multivariate analyses zijn vooral bedoeld als voorbereiding op de toetsing van structurele modellen. Naast discriminant analyse, heb ik logistische - en meervoudige regressie- en variantie-analyses uitgevoerd. Ik kijk vooral naar verschillen tussen mannen en vrouwen in hun houding ten aanzien van statistiek, naar relaties met bijvoorbeeld zelfvertrouwen, en naar enkelvoudige verschilscores. De resultaten zijn vooral bedoeld voor docenten die kennis willen nemen van de meningen, houding en gedrag van studenten statistiek.

Verder heb ik een validiteitsanalyse uitgevoerd, met als doel Nederlandse en Vlaamse uitkomsten te vergelijken met die van een Amerikaanse studie, verricht

door (o.a.) de ontwikkelaar van de SATS36© vragenlijst. Ik heb gekeken of ik deze 36 vragen (ook wel 'items' genoemd) op dezelfde wijze in de zes eerder genoemde componenten kan samenvatten, als in de Amerikaanse studie is gedaan. De analyse bestaat uit het repliceren van factormodellen zoals die in de Amerikaanse studie in 2004 zijn gerapporteerd, waarbij ik gebruik gemaakt heb van bundels van items (ook wel 'parcels' genoemd), in plaats van afzonderlijke items per component.

Het tweede deel van de analyses (beschreven in hoofdstuk 8) is gewijd aan meer geavanceerde vormen van modelbouw, waarbij ik gebruik maak van de analyse van covariantie structuren (ofwel 'Structural Equation Modeling'). Het analyseren van enkelvoudige verschilscores is wel mogelijk, maar alleen in een opzet met een compleet databestand en gegevens die verzameld zijn uit een gerandomiseerde opzet. Dat is hier niet het geval. Bovendien zouden er in de periode tussen de voor- en nameting omstandigheden geweest kunnen zijn die een methode-effect veroorzaken. Het doel van de geavanceerde analyses is om een zuiver inzicht in de verandering van houding ten aanzien van statistiek te verkrijgen (zogeheten intra-individuele verandering), eventuele methode-effecten uit de modellen filteren en te analyseren hoe deze 'latente verandermodellen' beïnvloed worden door individuele en institutionele factoren. De analyses zijn uitgevoerd met zogeheten 'Latent Change Method Effect Models', afgekort tot LCMEM.

In de analyse met LCMEM's zijn individuele en institutionele factoren niet op hun afzonderlijke effect geanalyseerd, maar is gekozen voor een combinatie van de sterkste indicatoren in één effectscore. Deze score is aangemaakt met behulp van de propensity score methode. Deze methode wordt oorspronkelijk gebruikt in ontwerpen waar geen gerandomiseerde opzet gekozen kon worden, waar incomplete (missing) cases zijn en waar zodoende verschillen in achtergrond zouden kunnen ontstaan tussen een controle- en de experimentele groep. Om vertekening naar aanleiding van deze verschillen te voorkomen wordt een zogeheten propensity score berekend die de kans voorspelt om bij een van deze twee groepen te horen. Door de score als covariaat aan de latente verschilmodellen toe te voegen wordt gecorrigeerd voor eventuele verschillen tussen de genoemde groepen.

Ik heb een afgeleide van de propensity score methode gebruikt, door niet zozeer de correctie voor eventuele verschillen tussen groepen te benadrukken, maar door het covariaat te interpreteren als één gebundelde effectvariabele; de score heeft derhalve een andere naam: 'propensity *related* score'. Voordeel van deze toepassing is dat een 'slank' model wordt geschat met een goed verklarend vermogen (ofwel 'power'). Als basis voor het berekenen van de propensity score werd tot nu toe de experimentele- en controle groep gebruikt. In mijn onderzoek is echter alleen sprake van gepaarde metingen. Om redenen van beschikbaarheid van deze groepen is ervoor gekozen om de propensity score methode op de gepaarde groepen toe te passen. Bovendien bevatten de groepen voldoende verschillen (zie eerder genoemde missende waarden) om te controleren voor eventuele vertekening door selectie.

Ten slotte zijn de uitkomsten van de latente verschilmodellen gebruikt om een structureel regressiemodel te schatten, waarbij het effect wordt nagegaan

van afzonderlijke individuele en/of institutionele indicatoren op de (latente) attitudeverandering, op verwachtingen, op leerhouding en cursusuitkomsten.

Wat zijn de uitkomsten van de kwantitatieve analyses?
Uit de resultaten (in hoofdstuk 7 en 9) blijkt dat houding ten aanzien van statistiek over het algemeen niet erg negatief is, maar dat ze wel negatiever wordt gedurende de cursus. Dat geldt met name voor de componenten 'Value' en voor 'Interest', maar niet voor 'Cognitive Competency' en 'Effort'. Studenten vinden dat ze aan het eind van de cursus competenter geworden zijn. Ook geven ze aan dat ze minder moeite hoefden te doen om het vak onder de knie te krijgen dan ze oorspronkelijk hadden verwacht. Hun positieve en negatieve gevoelens voor statistiek ('Affect') veranderen nauwelijks gedurende de cursus. Dat geldt ook voor 'Difficulty' (moeilijkheidsgraad), de enige factor die studenten niet positief waarderen: ze blijven het vak moeilijk vinden.

Cognitive Competency draagt het meeste bij aan de verklaring van cursusuitkomsten. Effort neemt een bijzondere positie in. Het lijkt erop dat Effort twee gezichten laat zien: die van studenten die met zo weinig mogelijk moeite het vak willen halen en die een oppervlakkige leerhouding aannemen en die van studenten die intrinsiek gemotiveerd zijn om kennis te nemen van statistiek en die een meer intensieve (diepe) leerhouding aannemen. Nader onderzoek zal hierover meer inzicht kunnen verschaffen. Een vergelijking met de Amerikaanse analyses laat een goede constructvaliditeit zien. Tevens zijn de gedeeltelijke verschillen tussen mannen en vrouwen bevestigd (zie ook 'individuele factoren').

Institutionele verschillen
Instituten verschillen enorm als het gaat om de houding van studenten ten aanzien van statistiek, de organisatie van cursussen, toetsmethoden, lengte en frequentie van cursusonderdelen en klassengrootte. Deze diversiteit maakte de analyses van institutionele verschillen wat betreft de geavanceerde modellen complex. De conclusies omtrent mogelijke effecten op cursusuitkomsten zijn daarom indicatief en voornamelijk kwalitatief van aard. Een verklaring voor het uitblijven van institutionele effecten kan gelegen zijn in een mogelijk selectie-effect. Elk instituut doceert groepen studenten met specifieke eigenschappen, deze studenten hebben de studie en het instituut om bepaalde redenen gekozen. In individuele factoren liggen dus institutionele factoren besloten, genest dus. De steekproef is te klein voor een multilevel analye (N=11) en bij vervolgonderzoek is een multilevel design zeker aan te raden.

Bivariate analyses tonen aan dat de meest in het oog lopende verschillen zich afspelen tussen universiteiten en colleges en in mindere mate tussen Vlaamse en Nederlandse universiteiten. Zo wordt aan colleges veel projectonderwijs in kleine groepen gegeven, worden meerdere toetsmomenten ingebouwd en is er veel interactie met docenten. Studenten aan colleges rapporteren dan ook een groter zelfvertrouwen, en een hogere Cognitive Competency en Effort. Zoals ik al heb aangegeven, kunnen deze verschillen het gevolg zijn van selectie aan de poort. Ook groepsgrootte verschilt, maar uit de analyses is niet komen vast te staan dat dit een effect heeft op cursusuitkomsten. Eerder zijn indicaties gevonden dat de

samenhang met de kwaliteit van de docent, de mate van interactie (logischerwijs meer in kleine groepen) en de persoonlijke aanpak een rol spelen bij de houding van de student. Bovendien wordt in andere studies aangegeven dat een effect van kleine groepen pas zichtbaar wordt in groepen van minder dan 15 studenten.

Verschillen tussen Vlaamse en Nederlandse studenten spelen zich af op het vlak van houding ten aanzien van statistiek. Nederlandse studenten rapporteren een positievere houding, met uitzondering van Effort, waar Vlaamse studenten een hogere score laten zien. Verder is uit de kwalitatieve analyses komen vast te staan dat de interactie tussen Vlaamse studenten en hun docent tijdens de bijeenkomsten veel formeler is dan in de Nederlandse collegezalen.

Individuele factoren

Individuele factoren, zo blijkt in hoofdstuk 9, hebben de meeste invloed op de houding ten aanzien van statistiek, op verwachtingen rondom de uitkomsten, werkhouding en op cursusuitkomsten. Met name leeftijd, geslacht, zelfvertrouwen en ervaring met wiskunde op de middelbare school zijn van belang. Nationaliteit speelt ook een rol, zij het minder sterk en waarschijnlijk ontstaan door verschillen in curricula op middelbare scholen. Zo laat een deelanalyse zien dat Duitse studenten (met meer statistiek op de middelbare school) een hoger verwachtingspatroon hebben, meer zelfvertrouwen en meer positieve houding ten aanzien van statistiek. Verder worden houding en verwachtingen positiever naarmate men ouder wordt, maar tegelijkertijd wordt dan de invloed van 'wiskunde op de middelbare school' kleiner. De invloed van zelfvertrouwen is positief.

Gender

Mannen en vrouwen verschillen in hun houding ten aanzien van statistiek: deze is voor vrouwen negatiever dan voor mannen. Verder verwachten vrouwen meer moeite te moeten doen dan mannen om het vak te halen. Ook het zelfvertrouwen is voor vrouwen lager dan voor mannen, en in tegenstelling tot de mannelijke studenten is er bij vrouwen een relatie tussen zelfvertrouwen en attitudeverandering. Ongeacht het zelfvertrouwen geven mannen een hoger verwacht cijfer voor de cursus dan vrouwen. Het effect van attitudeverandering op cursusuitkomsten is positief voor vrouwen en afwezig voor mannen. Kortom, voor mannen zijn veel correlaties in de geteste modellen niet groot of zelfs afwezig, terwijl ze veelal goed passen voor vrouwelijke studenten. Een verklaring hiervoor kan gelegen zijn in het feit dat er voornamelijk studenten *Sociale Wetenschappen* in de steekproef zitten, met tweederde vrouwen.

