KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT PSYCHOLOGIE EN PEDAGOGISCHE WETENSCHAPPEN
Centrum voor Methodologie van het Pedagogisch Onderzoek

STATISTICS ATTITUDES IN UNIVERSITY STUDENTS: STRUCTURE, STABILITY, AND RELATIONSHIP WITH ACHIEVEMENT

Proefschrift aangeboden tot het verkrijgen van de graad van Doctor
in de Pedagogische Wetenschappen door

Stijn Vanhoof

Promotor: Prof. Dr. Patrick Onghena
Copromotor: Prof. Dr. Lieven Verschaffel

2010

Stijn Vanhoof, Statistics attitudes in university students: Structure, stability and relationship with achievement
Doctoral dissertation submitted to obtain the degree of Doctor in Educational Sciences, October 2010
Supervisor: Prof. Dr. Patrick Onghena. Co-supervisor: Prof. Dr. Lieven Verschaffel

---

Both in scientific research and in everyday life we are increasingly faced with statistical facts, reasoning and figures. However, research in the area of learning and teaching statistics has shown that reasoning in situations involving variability and uncertainty is frequently not in agreement with formal theory. Even after following one or several statistics courses, many students continue to show misconceptions. When investigating students' correct and incorrect reasoning in the area of statistics, attitudes and other non-cognitive factors are increasingly considered important, especially since the reform movement in statistics education. Students are supposed to be active learners able to solve non-routine problems in a social environment, and they will develop positive or negative statistics attitudes as they encounter similar experiences with statistics repeatedly. It is believed that such attitudes may increase or decrease engagement and ability to solve statistics problems. Negative statistics attitudes are often considered to be related to poor learning or low course grades. Positive attitudes are believed to go together with chances of students in developing useful statistical reasoning skills.

Because in earlier studies statistics attitudes and their relationship with statistics achievement were almost exclusively investigated before and after following one introductory statistics course, little was known about the evolution of statistics attitudes during students' whole curriculum. Therefore, the main objective of the present doctoral dissertation was to address this lacuna in the research: Statistics attitudes of 785 students Educational Sciences, and Speech Pathology and Audiology of the Katholieke Universiteit Leuven were assessed five times during the first three years of their curriculum. In the present doctoral dissertation, four manuscripts are presented in which three major aspects with regard to statistics attitudes were investigated: Structure, stability and relationship with statistics achievement. In an introductory chapter (Chapter 1), these aspects are situated within the context of the reform movement in statistics education.

In the study reported in the first manuscript (Chapter 2), a Dutch translation of the Attitudes Toward Statistics scale (ATS; Wise 1985) was used to investigate the relationship between statistics attitudes and short- and long-term statistics exam results. The data for this study were pilot-data coming from another cohort than the participants of the other studies and it is the only study making use of the ATS scale. The findings extended the knowledge regarding the connection between statistics attitudes and statistics achievement to a longitudinal context. Moreover, attitude measures at the beginning of the curriculum appeared equally predictive for long-term achievement as cognitive measures.

The second manuscript (Chapter 3) focused on several unsolved questions with regard to the structure and item functioning of our translated Survey of Attitudes Toward Statistics (SATS-36; Schau et al.,1995). Because earlier studies used the technique of item parceling to analyse the factor structure of this instrument, individual item functioning had not been evaluated before. Based on confirmatory factor analysis using individual items, the results suggested that the SATS-36 can be improved by taking some error covariances into account and by removing poorly functioning items. Furthermore, it was suggested that depending on the goals of a specific study either six subscales could be used or three of them (Affect, Cognitive Competence, and Difficulty) can be combined into one subscale without losing much information.

To examine whether the SATS-36 has appropriate properties for longitudinal comparison and to investigate the stability of statistics attitudes, the third manuscript (Chapter 4) focused on longitudinal measurement invariance of the SATS-36. Increasingly restrictive invariance tests (invariance of factor configuration, factor loadings, indicator intercepts, error variances, factor variances and factor means) were performed. Evidence of weak invariance and partial strong invariance was found for all SATS-36 subscales except Effort, providing support for the SATS-36 as a useful instrument for comparing statistics attitudes across time. Latent attitude means about the statistics domain remained stable over time, while latent mean differences emerged for students' attitudes about themselves as learners of statistics.

The goal of the study presented in the fourth and final manuscript (Chapter 5) was to investigate the directionality of the relationship between statistics attitudes and statistics achievement. Previously, not supported by appropriate empirical data, many researchers assumed a unidirectional effect from statistics attitudes to statistics achievement. In this study, structural equation modeling was used to provide empirical evidence on the directionality of effects. A comparison of alternative plausible models showed results that were opposite from the common view: A unidirectional model with effects from statistics achievement to statistics attitudes was found for students' attitudes about themselves as learners of statistics. Regarding attitudes about the domain of statistics, no effects over and above the stability effect of attitudes and achievement were present during the progress of the students' curriculum. Based on these results, it was suggested that rather than fostering positive attitudes because of their effect on achievement, improving students' achievement in statistics is a strategy for eliciting positive statistics attitudes about themselves as learners.

Finally, in Chapter 6 the main results that emerged from this doctoral dissertation are summarized and discussed and recommendations for further research and for statistics education practice are presented, such as taking suggested modifications to the SATS-36 into account, analyzing both individual items and item parcels to profit from advantages of both approaches, including attitude assessments before and after exams and students' knowing their exam results, and establishing measurement invariance before investigating attitude change.

Stijn Vanhoof, Statistiekattitudes bij universiteitsstudenten: Structuur, stabiliteit en relatie met prestaties
Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Pedagogische Wetenschappen, Oktober 2010
Promotor: Prof. Dr. Patrick Onghena. Co-promotor: Prof. Dr. Lieven Verschaffel

---

Zowel in wetenschappelijk onderzoek als in het dagelijkse leven worden we steeds meer geconfronteerd met statistische feiten en redeneringen. Uit onderzoek blijkt echter dat heel wat mensen in situaties die gepaard gaan met variabiliteit en onzekerheid redeneringen maken die niet in overeenstemming zijn met de normatieve statistische theorie. In het onderwijs blijkt dat veel studenten - zelfs na het volgen van één of meerdere statistiekcursussen – misvattingen op het gebied van statistiek blijven vertonen.

Vooral sinds de recente hervormingen in het statistiekonderwijs wordt bij onderzoek over het statistisch redeneren van studenten steeds meer belang gehecht aan attitudes en andere niet-cognitieve factoren. Van studenten wordt verwacht dat ze actief leren en niet-routinematig problemen oplossen in een sociale omgeving. Ze zullen hierbij positieve en negatieve attitudes ontwikkelen waarvan wordt aangenomen dat ze het engagement en de capaciteit om statistische problemen op te lossen kunnen verhogen of verlagen. Negatieve attitudes worden gerelateerd aan inefficiënte leerprocessen en slechte examenresultaten. Van positieve attitudes wordt verondersteld dat ze samengaan met het ontwikkelen van efficiënte statistische vaardigheden.

Omdat in eerdere studies statistiekattitudes en de relatie met prestaties bijna uitsluitend onderzocht werden voor en na een inleidende cursus statistiek, was er weinig gekend over de evolutie van attitudes tijdens het volledige curriculum van studenten. Het voornaamste doel van het voorliggende doctoraatsproefschrift was daarom tegemoet te komen aan deze tekortkoming: statistiekattitudes van 785 studenten Pedagogische Wetenschappen en Logopedische en Audiologische Wetenschappen van de Katholieke Universiteit Leuven werden vijf keer gemeten tijdens de eerste drie jaren van hun curriculum. Vier manuscripten worden in dit doctoraatsproefschrift voorgesteld waarin drie aspecten van statistiekattitudes onderzocht worden: structuur, stabiliteit en relatie met prestaties. In een inleidend hoofdstuk (Hoofdstuk 1) worden deze aspecten gekaderd binnen de bredere context, namelijk de recente hervormingen in het statistiekonderwijs.

In het eerste manuscript (Hoofdstuk 2) werd een Nederlandstalige vertaling van de Attitudes Toward Statistics scale (ATS; Wise, 1985) gebruikt om de relatie tussen statistiekattitudes en korte- en lange-termijn examenresultaten te onderzoeken. Deze studie werd uitgevoerd op pilootgegevens van een andere cohorte studenten dan de deelnemers van de drie volgende studies; het is de enige studie waarin gebruik gemaakt wordt van de ATS. De resultaten breidden de kennis over de koppeling tussen statistiekattitudes en -prestaties uit naar een longitudinale context. Bovendien bleek dat attitudemetingen aan het begin van het curriculum latere prestaties even goed voorspelden als cognitieve metingen aan het begin van het curriculum.