Verwachtingen rondom cursusuitkomsten

Hoe positiever de houding is ten aanzien van statistiek, des te hoger zijn de verwachtingen van de cursusuitkomsten en het eindcijfer. Deze verwachtingen worden positief beïnvloed door eerdere ervaringen met statistiek en wiskunde. Ook is een positieve relatie met zelfvertrouwen aangetoond. Er is geen relatie tussen verwachtingen omtrent cursusresultaten en 'inspanning' (Effort) aangetoond, maar wel met het aantal gestudeerde uren. Hoe meer uren de student gestudeerd heeft, des te hoger zijn de verwachtingen omtrent cursusuitkomsten

en des te hoger zijn de daadwerkelijke cursusresultaten.

Conclusies en aanbevelingen
In hoofdstuk 10 bespreek ik de conclusies, doe ik aanbevelingen en kaart ik aantal discussiepunten aan. De beste verklaring voor cursusuitkomsten wordt gegeven in een structureel regressiemodel met de individuele factoren leeftijd, geslacht, zelfvertrouwen, wiskunde ervaring, de veranderde houding ten aanzien van statistiek, Effort, het aantal gewerkte uren en verwachtingen omtrent de uitkomsten. In dit opzicht kan het Expectancy Value model worden toegepast op de houding ten aanzien van statistiek en daaraan gerelateerde cursusuitkomsten. Over het algemeen valt het wel mee met de negatieve grondhouding van studenten als ze verplicht een cursus statistiek volgen. De gemiddelde waardering is goed. Studenten vinden het vak onverminderd moeilijk, maar ze vinden wel dat ze er iets van leren. Zorgelijker is de negatieve verandering in die houding gedurende de cursus. Docenten zouden daar alert op kunnen zijn en hun cursus(materiaal) daarop aanpassen.

De aanbevelingen gaan over de ontwikkeling en inrichting van het vak statistiek. Te denken valt aan een afwisseling van didactische methoden, evenals een persoonlijke en interactieve aanpak in kleine groepen. Belangrijk is ook de kwaliteit lesgeven, een aspect dat in deze studie niet is meegenomen. Voor een docent is het essentieel om studenten te kunnen stimuleren en motiveren, zodat ze een vorm van statistisch redeneren ontwikkelen (bijvoorbeeld door ze projectmatig te laten werken met levensechte designs en data). Verder kunnen docenten veel doen om het zelfvertrouwen van hun studenten te vergroten, met name bij vrouwelijke studenten, bijvoorbeeld door positieve feedback te geven, beschikbaar te zijn voor uitleg en/of begeleiding en projecten te organiseren waar studenten het vak als het ware 'in de vingers' krijgen. Ook het instellen van meerdere toetsmomenten helpt, omdat dan niet alles afhangt van één of twee toetsmomenten. Verder kan peer-assessment worden gebruikt.

Een element van onzekerheid en daardoor negatieve verwachtingen is de mate waarin de student geïnformeerd is over het vak. Docenten kunnen de voorinformatie verbeteren door tijdig uitgebreide cursusdocumentatie te verstrekken en informatiesessies bij de start van de cursus te organiseren. Op hun beurt zouden studenten ter voorbereiding, voor aanvang van de cursus het cursusmateriaal moeten doorlezen. Zo ontwikkelen ze een 'geïnformeerde verwachting' en dat reduceert onzekerheid. Aangetoond is dat studenten meer zelfvertrouwen hebben naarmate ze meer ervaring hebben met statistiek en wiskunde. Een aanvulling wat betreft statistiek op Nederlandse en Vlaamse middelbare schoolcurricula is daarom aan te bevelen.

Toekomstig onderzoek?
Onderzoeksaanbevelngen gaan over de bijzondere positie van 'Effort', over het meenemen van evaluaties van docenten en het analyseren van het effect van 'wiskunde-ervaring' op prestaties. In de methodologische opzet wordt een longitudinale studie en een multilevel analyse aanbevolen, evenals het verder uitwerken van de latente verschilmodellen met methode-factor (LCMEM) en de propen-

sity gerelateerde methode. Deze toepassingen staan nog in de kinderschoenen en aanvullende analyses kunnen de robuustheid van deze methodes aantonen. Met name het verbeteren van de interpretatie van de 'propensity gerelateerde covariaat' en de keuze voor 'gepaarde groepen' vragen aanvullende analyse.

Wat is de toegevoegde waarde van deze studie?
Voor het eerst is houding ten aanzien van statistiek vergeleken tussen een aantal instituten, tussen universiteiten en colleges en tussen Nederlandse en Vlaamse instellingen van hoger onderwijs. Uit de resultaten blijkt verder dat, evenals in de Amerikaanse analyses, de SATS36© een stabiel en valide instrument is om de houding ten aanzien van statistiek te meten. Aanbevolen wordt om dit instrument in evaluatiestudies te (gaan) gebruiken. Effort neemt een bijzondere plaats in, aangezien een tweedeling in het gedrag van dit construct is waar te nemen: een oppervlakkig en een dieper leergedrag. Ten slotte kan een aantal aanbevelingen worden gedaan die docenten helpt het vak statistiek verder te ontwikkelen en aan te passen, zodat de houding ten aanzien van statistiek een positieve verandering doormaakt, studenten daardoor een betere leerhouding hebben met positieve gevolgen voor de cursusuitkomsten.

Ook methodologisch is er winst te behalen. Deze studie laat allereerst een alternatief zien voor standaard analysemethoden, waar methode-effecten niet direct zichtbaar kunnen worden gemaakt. De latente verschilmodellen, voorzien van een methodefactor, bieden uitkomst. In alle gevallen is een methode-effect aangetoond: controle voor dit methode effect levert hogere pretest- posttest correlaties op. Mogelijke oorzaken van deze methode-effecten zijn gelegen in de verschillen in setting tussen de voor- en de nameting, verschillen in de omstandigheden, de ruis die optreedt door de verschillen in cursusduur en het gebruik van hetzelfde meetinstrument voor voor- en nameting.

Ten tweede zijn propensity gerelateerde covariaten ingevoerd, naar de propensity score methode. Dat was nodig omdat de data niet compleet zijn en niet afkomstig van een a-selecte steekproef. Behalve het aanbrengen van balans in eventuele verschillen tussen groepen die aan de voor- en/of nameting hebben deelgenomen, heeft deze methode nog een ander voordeel: groepen indicatoren kunnen gecombineerd worden en samen als een covariaat worden getest op hun effect op cursusuitkomsten. Door de propensity score methode als basis te gebruiken worden de sterkste indicatoren geselecteerd en zo krijg je een slank, goed verklarend en robuust model voor cursusuitkomsten. Het laat ook zien dat institutionele variabelen op een ander niveau zijn gemeten dan individuele variabelen. De laatste passen beter op de latente verschilmodellen en ze laten zich goed interpreteren in een structureel regressiemodel.

De belangrijkste aanbeveling die gedaan kan worden is dat docenten hun studenten moeten kunnen motiveren en enthousiasmeren zodat ze leren kritisch met statistiek om te gaan, zowel tijdens hun studie als in de (beroeps)praktijk.

*

Appendix A

Topic list

All in-depth interviews were held in Dutch. In this appendix the topics for these interviews are described, as is the interview-structure.

Introduction The interviewer introduces herself and informs the interviewee of the content and objectives of this interview. The objective is to explore experiences and to conceptualize. A.k.a. as getting acquainted with the concepts used in 'teaching statistics', getting familiar with the stats jargon. The topic is explained and the boundaries are set: this interview deals with experiences in first year mandatory introduction in Statistics. The interviewer gives an indication of the duration of the interview, promises confidentiality of the results and offers to keep the interviewee informed on the progress. Also a copy of the dissertation is offered.

Interview topics The interviewer uses a topic list instead of a structured questionnaire, the aim being to give the interviewee the utmost opportunity to give his or her own insight or opinion. Main topics / questions are:

- personal experience
  - experience in teaching and researching Statistics education
  - broad or narrow (number of years at same institute)
  - comparison teaching methods start of career - now
  - comparison own teachership start of career - now
  - comparison entrance level students start of career - now (with respect to the changes in secondary education, 'leerhuis', level and teaching methods mathematics, et cetera)
- What is the current situation on teaching Introductory Methods & Statistics in the college / university?
- what does their student population look like?

- Is it (in your opinion)necessary to make the introductory courses Methods & Statistics compulsory for all freshmen who major in Social Science? If yes, why?
  - compulsory for all freshmen?
  - what are the objectives?
  - how are these objectives measured?
  - to what extend are these objectives met and why?
  - what are the objectives?
  - how are these objectives measured?
  - to what extent are these objectives met and why?
- How are the introductory courses Methods & Statistics organized with respect to:
  - schedule / course outline
  - length course (in weeks / months)
  - teaching methods
  - what subjects are taught? Key concepts?
  - class size / group size
  - learning goals, if any? Used? Results? Perceptions (keep asking)
  - methods of assessment and grading
  - division lectures / group discussions / assignments / tests / projects (dependent on the answer given earlier).
  - methods of course-evaluation
  - network of colleagues? Available? Regular meetings? What about? (keep asking)
- in order not to annoy the interviewees and to show interest in the department under study, a few confirmatory questions will be prepared, using the Internet as a source. Main subject: educational factors, teaching methods.
- What is the current situation on course outcomes with respect to Introductory courses Methods & Statistics?
- what do the interviewees expect with respect of future developments of courses Methods & Statistics:
  - innovation teaching methods
  - innovation assessment & grading
  - linkage between Methods & Statistics and the curriculum as a whole.
  - requirements with respect to Methods & Statistics

Appendix B

Teacher's questionnaire

Introductie
Deze korte vragenlijst hoort bij het onderzoek naar Kwaliteit in Statistiek Onderwijs aan Eerstejaars studenten aan Universiteiten en Colleges. Behalve naar meningen, motivatie en houdingen van studenten zijn wij ook benieuwd naar de inrichting van het Statistiek onderwijs aan de instellingen die aan deze studie meedoen. Enige tijd geleden is daarvoor al een serie diepte-interviews georganiseerd. Om een betrouwbare en valide analyse te kunnen doen, wil ik u vragen om deze korte vragenlijst in te vullen. Het betreft een aantal cursusgegevens. Het invullen van de vragenlijst kost ongeveer 5 minuten. Alvast hartelijk dank voor de genomen moeite.