Het tweede manuscript (Hoofdstuk 3) focuste op enkele onopgeloste vragen over de structuur en de psychometrische eigenschappen van individuele items van onze vertaalde Survey of Attitudes Toward Statistics (SATS-36; Schau et al., 1995). Omdat in eerdere studies item parceling gebruikt werd om de factorstructuur te analyseren, konden individuele items nog niet onderzocht worden. Op basis van confirmatorische factoranalyses op de individuele items, bleek uit onze studie dat de SATS-36 verbeterd kan worden door enkele errorcovarianties in rekening te nemen en door slecht functionerende items te verwijderen. Bovendien bleek dat, afhankelijk van de doelen van een specifieke studie, de zes subschalen van de SATS-36 kunnen gebruikt worden of dat drie ervan (Affect, Cognitieve Competentie en Moeilijkheid) kunnen gecombineerd worden in één subschaal zonder dat er veel informatie verloren gaat.

Om na te gaan of de SATS-36 geschikt is voor longitudinale vergelijkingen en om de stabiliteit van statistiekattitudes te onderzoeken, werd in het derde manuscript (Hoofdstuk 4) de longitudinale meetinvariante van de SATS-36 onderzocht. Er werden sdeeds meer restrictieve tests uitgevoerd: invariantie van factorconfiguratie, factorladingen, indicatorintercepts, errorvarianties, factorvarianties en factorgemiddeldes. Evidentie voor weak invariance en partial strong invariance werd gevonden voor alle SATS-36 subschalen behalve Inzet. De SATS-36 blijkt geschikt om statistiekattitudes over te tijd te vergelijken. Latente attitudegemiddeldes over het domein statistiek waren stabiel over de tijd. Latente gemiddeldes voor de attitudes van studenten over zichzelf als lerenden van statistiek varieerden echter wel over de tijd.

De studie die gepresenteerd wordt in het vierde en laatste manuscript (Hoofdstuk 5) had als doel de richting van de relatie tussen statistiekattitudes en statistiekprestaties te onderzoeken. In eerdere studies namen veel onderzoekers aan dat er een unidirectioneel verband bestaat van statistiekattitudes naar statistiekprestaties, zonder hiervoor gepaste empirische evidentie te hebben. In de huidige studie werden structurele vergelijkingsmodellen gebruikt om de richting van de effecten empirisch te onderzoeken. De vergelijking van alternatieve mogelijke modellen leverde resultaten op die tegengesteld waren aan de gangbare opvatting: Een unidirectioneel model met effecten van statistiekprestaties naar statistiekattitudes werd gevonden voor attitudes van studenten over zichzelf als lerenden van statistiek. Wat attitudes ten opzichte van het domein statistiek betreft, werden gedurende het curriculum van de studenten geen effecten gevonden bovenop de stabiliteitseffecten van attitudes en prestaties. Op basis van deze resultaten werd gesuggereerd dat verbetering van de prestaties van studenten kan leiden tot positievere attitudes, eerder dan omgekeerd.

Ten slotte worden in Hoofdstuk 6 de voornaamste resultaten van dit doctoraatsproefschrift samengevat en bediscussieerd. Ook worden aanbevelingen voor verder onderzoek en voor de praktijk gepresenteerd, zoals het analyseren van zowel individuele items als item parcels om te kunnen profiteren van de voordelen van beide technieken, het invoeren van attitudemetingen voor en na examens en voor en na studenten hun examenresultaten kennen, en het nagaan van meetinvariantie alvorens attitudeveranderingen over de tijd te bespreken.

Dankwoord

Patrick en Lieven, bedankt voor jullie deskundige en enthousiaste begeleiding. Ik heb het vertrouwen dat jullie in me hadden, ook op momenten dat het moeilijker ging, zeer gewaardeerd. Jullie gaven me veel vrijheid maar waren altijd beschikbaar wanneer ik grote of kleine vragen had. Verrijkend waren jullie complementaire commentaren op teksten. Opvallend was jullie beider scherp oog voor detail.