1. Voor welke opleiding wordt de Introductiecursus Statistiek gegeven? .....
2. Hoeveel ECTS vertegenwoordigt deze introductiecursus Statistiek? ..... ECTS.
3. Hoeveel studenten nemen dit keer in totaal aan deze cursus deel? ..... studenten.
4. Kunt u het aantal groepen aangeven waarin de cursisten worden ingedeeld? ... groepen ... studenten per groep.

Nu volgen een paar vragen over de wijze waarop de cursus gegeven wordt.

5. Hieronder wordt een aantal onderwijsvormen genoemd. Kunt u (in volgorde van gebruik) de drie belangrijkste onderwijsvormen aangeven die voor deze cursus worden gebruikt?

Onderwijsvorm	**Aantal malen per week**	**Aantal uren per keer**
Hoorcolleges		
Werkcolleges		
Werkgroepjes		
Individuele begeleiding		
Anders, namelijk		

6. Hoe lang duurt de cursus in totaal? ..... weken.

Ten slotte wil ik graag van u weten welke beoordeling u voor de Introductiecursus Statistiek hanteert.

7. Hieronder volgt een aantal beoordelingsvormen. Kunt u aangeven welke vormen u gebruikt voor het beoordelen van uw studenten voor de cursus Statistiek? Er is meer dan één antwoord mogelijk.

Beoordelingsvorm	**Aantal**	**Weging**
Tentamen(s) multiple choice		
Tentamen(s) open vragen		
Tentamen(s) met zowel open vragen als multiple choice		
Huiswerkopdrachten		
Papers		
Studentenproject		
Presentatie		
Anders, te weten		

Appendix C

Pretest and posttest questionnaire

Pretest items

DIRECTIONS USED TO INTRODUCE THE QUESTIONNAIRE*[1]*: The statements below are designed to identify your attitudes about statistics. Each item has 7 possible responses. The responses range from 1 (strongly disagree) through 4 (neither disagree nor agree) to 7 (strongly agree). If you have no opinion, choose response 4. Please read each statement. Mark the one response that most clearly represents your degree of agreement or disagreement with that statement. Try not to think too deeply about each response. Record your answer and move quickly to the next item. Please respond to all of the 51 statements/questions.

We would like to have your permission to use your final grade of this Introduction course for our research. If you want to give your permission, please fill in your student number below. We will only use your final grade from the Introduction course Statistics together with student number (i.e. no name) and process all information confidentially.

_____ studentnumber.

DIRECTIONS: For each of the following statements mark the one best response.

The following items with regards to 'attitudes toward statistics' are used. The answers range from 1 (totally disagree) to 7 (totally agree):

[1] *PLEASE NOTE THAT AN EXAMPLE OF THE QUESTIONNAIRES USED CAN BE REQUESTED AT N.Verhoeven@roac.nl.*

1. I plan to complete all of my statistics assignments
2. I plan to work hard in my statistics course
3. I will like statistics
4. I will feel insecure when I have to do statistics
5. I will have trouble understanding statistics because of how I think
6. Statistics formulas are easy to understand
7. Statistics is worthless
8. Statistics is a complicated subject
9. Statistics should be a required part of my professional training
10. Statistical skills will make me more employable
11. I will have no idea of what's going on in this statistics course
12. I am interested in being able to communicate statistical information to others
13. Statistics is not useful to the typical professional
14. I plan to study hard for every statistics test
15. I will get frustrated going over statistics tests in class
16. Statistical thinking is not applicable in my life outside my job
17. I use statistics in my everyday life
18. I will be under stress during statistics class
19. I will enjoy taking statistics courses
20. I am interested in using statistics
21. Statistics conclusions are rarely presented in everyday life
22. Statistics is a subject quickly learned by most people
23. I am interested in understanding statistical information
24. Learning statistics requires a great deal of discipline
25. I will have no application for statistics in my profession
26. I will make a lot of math errors in statistics
27. I plan to attend every statistics class session

28. I am scared by statistics
29. I am interested in learning statistics
30. Statistics involves massive computations
31. I can learn statistics
32. I will understand statistics equations
33. Statistics is irrelevant in my life
34. Statistics is highly technical
35. I will find it difficult to understand statistical concepts
36. Most people have to learn a new way of thinking to do statistics

Global attitude questions:

37 How well did you do in your high school mathematics courses? ('very poorly' to 'very well')[2]

38 How good at mathematics are you? ('very poorly' to 'very well')

39 In the field in which you hope to be employed when you finish school, how much will you use statistics? ('Not at all' to 'A great deal')

40 How confident are you that you can master introductory statistics material? ('Not at all confident' to 'very confident')

Background questions pretest:

41 Your sex (male, female)

42 Do you have a Dutch passport? (Dutch, non-Dutch or both)

43 Does one of your parents (or both) originate from a country other than the Netherlands? (Yes, ... (country), or 'no')

44 Age (in years): ....

45 Degree you are currently seeking. (Bachelor, master, doctorate)

[2]The labels for the scale on each of the items for questions 37 to 40 differ from those used previously.

46 What grade do you expect to receive in this course?

A	A-	B+	B	B-	C+	C	C-	D	F

47 Add major or department here: .....

48 Number of years of high school mathematics taken .....

49 Number of college mathematics and/or statistics courses completed (don't count this semester): .....

50 What grade on a numerical scale from 1 - 10 do you expect to receive in this course? .....

THANKS FOR YOUR HELP!

Posttest questionnaire

The items for the posttest questionnaire differ from the pretest questionnaire with regards to time. The answering possibilities remained the same. One question (q51) was added to measure the perceived hours studied.

Attitude questions:

1. I tried to complete all of my statistics assignments
2. I worked hard in my statistics course
3. I like statistics
4. I feel insecure when I have to do statistics problems
5. I have trouble understanding statistics because of how I think
6. Statistics formulas are easy to understand
7. Statistics is worthless
8. Statistics is a complicated subject
9. Statistics should remain being a required part of my professional training
10. Statistical skills will make me more employable
11. I have no idea of what's going on in this statistics course
12. I am interested in being able to communicate statistical information to others

13. Statistics is not useful to the typical professional
14. I tried to study hard for every statistics test
15. I get frustrated going over statistics tests in class
16. Statistical thinking is not applicable in my life outside my job
17. I use statistics in my everyday life
18. I am under stress during statistics class
19. I enjoy taking statistics courses
20. I am interested in using statistics
21. Statistics conclusions are rarely presented in everyday life
22. Statistics is a subject quickly learned by most people
23. I am interested in understanding statistical information
24. Learning statistics requires a great deal of discipline
25. I will have no application for statistics in my profession
26. I make a lot of math errors in statistics
27. I tried to attend every statistics class session
28. I am scared by statistics
29. I am interested in learning statistics
30. Statistics involves massive computations
31. I can learn statistics
32. I understand statistics equations
33. Statistics is irrelevant in my life
34. Statistics is highly technical
35. I find it difficult to understand statistical concepts
36. Most people have to learn a new way of thinking to do statistics

'Global attitude'questions:

37 How well did you do in your high school mathematics courses?

38 How good at mathematics are you?

39 In the field in which you hope to be employed when you finish school, how much will you use statistics?

40 How confident are you that you have mastered introductory statistics material?

Background questions posttest:

41 Your sex (male, female)

42 Do you have a Dutch passport? (Dutch, non-Dutch or both)

43 Does one of your parents (or both) originate from a country other than the Netherlands? (Yes, ... (country), or 'no')

44 Age (in years): ....

45 Degree you are currently seeking. (Bachelor, master, doctorate)

46 What grade do you expect to receive in this course?

A	A-	B+	B	B-	C+	C	C-	D	F

47 Add major or department here: .....

48 Number of years of high school mathematics taken .....

49 Number of college mathematics and/or statistics courses completed (don't count this semester): .....

50 What grade on a numerical scale from 1 - 10 do you expect to receive in this course? .....

51 In a usual week, how many hours did you spend outside of class studying statistics?

(a) less than 3 hours

(b) 3 - 6 hours

(c) 7 - 10 hours

(d) 11 - 14 hours

(e) 15 hours or more

THANKS FOR YOUR HELP!

Appendix D

SATS Items

In this appendix, the original scoring on SATS Items is shown. This information was retrieved from the web on July 4th, 2005 [1]. The SATS-scores, questionnaire and a list of references is shown on www.evaluationandstatistics.com.

Candace Schau (2005) developed the SATS measurement instrument in 1995 (1995, 1999; See also Dauphinee et al., 1997; Hilton et al., 2004, 2005). In order to conduct the analysis in the same direction Schau did, we have to reverse a number of items, according to the list below (Schau, 2005). The numbers with an asterix have been reversed.

Component (subscale) scores on the SATS are formed by reversing the responses (1 becomes 7, 2 becomes 6, etc.) to the items indicated with an * and summing the items within each component. Using our 7-point response scale, higher scores then correspond to more positive attitudes. This original information is used to validate the Dutch version of the SATS and test whether the reliability and concurrent validity measure up to the original findings.

Our scale contains 36 items grouped into six components identified through a development process. Prescore data from a sample of undergraduate students supported this component structure (Dauphinee, Schau & Stevens, 1997; Schau, Stevens, Dauphinee, & Del Vecchio, 1995).

The following lists the items in my component structure. together with the component-descriptions. Alpha values are from the pretest version.

Affect: Positive and negative feelings concerning statistics.

- 3. I will like statistics.
- 4.*I will feel insecure when I have to do statistics problems.
- 15.*I will get frustrated going over statistics tests in class.
- 18.*I will be under stress during statistics classes.

[1] Source: Dauphine, Schau & Stevens, 1997.

- 19.I will enjoy taking statistics courses.
- 28.*I am scared by statistics.

Cognitive Competence: Attitudes about intellectual knowledge and skills when applied to statistics.

- 5.*I will have trouble understanding statistics because of how I think.
- 11.*I will have no idea of what's going on in statistics.
- 26.*I will make a lot of math errors in statistics.
- 31.I can learn statistics.
- 32.I will understand statistics equations.
- 35.*I will find it difficult to understand statistics concepts.

Value: Attitudes about the usefulness, relevance, and worth of statistics in personal and professional life.

- 7.*Statistics is worthless.
- 9.Statistics should be a required part of my professional training.
- 10.Statistical skills will make me more employable.
- 13.*Statistics is not useful to the typical professional.
- 16.*Statistical thinking is not applicable in my life outside my job.
- 17.I use statistics in my everyday life.
- 21.*Statistics conclusions are rarely presented in everyday life.
- 25.*I will have no application for statistics in my profession.
- 33.*Statistics is irrelevant in my life.

Difficulty: Attitudes about the difficulty of statistics as a subject.

- 6.Statistics formulas are easy to understand.
- 8.*Statistics is a complicated subject.
- 22.Statistics is a subject quickly learned by most people.

- 24.*Learning statistics requires a great deal of discipline.
- 30.*Statistics involves massive computations.
- 34.*Statistics is highly technical.
- 36.Most people have to learn a new way of thinking to do statistics.

Effort: Amount of work the student expends to learn statistics.

- 12.I plan to complete all of my statistics assignments.
- 20.I plan to work hard in my statistics course.
- 23.I plan to study hard for every statistics test.
- 29.I plan to attend every statistics class session.

Interest: Students level of individual interest in statistics.

- 1.I am interested in being able to communicate statistical information to others.
- 2.I am interested in using statistics.
- 14.I am interested in understanding statistical information.
- 27.I am interested in learning statistics.

Appendix E

Parcels

Ways of evaluating whether mini-scales (i.e. parcels) can be combined are: combinations of positively and negatively worded items, creating (if possible) pairs with opposite skew, or interpretation the (theoretical) content of the items.

In this study parceling was done according to Schau's method (Dauphinee et al., 1997; also see Little et al., 2002). This means that first the negatively worded items were reversed and then for each component, 2 or 3 parcels were computed, taking the average scores. As was described in chapter 6, before the parceling procedure started, the data were cleaned for item-nonresponse. This parceling procedure only shows pretest-results, as posttest parcels have been computed the same way.

In order to perform a thorough replication of the American analysis, pretest scores were analyzed first (N=1976), and then posttest scores were analyzed (N=1511) in the same manner. This means the missing values regard item-nonresponse as mentioned above. As item non-response concerned $< 5\%$ of the data, they were deleted listwise.

For this study the following parcels are used (numbers refer to question numbers, see appendix D):

Affect[1]
PARCEL1: 3, 4R, 28R; **PARCEL2**: 15R, 18R, 19

Cognitive Competence[2]
PARCEL1: 31, 5R, 26R; **PARCEL2**: 32, 35R, 11R

Difficulty[3]
PARCEL1: 8R, 22, 30R, 36R; **PARCEL2**:6, 24R, 34R

Value[4]
PARCEL1: 7R, 17, 21R; **PARCEL2**: 10, 13R, 33R; **PARCEL3**: 9, 16R, 25R

Interest[5]
PARCEL1: 12, 23; **PARCEL2**: 20, 29

Effort[6]
PARCEL1: 1, 2; **PARCEL2**: 14, 27.

[1] α.847. The skewness and kurtosis stay within range.
[2] α.777. Skewness and kurtosis are ok.
[3] α.817. Skewness and kurtosis, although a slight skew to the right.
[4] α.797. Parcels 2 and 3 are skewed to the left.
[5] α.836. Both parcels appear non-normal and skewed to the left.
[6] α.744. Both parcels are not normally distributed and highly skewed to the left.

Appendix F

Response Rate

Tables F.1 and F.2 contain the response rates and percentages for pretest and posttest responses split up for separate institutions. They show separate pretest and posttest responses, as well as the combined response, both before and after cleaning (removing cases missing on ≥ 2 SATS items, referred to as 'useful'). The population totals were based upon the number of enrollments given by the teachers of the courses in question. The true population is unknown.

Table F.1: Response rates in absolute numbers[1]

		pretest		**posttest**		**combi**	
Institute	(sub) population	1^{st}	useful	1^{st}	useful	useful pre + post	useful pre + / post
A	70	63	62	89	89	27	121
B	163	156	154	129	126	76	181
C	440	254	236	326	315	132	298
D	100	87	86	16	12	12	88
E	42	41	41	40	40	40	41
F	322	244	242	196	193	166	293
G	475	333	314	112	112	63	357
H	250	165	162	212	206	67	290
I	500	435	431	227	227	149	496
J	280	233	232	178	175	123	269
K	25	16	16	16	12	6	21
Total	2667	2027	1976	1536	1511	861	2555

[1]PLEASE NOTE that 'Useful Pre + Post' means that only cleaned scores on both pretest and posttest are taken into account. Pre- and/or post means that all combinations are taken into account including incomplete cases.

Table F.2: Response rates in percentages

	pretest		**posttest**		**combi**
Institute	1^{st}	useful	1^{st}	useful	useful pre + post
A	.90	.89	1.27	1.27	.39
B	.96	.94	.79	.77	.47
C	.58	.54	.74	.72	.30
D	.87	.86	.16	.12	.12
E	.98	.98	.95	.95	.95
F	.76	.75	.60	.52	.52
G	.70	.66	.24	.13	.13
H	.66	.65	.85	.82	.27
I	.87	.86	.45	.45	.30
J	.83	.83	.64	.63	.44
K	.64	.64	.48	.48	.24
Total	.76	.74	.54	.52	.32

For the complete data collection, the useful response (after cleaning) is as follows:

- Pretest: N = 1,976 (i.e. 74% of the population.)
- Posttest: N = 1,511 (i.e. 52% of the population.)
- Combination of pretest and posttest data, excluding incomplete cases: N = 861 (i.e. 32% of the population.)
- Combination of pretest and posttest data, including incomplete cases: N = 2,555.

The response rate differs a lot across institutions, the highest rating being almost 100% and the lowest being only 12%: at one instutite the questionnaire was filled in along with other course-questionnaires via intranet in close monitoring and at another college the questionnaire was handed out in a large lecture hall with a few hundred students, resulting in a lower response rate than when closely monitored in small groups. In two institutions an even lower response rate was reported. In one institute the respons was low because the course was in its final semester before a big curricular change and contrary to other semesters, a number of students did not show up for the first meeting. At one of the other institutions a few groups were accidentally left out of the posttest measurement. Lastly, the posttest response in institution A is higher than 1. This can be explained by the fact that many students who did not enroll in the course, came to do a retake of the exam and therefore filled in the posttest questionnaire. They took the course at an earlier moment that year.

Appendix G

Univariate- and bivariate results

G.1 Component reliability

The reliability results[1] in the first column below represent range of coefficient alpha values for each component from results reported in studies that have used the SATS©28 (see Schau, 2003, for a list of these studies and for more information).

Tempelaar reports reliability coefficients (2007a, p. 58) on his pretest data (posttest data reliability was not computed). The results are shown in table G.2. As is explained in appendix E, Schau only reports on pretest data and on 4 out of 6 components, as is shown in table G.1. The coefficients resemble the Schau results a great deal.

In the data from my study, the reliability coefficients have been calculated for both pretest and posttest data, on six components. Tables G.3 and G.4 show that the reliability of all components are similar to those of previous studies.

[1] The assumption is that the reliability of my components is equally good as the reliability of the components that Schau et al. (1995) and Tempelaar (2007a) found. A reliability that is no more than 5 points ($\triangle\alpha < 0.05$) different from Schau's and Tempelaar's findings is considered acceptable.

Table G.1: Reliability Schau data

Component	**Total**	**Female**	**Male**
Affect	.80 to .89	.84	.81
Cognitive Competence	.77 to .88	.83	.77
Value	.74 to .90	.85	80
Difficulty	.64 to .81	.77	.64

Table G.2: Reliability Tempelaar data

Component	**Total**
Affect	.82
Cognitive Competence	.78
Value	.78
Difficulty	.68
Interest	.80
Effort	.76

Table G.3: Reliability Verhoeven Data **PRETEST**

Component	**Total**	**Female**	**Male**
Affect	.80	.81	.77
Cognitive Competence	.77	.76	.78
Value	.78	.78	77
Difficulty	.71	.71	.70
Interest	.83	.83	.85
Effort	.80	.78	.81

The same result is shown for differences between males and females.

In sum the reliability analyses shows the first sign that the construct measures consistently in the Netherlands and Flanders and it shows approximately the same results as the Maastricht and the US data. This forms a good basis for the additional construct validity testing, by means of measurement models.

Table G.4: Reliability Verhoeven Data **POSTTEST**

Component	**Total**	**Female**	**Male**
Affect	.82	.83	.78
Cognitive Competence	.82	.82	.81
Value	.82	.83	80
Difficulty	.75	.75	.73
Interest	.84	.85	.83
Effort	.80	.79	.76

Table G.5: Descriptives Individual Variables

variable name	**sample size**	**mode**	**median**	**mean**	s^2	**normal distr.**
Gender Female	2551	female (74.4%)	n.a.	n.a.	n.a.	n.a.
Age	2514	18	18.5	19.72	14.01	0.000
Institution	2555	KUL (19.4%)	n.a.	n.a.	n.a.	n.a.
Major[2]	2521	Flanders Comm.Sc. (21.9%)	n.a.	n.a.	n.a.	n.a.
Nationality	2523	Dutch (56.0%)	n.a.	n.a.	n.a.	n.a.
Expected grade pretest	1913	7.00	6.00	6.13	1.28	0.000
Expected grade posttest	1434	6.00	6.00	5.84	2.21	0.000
Final grade	1392	6.00	6.00	5.79	4.80	0.000
Good in Math in Highschool	2543	5.00	5.00	n.a.	n.a.	n.a.
Good in Math now	2555	5.00	4.00	n.a.	n.a.	n.a.
STATS-use future job	2540	4.00	4.00	n.a.	n.a.	n.a.
Self-confidence master STATS	2536	5.00	4.00	n.a.	n.a.	n.a.
How many hours p.w.	1453	4.50	4.50	6.05	18.89	0.000
STATS-experience	2509	0.00	0.00	0.31	0.59	0.000
MATH-experience	2504	6.00	6.00	5.86	1.50	0.000

Table G.6: Correlation[3]analysis Grades

Spearman's Rho	**Final Grade**	**Expected Grade Pretest**	**Expected Grade Posttest**
Final grade		0.211**	0.335**
Exp.Grade Pre	0.211**		0.471**
Exp.Grade Post	0.335**	0.471**	

G.2 Uni- and bivariate results

Table G.7: Mean final grade across institutions

Institution	Final grade M	SD	N
A	5.97	1.93	92
B	6.80	1.22	132
C	6.26	1.11	105
D	7.01	1.15	51
E	6.37	1.37	31
F	5.89	1.69	244
G	7.02	1.78	281
H	5.55	2.13	193
I	3.48	2.26	263

Table G.8: Mean expected grade across nationality[4]

Nationality	Dutch	Belgium	Other	Total	F-value	P-value
Pretest grade	6.25 (1032)	5.78 (666)	6.68 (214)	6.13 (1912)	68.123	0.000
Posttest grade	6.03 (843)	5.40 (426)	6.03 (165)	5.84 (1434)	27.901	0.000
Final grade	6.34 (904)	3.66 (275)	6.16 (191)	5.78 (1370)	208.835	0.000

[2]For the variable 'Major' a number of main categories were computed. In case of two majors (a so-called 'double major'), the first one was used. Flanders and The Netherlands use a different system of majors. As 'communication science' is a topic in General Social Science in the Netherlands, in Flanders it falls under Sociology. For the Flemish majors, therefore separate categories have been computed.

[3]** indicates that the correlation is significant at 0.001 level (2-tailed).

[4]Note that sample sizes are in parentheses.

Appendix H

Measurement models

H.1 Model fit indices

- overall fit index: χ^2. This is considered the most basic model fit index:

$$(N-1)F_{ML} \tag{H.1}$$

  It is based upon the overall degrees of freedom in the sample and the minimized Maximum Likelihood estimation (Kline, 2005, 135). A small and insignificant value indicates that the model fits the data well. However this index is influenced a great deal by its sample size, tending to 'blow up' with large sample sizes. Furthermore, with more parameters (as the model becomes more complex) the model fit also tends to blow up. Therefore, alternative fit indices should be looked at, in order to assess the fit of these models.

- TLI, Tucker-Lewis Index also known as the NNFI. This index is sample based, parsimony adjusted and the value has a range between 0 and 1. Values above .90 are considered to indicate a good fit.

$$\text{TLI} = 1 - \frac{NC_M}{NC_B} \tag{H.2}$$

- CFI, Comparative Fit Index. This comparative fit index compares the real data matrix with the imposed data matrix. CFI penalizes for sample size, hence it does not have the same problem as our chi-square. CFI ranges between 0 and 1 and a value above 0.90 indicates a good to very good fit.

$$\text{CFI} = 1 - \frac{\hat{\delta}_M}{\hat{\delta}_B} \tag{H.3}$$

- RMSEA, Root Mean Square of Approximation. According to Kline (2005) this fit index actually represents the 'badness of fit' because the larger the

value, the worse the fit. The index is parsimony adjusted, meaning that built in is a correction for model complexity. A value below 0.06 indicates a good fit, a value between 0.06 and 0.08 indicates a fair fit. Values above 0.10 indicate not such a good fit.

$$\text{RMSEA} = \sqrt{\frac{\hat{\delta}_M}{df_M(N-1)}} \tag{H.4}$$

- SRMR, Standardized Root Mean Square Residual. This index also describes the 'badness-of-fit'. Values close to '0' indicate that the model fits well, and the bigger the value becomes, the worse the fit is. This SRMR represents the value of the mean absolute correlation residual. Originally measure (RMR) was based on unstandardized variables, but the transformation of the sample- and model- covariance matrix makes interpretation much easier: the difference between observed and predicted correlations (Kline, 2005, 141). A value < 0.08 will be considered a good fit.

H.2 Results CFA

Table H.1: Fit measurement models

Model	**sample** size	χ^2 **value**	**TLI**	**CFI**	**rmsea**
Pretest	1976	490.43**[1](50)	0.945	0.964	0.067
Posttest	1511	373.48** (50)	0.954	0.970	0.065
Pre- + Posttest combined	2555	980.20** (220)	0.952	0.970	0.037

[1] ** indicates significance at a 0.001 level. DF can be found in parentheses.

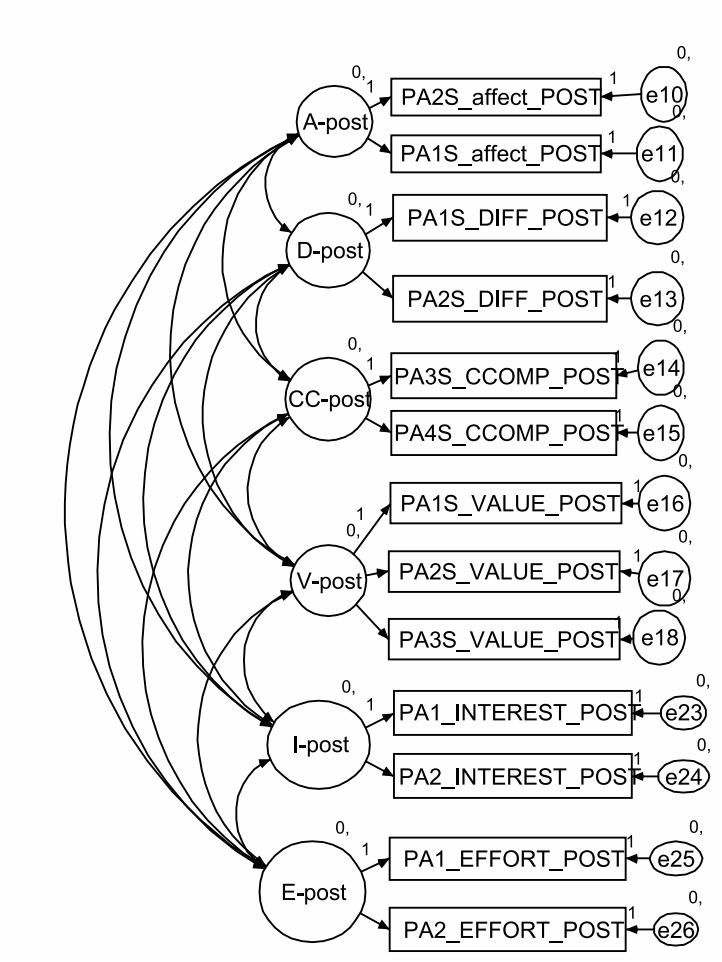

Figure H.1: model posttest

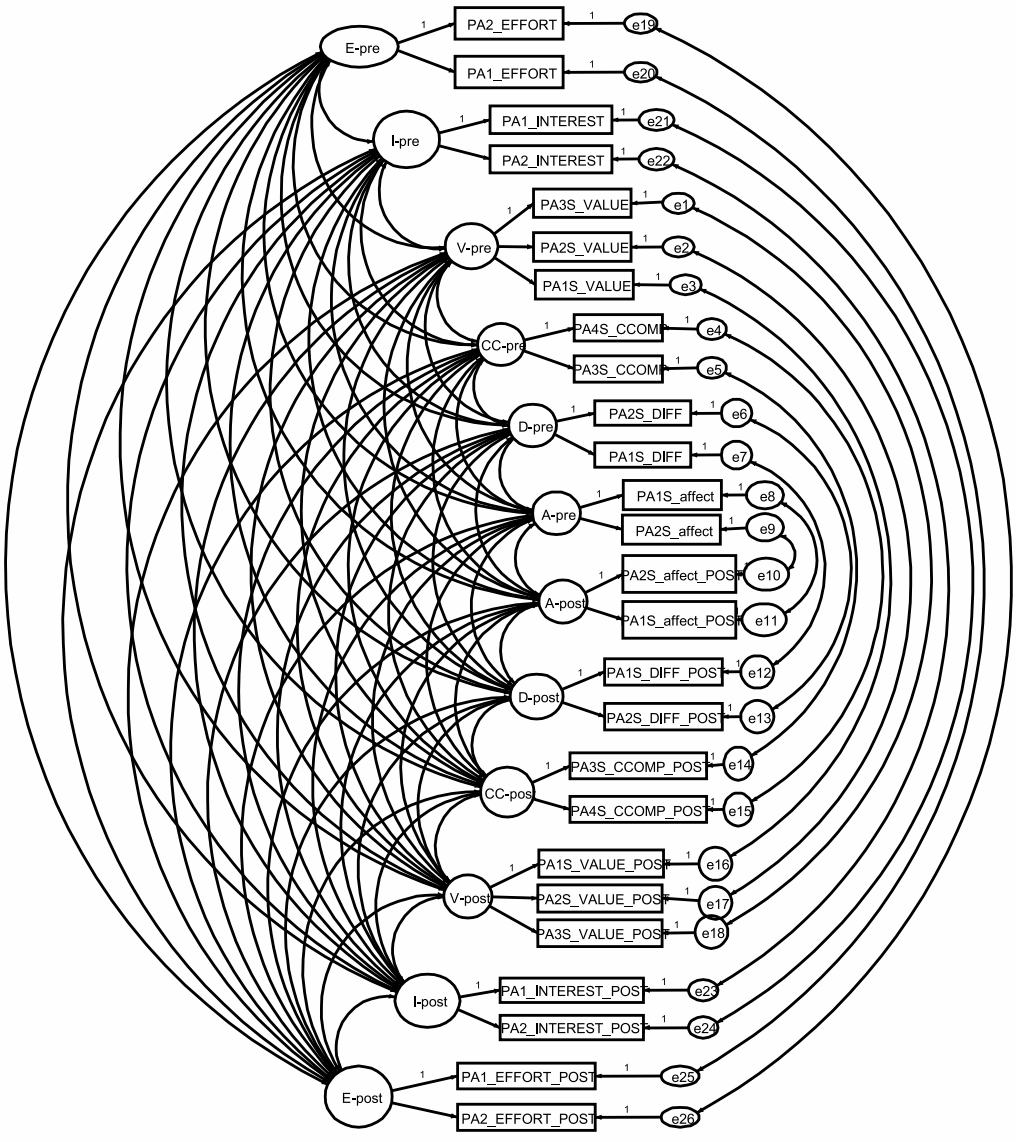

Figure H.2: pretest-posttest model

Appendix I

Institutional differences

Table I.1: Institutional differences PRETEST and ANOVA-tests (N=1980)-first part

Attitude	**A** n=62	**B** n=151	**C** n=237	**D** n=86	**E** n=41	**F** n=242	**G** n = 315
Affect	3.71	3.85	4.15	4.01	4.26	3.87	4.37
Ccomp	3.90	4.57	4.37	4.82	4.79	4.10	4.53
Value	4.56	4.78	4.99	4.62	5.11	4.66	4.73
Difficulty	2.94	3.39	3.29	3.75	3.52	3.25	3.40
Interest	4.61	4.53	4.63	4.18	4.95	4.39	4.43
Effort	6.18	6.33	5.91	5.49	5.98	5.95	5.68

Tables I.1 and I.2 show the scores on attitude-components per institution.

As it is not my intention to make a subjective comparison between colleges and universities, the institution-names have been changed to institutions A to K. This change into letters (to some extent) ensures the confidentiality that was promised at the start of the data collection. Upon request, separate colleges and universities can obtain more college specific information.

More than half of the attitudes show the lowest score (i.e. most negative) for institution A[1], followed by J(2x)[2] and K (1x)[3]. The highest score is reported with K in half of the attitudes[4], followed by D(2x)[5] and C (1x)[6]. College no. 2 scored the highest average Effort (6.33).

College K reports the biggest change[7]. College J reports a very small posi-

[1] 3.9 for Cognitive Competency, 4.56 for Value and 2.94 for Difficulty.
[2] 3.67 for Affect and 4.21 for Interest.
[3] 6.08 for Effort.
[4] 4.93 for Affect, 4.99 for Value together with C and 5.08 for Interest.
[5] 4.82 for Cognitive Competency and 3.75 for Difficulty.
[6] 4.99 for Value.
[7] -2.23 points; note that the sample size for this college is very small.

Table I.2: Institutional differences PRETEST and ANOVA-tests (N=1980)- second part

Attitude	**H** n=162	**I** n=433	**J** n=235	**K** n=16	*F-value*	*P-value*
Affect	4.14	3.73	3.67	4.93	13.06	0.000
Ccomp	4.22	4.09	4.19	4.73	12.85	0.000
Value	4.77	4.78	4.60	4.99	4.91	0.000
Difficulty	3.34	3.14	3.24	3.68	9.58	0.000
Interest	4.63	4.62	4.21	5.08	5.00	0.000
Effort	5.89	6.08	6.13	6.08	12.30	0.000

Table I.3: Institutional differences POSTTEST and ANOVA-tests (N=1511) - first part

Attitude	**A** n=89	**B** n=126	**C** n=315	**D** n=16	**E** n=40	**F** n=193	**G** n=112
Affect	3.97	3.86	4.27	4.12	4.65	4.00	4.76
Ccomp	4.37	4.62	4.66	5.07	5.08	4.57	5.01
Value	4.58	4.65	4.65	4.48	5.23	4.66	4.92
Difficulty	2.83	3.37	3.22	3.63	3.74	3.32	3.68
Interest	4.48	4.20	4.01	3.64	4.90	4.25	4.54
Effort	5.24	5.82	4.93	5.39	5.48	5.56	4.98

tive change, almost neglectible. College I reports the most negative change in Interest(-0.71 points). A negative change in Interest indicates that despite of the course, the Interest in Statistics topics has declined.

For the remaining four attitudes, changes are reported both negative and positive, hence a bit diffuse. The biggest negative changes are reported by Flemish colleges I(-0.32 for Value), J (-0.28 for Difficulty) and K(-0.41 for Affect). College D reports the biggest negative change in Cognitive Competency(-0.11), compared to the other institutions. College E reports positive changes in Affect(0.41) and Value(0.16), whereas College A reports a positive change in Cognitive Competency(0.53), and G in Difficulty(0.47). All in all the changes are small, yet they differ significantly across institutions (p=0.00). Fig. I.1 to I.6 envision these results.

The highest posttest averages are reported by College E[8]. The highest values are reported on Effort(6.19) and Affect(4.76) for Colleges J and G respectively. Lowest scores show a more diffuse pattern, as colleges I, J and K show the lowest average on Value (4.42), Affect(3.51) and Effort(4.00) respectively. College A scores 2.83 on Difficulty, and 4.37 on Cog.Competency. Interest scores the lowest at college D (3.64).

[8]5.08 on Cog. Competency, 5.23 on Value, 3.74 on Difficulty and 4.90 on Interest.

Table I.4: Institutional differences POSTTEST and ANOVA-tests (N=1511) - second part

Attitude	**H** n=206	**I** n=227	**J** n=175	**K** n=12	*F-value*	*P-value*
Affect	4.53	3.61	3.51	4.68	18.22	0.000
Ccomp	4.76	4.37	4.51	4.97	5.61	0.000
Value	4.77	4.42	4.49	5.07	5.18	0.000
Difficulty	3.30	3.00	3.00	3.69	12.50	0.000
Interest	4.35	3.89	4.05	4.13	6.17	0.000
Effort	5.11	4.81	6.19	4.00	21.38	0.000

Table I.5: Institutional differences PRETEST - POSTTEST CHANGE and ANOVA-tests (N=936) - first part

Attitude	**A** n=30	**B** n=96	**C** n=154	**D** n=14	**E** n=40	**F** n=142	**G** n=70
Affect	.16	-.12	.16	.22	.41	.06	.20
Ccomp	.53	-.10	.33	-.11	.30	.45	.47
Value	-.21	-.11	-.22	-.25	.16	-.01	.07
Difficulty	-.03	-.07	-.06	-.02	.21	.10	.47
Interest	-.53	-.31	-.45	-.54	-.04	-.15	-.20
Effort	-1.51	-.48	-.70	-.41	-.50	-.46	-.91

Table I.6: Institutional differences PRETEST - POSTTEST CHANGE and ANOVA-tests (N=936) - second part

Attitude	**H** n=78	**I** n=164	**J** n=141	**K** n=7	*F-value*	*P-value*
Affect	.22	-.12	-.17	-.41	2.74	0.002
Ccomp	.52	.27	.27	-.07	3.62	0.000
Value	-.11	-.32	-.07	-.03	2.49	0.006
Difficulty	-.07	-.12	-.28	.06	5.90	0.000
Interest	-.40	-.71	-.13	-.64	4.05	0.000
Effort	-0.92	-1.29	.07	-2.23	15.95	0.000

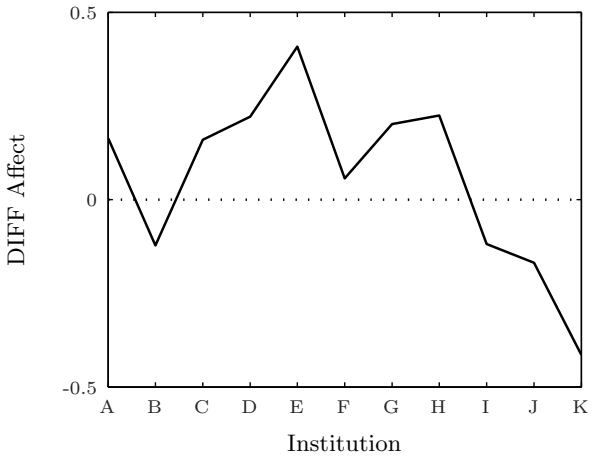

Figure I.1: Affect - average difference scores across institutions

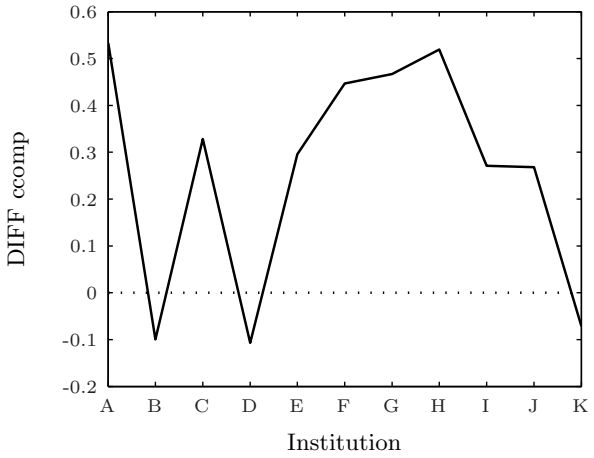

Figure I.2: Ccomp - average difference scores across institutions

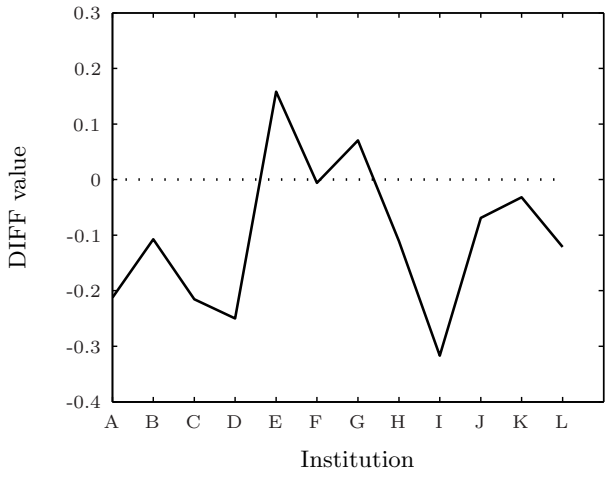

Figure I.3: Value - average difference scores across institutions

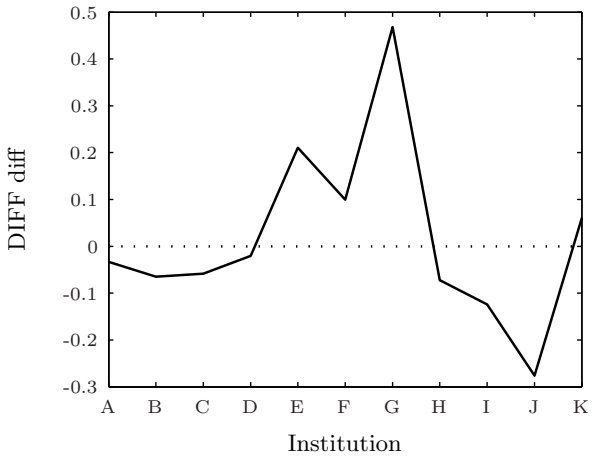

Figure I.4: Difficulty - average difference scores across institutions

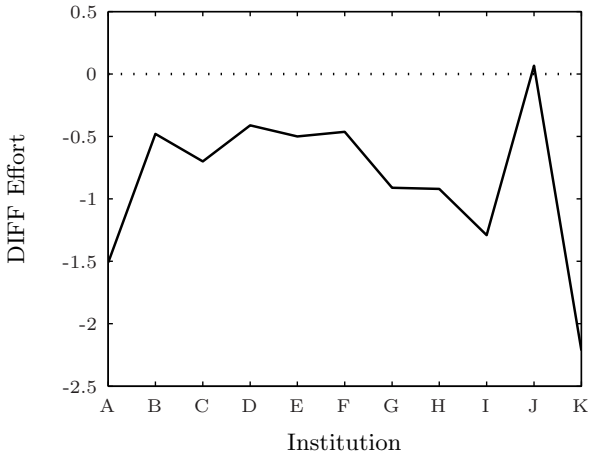

Figure I.5: Effort - average difference across institutions

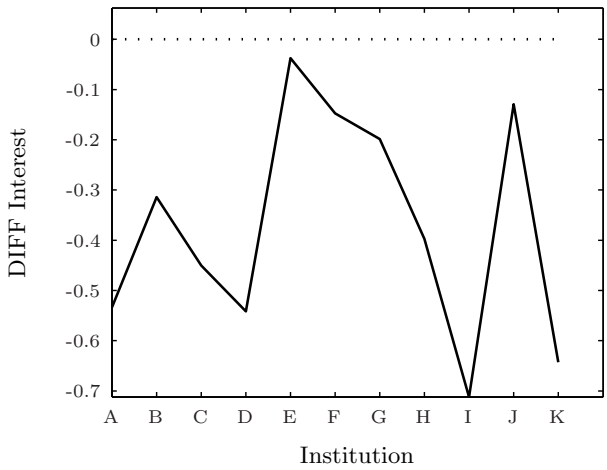

Figure I.6: Interest - average difference scores across institutions

Appendix J

Latent Change Method Effect Models

J.1 Identification procedure for LCMEM

In order to identify the model in figure 7.3 all path loadings of the items are set to unity and all intercepts are set to zero. In cases of bad fit, the alternative would be to set one intercept free. In that way the fit would become perfect, because there will be no constraints on the means nor on the variances and covariances if the measurement error variances are all free (Steyer, 2007). This option is not chosen here because it does not seem a good idea to use up all the degrees of freedom just to obtain a perfect fit. Furthermore, means of the latent factors are labeled in order to have them estimated. By default the means of the error terms are set to zero. For each of the six components, similar LCMEM are tested.

J.2 Results LCMEM

Table J.1: Model fit Latent Change Method Effect Model

	Affect	**Cogn. Comp.**	**Difficulty**	**Value**	**Interest**	**Effort**
TLI	0.885	0.616	0.984	0.863	0.987	0.240
CFI	0.988	0.962	0.998	0.922	0.999	0.924
RMSEA	0.118	0.197	0.036	0.101	0.038	0.243
Chi-square	36.43	99.81	4.36	327.36	4.31	151.43
Df	1	1	1	12	1	1

Table J.2: Means and Variances of latent attitude constructs

Construct	Means	S.E.	P	Variance	S.E.	P
Affect						
T1	4.150	0.023	0.000	0.843	0.042	0.000
T2-1	0.068	0.029	0.017	0.688	0.047	0.000
Method	-0.351	0.016	0.000	0.225	0.026	0.000
Cogn. Comp.						
T1	4.279	0.023	0.000	0.931	0.043	0.000
T2-1	0.317	0;026	0.000	0.548	0.040	0.000
Method	0.014	0.016	0.388	0.196	0.024	0.000
Difficulty						
T1	3.198	0.016	0.000	0.410	0.024	0.000
T2-1	-0.058	0.021	0.007	0.377	0.028	0.000
Method	0.196	0.013	0.000	0.125	0.018	0.000
Value						
T1	4.532	0.020	0.000	0.578	0.032	0.000
T2-1	-0.114	0.023	0.000	0.417	0.031	0.000
Method	0.324	0.016	0.000	0.208	0.022	0.000
Effort						
T1	5.878	0.020	0.000	0.388	0.045	0.000
T2-1	-0.674	0.033	0.000	0.815	0.069	0.000
Method	0.119	0.015	0.000	-0.226	0.033	0.000
Interest						
T1	4.462	0.024	0.000	0.985	0.047	0.000
T2-1	-0.344	0.030	0.000	0.798	0.053	0.000
Method	0.069	0.016	0.000	0.187	0.027	0.000

Appendix K

LCMEM with Covariates

K.1 Discriminant Analysis - procedure

In order to prepare these individual and institutional PRC's, a Discriminant Analysis is run, using SPSS:

- A discriminant analysis is conducted (Rosenbaum & Rubin, 1983). The reason for doing this is that the grouping variable has two values: either a student participated in the pretest or at least the posttest[1] in the posttest. This results in one variate for group membership.

- use a stepwise method (Wilks) in order to obtain strong and significant Discriminant scores.

- save the discriminant scores (k-1)

- repeat the DA with the strongest factors

- rename the discriminant scores according to the result of the Structure matrix. This matrix contains the discriminant loadings (i.e. the correlation of each item with the DF) and this can be used to assign labels to the Discriminant Functions.

- interpret the Standardized Canonical Discriminant Function Coefficients table. The coefficients represent semi-partial correlations between each item and the DF, controlling for the other items. The matrix gives information on the relative importance of the item's contribution to the variate.

[1] This means that the student only participated in the posttest, or in both post- and pretest. Taking part in the posttest measurement indicates that at some point in time, a student did the whole course and if there was a change in attitudes, it has already taken place.

K.2 Identification procedure for LCMEM with PRC

Identification of the LCMEM with covariates is obtained in the same way as for the models without the covariates. In case the fit of the models is not good, the possibility arises to set one intercept free, meaning that this intercept will be freely estimated. A bad fit is sometimes caused by the strictness of setting parameters to equal, to unity or to zero. In order to 'give the model more flexibility' (and test hypotheses of parallel loadings across measurement periods) one intercept is set free and the fit is assessed. This option was chosen for all models with propensity related covariates and as a result the models fit well to very well. Freeing intercepts would result in no constraints on the means nor on the variances and covariances if the measurement error variances are all free[2], and means and variances of the state- , change- and method factor were labeled in order to be estimated. This resulted in df=1 for five out of six component models.

In chapter 7, the results of the missing values analysis showed that most of the incomplete cases are 'missing at random'. In AMOS, Full Information Maximum Likelihood is used, which assumes incomplete cases to be MAR. Therefore the dataset including incomplete cases (N=2,555) was used here.

K.3 LCMEM with PRC - figures and results

Table K.1: Model fit LCMEM with both Institutional and Individual propensity related covariate

Fit Index	**Affect**	**Cogn. Comp.**	**Difficulty**	**Value**	**Interest**	**Effort**
TLI	0.969	0.995	0.944	0.924	0.977	0.610
CFI	0.993	0.999	0.987	0.960	0.995	0.907
RMSEA	0.043	0.016	0.047	0.058	0.036	0.127
Chi-square	28.35	28.05	33.40	183.23	21.36	212.16
P-value	(0.000)	(0.000)	(0.00)	(0.000)	(0.000)	0.000)
Df	5	5	5	19	5	5

[2]In the models for the 'engine' and 'change' levels this option has not been chosen, because we don't want df=0 (i.e. a 'perfect fit'), but we want to be able to assess the fit indices instead.

[3]Constraints have been added to test the hypothesis that the loading onto the pretest and posttest attitudes is equal, the alternative being that they are unequal. Furthermore standardized estimates of the PRC's show the relative importance of those covariates onto the model.

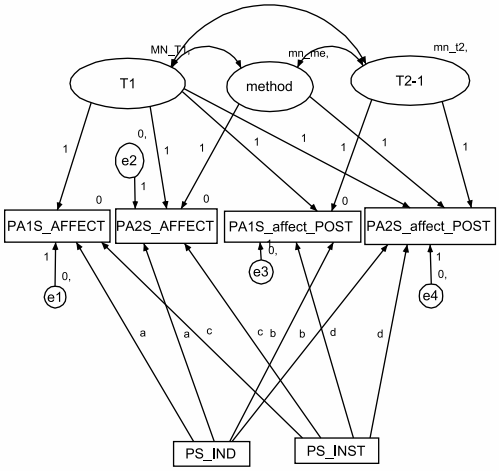

Figure K.1: Propensity Related Covariates[3]

Table K.2: Model fit LCMEM with propensity related covariate and number of hours

Fit Index	**Affect**	**Cogn. Comp.**	**Difficulty**	**Value**	**Interest**	**Effort**
TLI	0.810	0.778	0.708	0.813	0.749	0.509
CFI	0.932	0.921	0.896	0.892	0.927	0.825
RMSEA	0.095	0.093	0.098	0.084	0.098	0.136
Chi-square	238.78	237.84	256.14	492.81	247.67	481.43
P-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	0.000)
Df	10	10	10	26	10	10

Table K.3: Means and Variances of latent attitude constructs; LCMEM with individual, institutional propensity related covariate and n-hours

Construct	**Means**	**S.E.**	**P - value**	**Variance**	**S.E.**	**P - value**
Affect						
T1	4.174	0.023	0.000	0.833	0.042	0.000
T2-1	-0.023	0.048	0.629	0.698	0.047	0.000
Method	-0.420	0.020	0.000	0.240	0.026	0.000
Cogn.Comp.						
T1	4.335	0.024	0.000	0.939	0.042	0.000
T2-1	0.261	0.043	0.000	0.566	0.039	0.000
Method	-0.109	0.020	0.000	0.215	0.024	0.000
Difficulty						
T1	3.198	0.016	0.000	0.400	0.023	0.000
T2-1	0.059	0.035	0.093	0.377	0.028	0.000
Method	0.177	0.016	0.000	0.128	0.018	0.000
Value						
T1	4.522	0.021	0.000	0.571	0.032	0.000
T2-1	-0.048	0.040	0.254	0.421	0.031	0.000
Method	0.339	0.019	0.000	0.212	0.022	0.000
Effort						
T1	5.965	0.019	0.000	0.594	0.040	0.000
T2-1	-1.606	0.056	0.000	0.976	0.064	0.000
Method	0.002	0.018	0.924	0.198	0.032	0.000
Interest						
T1	4.474	0.025	0.000	0.981	0.047	0.000
T2-1	-0.495	0.050	0.000	0.807	0.054	0.000
Method	0.041	0.021	0.046	0.188	0.027	0.000

Appendix L

Models with 'final grade'

L.1 Analysis procedure for the hybrid approach

In order to assess the contribution of separate variables to 'course outcomes' a hybrid regression model is estimated. The model (see fig. 8.2 consists of the following parts:

- The dependent variable is '**final grade**'.
- Attitudes: for each attitude-component at T_{2-1}, the Factor Score weight from the LCMEM acts as a starting point to calculate the weighted score for the component at T_{2-1}, taking into account the method effect, and the parcels[1].
- Individual covariates: the variables that loaded most highly onto the Discriminant Function, are added onto the model. These variables are: Age, 'perception of level of mathematics at high school', self confidence, statistics experience.
- As has been shown before, Effort is considered to take a separate place in statistics attitudes, that is why it has been placed on a different position in the model.
- Gender has a special position in this model. Previously it was added to the Discriminant Analysis as a dummy-variable. This time it will be used in a multigroup analysis in order to test whether the model is the same for male and female students.

[1] For instance the weighted variable 'Affect at T_2' will be computed as follows: COMPUTE AFFECT-T2 = 0.370*PA1S-affect-POST + 0.350*PA2S-affect-POST + -0.322*PA1S-affect + -0.348*PA2S-affect.
EXECUTE .
All other change factors have been computed from the factor score weights in the same manner.

- 'Number of hours studied' is assumed to correlate with 'Effort'[2]. Furthermore it is assumed that it more likely has a direct effect on 'expectations' (i.e. expected grade) and indirect on 'final grade'.
- Pretest expected grade. The latter was chosen over the posttest expected grade, because pretest expectations are supposed to hold their influence throughout the model. Seeing as the prerequisite of 'proceeding the effect in time' could be violated, it has been chosen to represent the possible relation between 'attitude' and 'expected grade' by means of a correlation (i.e. correlated error terms).
- As was done with the complete pretest CFA-model, components are assumed to correlate, but as the components have been computed from the factor score weights, they have become indicators instead. Therefore correlations are assumed to go through the error terms. This time however, a slightly data driven approach is chosen, leaving out those correlations that are not significant.

Institutional variables

It has been decided not to add institutional variables to this last regression analysis. The reason for this decision has been thoroughly discussed in chapters 7 to 9. In sum there are two reasons for this decision. First the level at which institutional variables operate differs from the student level. Second, the institutional differences are too big for a consistent result to be presented. Institutional factors will be addressed again in a more qualitative approach in the conclusion chapter.

Interpreting regression outcomes

The structural model is drawn in AMOS, because I want to simultaneously estimate parameters, look at latent variables and error variance. Besides fit indices (RMSEA, TLI, CFI and χ^2) hypotheses will be tested using unstandardized coefficients. Furthermore the relative contributions will be assessed, using standardized coefficients and squared multiple correlations. As the datafile has missing values, multivariate normality cannot be checked, as the solution of bootstrapping cannot be chosen. Therefore normality is checked using SPSS and the consequences for any nonnormal data will be described. Outliers will be assessed by checking the Mahalanobis' distance.

[2]Correlations run through the error term.

Table L.1: Complete LCMEM with all covariates, expected grade and 'Grade'

Fit Index	**Affect**	**Cogn.Comp.**	**Difficulty**	**Value**	**Interest**	**Effort**
TLI	.774	.764	.679	.776	.741	.502
CFI	.895	.890	.850	.858	.879	.768
RMSEA	.090	.088	.091	.079	.090	.114
Chi-square	453.75	432.80	460.46	708.87	453.45	723.96
P-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Df	21	21	21	42	21	21

Table L.2: Leanest LCMEM and institutional invariance tests

Component	χ^2 Regional Baseline model	χ^2 Regional metric invariance	$\Delta\chi^2$ Δdf=3	χ^2 Univ. type Baseline	χ^2 Univ. type metric invariance	$\Delta\chi^2$ Δdf=3
Affect	254.43	260.00	5.58	183.89	199.86	16.97
P-value			0.134			0.001
Cogn. Comp.	206.68	223.85	17.16	113.46	136.83	23.37
P-value			0.001			0.000
Difficulty	142.56	150.97	8.41	140.99	154.30	13.31
P-value			0.038			0.004
Value	276.93	281.84	4.88	306.07	315.41	9.33
P-value			0.181			0.025
Interest	90.85	95.50	4.62	139.44	145.44	6.03
P-value			0.199			0.110
Effort	125.88	149.29	23.41	131.53	137.17	5.65
P-value			0.000			0.131

Table L.3: Leanest LCMEM and multiple squared correlations

Component	comparison	Expected grade	Grade
Affect		0.203	0.161
Affect	university	0.244	0.201
	LAS college	0.218	0.159
Affect	Flemish	0.169	0.191
	Dutch	0.191	0.106
Cognitive Competency		0.263	0.197
Cognitive Competency	university	0.290	0.213
	LAS college	0.285	0.212
Cognitive Competency	Flemish	0.192	0.214
	Dutch	0.277	0.135
Difficulty		0.143	0.165
Difficulty	university	0.150	0.180
	LAS college	0.191	0.175
Difficulty	Flemish	0.104	0.176
	Dutch	0.141	0.113
Value		0.034	0.135
Value	university	0.038	0.179
	LAS college	0.038	0.085
Value	Flemish	0.019	0.149
	Dutch	0.044	0.074
Interest		0.020	0.113
Interest	university	0.021	0.162
	LAS college	0.025	0.069
Interest	Flemish	0.012	0.150
	Dutch	0.026	0.048
Effort		0.012	0.122
Effort	university	0.008	0.154
	LAS college	0.035	0.077
Effort	Flemish	0.001	0.233
	Dutch	0.009	0.084

Curriculum Vitae

Pieternel Verhoeven was born on March 31, 1961, in Utrecht. She attended the Oosterlicht College in Utrecht from 1972-1979, after which she studied Law for one year. After working as a management-assistant for a number of years, she then decided to pursue an academic career in Sociology (part-time) in 1991.

Pieternel holds two Bsc.'s in General Social Science and Sociology and an Msc. in Sociology. She is specialized in Methods & Statistics and for her masters' thesis she performed a secondary analysis on network data, focusing on differences between male and female network contacts with respect to career perspectives. After working as research assistant at the Sociology Department of Utrecht University she founded her own research agency in 1998. This agency successfully focused on applied social scientific research, using both qualitative and quantitative strategies and triangulated designs. Pieternel Verhoeven conducted a large number of research projects for clients in governmental and non-profit organizations.

She has also developed and taught courses in Methods & Statistics for colleges and universities, with a special interest in innovative methods such as distant learning, development of Intranet courses and problem-based (competency-based) learning. She held teaching posts at the Open University, several departments of INHOLLAND College and the International School of Hospitality Management. In addition, she has coached graduate students from Utrecht University. Pieternel combined her knowledge of Methods & Statistics with her management experience in a number of large-scale projects for the Open University. This experience led to the publication of two textbooks, one on Methods & Statistics and one on management and an online course in Methods & Statistics.

Pieternel's interest lies in the development and teaching of Statistics for students from different educational backgrounds and with different entry levels. As Statistics is a mandatory course for first-year students, special teachers' skills are needed in order to accomplish long-term retention of statistical competencies.

Currently Pieternel is a university teacher and coordinator of Methods & Statistics at the Roosevelt Academy in Middelburg. Besides finishing her Phd-thesis, she functions as a coordinator for the Community Research Center 'Eleanor', a Roosevelt Academy office that aims to set up applied community based research projects for students, and thereby helps developing undergraduate research.