I thank the members of my doctoral committee Prof. Eva Ceulemans, Prof. Bob delMas, Prof. Dirk Tempelaar, Prof. Wim Van den Noortgate and Prof. Wim Van Dooren for their constructive feedback and suggestions. Also, I am honoured that Prof. Bob delMas, Prof. Dirk Tempelaar and Prof. Bieke De Fraine have agreed to be jury member of my doctoral defense. Bedankt ook aan alle co-auteurs van de manuscripten en alle collega's waarmee ik verschillende "nevenprojecten" heb kunnen aanvatten.

Twee collega's dank ik in het bijzonder. Ana, ik heb erg genoten van onze nauwe samenwerking bij verscheidene projecten. Onvergetelijk zijn onze reizen samen. Sofie, dankzij jou heb ik de laatste jaren enkele versnellingen hoger kunnen schakelen. Jouw steun en vertrouwen en onze gesprekken over (kwantitatief) onderzoek hebben me doen doorzetten.

Collega's van "The Gang", het was zeer fijn samenwerken met jullie. We hebben de leukste en zotste momenten samen beleefd, maar ook moeilijke momenten. In beide gevallen was ik blij dat het samen met jullie was. Ana, Bartel, Goele, Hannelore, Inge, Ilse, Isis, Sigrid, Sofie, Wilfried, mijn bureaugenootje Eva en alle anderen: bedankt!

Bedankt, familie en vrienden, voor de 'gedachten-verzettende' momenten en zoveel meer. Bedankt, Jan, om me in de beginperiode de knepen van het vak te leren en raad te geven wanneer ik die nodig had.

Katrien, het doctoraat afwerken was zwaar en de druk bleef vaak ook na de werkuren hangen. Ik ben blij dat ik dit samen met jou kon trotseren.

Table of Contents

Chapter 1

General introduction

1 General background: Reform movement in statistics education

This chapter introduces the background, aims, and outline of four studies that are presented in this dissertation on university students' statistics attitudes. We start with describing the background of the studies, namely the international reform movement in statistics education (Ben-Zvi & Garfield, 2004; Moore, 1997; Shaughnessy, 2007).

For a long time, the *content* of statistics lessons was rather "traditional", with an emphasis on teaching probability theory, learning specific statistics procedures and the studying statistics from a mathematical perspective. The *goal* of this approach was accumulating statistical knowledge, the memorization of facts and formulas, and the ability to follow rules and execute procedures in rather standard contexts. The *didactic approach* to statistics was mainly characterized by an "information transfer" model with the teacher presenting clear, step-by-step demonstrations of procedures and by a lack of active student participation. In recent years, however, considerable attention has been paid by researchers, policy makers, and statistics teachers to the limitations of these traditional statistics courses. In the following paragraphs we successively describe developments in social, technological and educational areas that (together with a similar re-examination of the field of mathematics education[1]) have led to a reform movement in statistics education (Ben-Zvi &

[1] The relation between statistics and mathematics education is a complex issue. In the present doctoral dissertation, statistics is considered to be an independent field dealing with variability and uncertainty in context. The statistics field is considered to be closely related to mathematics because mathematical concepts and procedures are often used as part of the solution of statistical problems (e.g., see Cobb & Moore, 1997; Garfield, 2003; März, Vanhoof, & Onghena, 2010).

Garfield, 2004; Moore, 1997; Shaughnessy, 2007). First, because of social developments, the international research literature has argued that greater attention should be paid to statistics education. We are living in a knowledge-based society in which statistics is more than ever intertwined with daily life. For instance, inference or the drawing of a conclusion from data-based evidence abounds in the media, in the labor market, or even in the doctor's office (Ben-Zvi & Garfield, 2004; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2008; Greer, 2000; Shaughnessy, 2007). As a result, the acquisition of analytical and quantitative skills and statistical literacy has become more important. This is reflected in many higher education curricula, in which statistics courses have become essential and mandatory (e.g., Callaert, 2004; Cobb, 2005; Zieffler, 2006).

Second, the reform movement in the teaching of statistics is further stimulated by the introduction of modern technologies in the classroom (such as graphic calculators and simulation software) (Ben-Zvi, 2000; Biehler, 1993; Mills, 2002). Specifically, it has created a shift in teachers' attention from procedural to conceptual learning (Ben-Zvi, 2000). It is stated that the instructional use of simulations promises to provide students with deeper conceptual understandings, because it allows to visualize concepts such as sampling distributions and to complete computational tasks more quickly and efficiently so that students can focus more on the understanding of statistical concepts (Hodgson & Burke, 2000; Mills, 2002; Simon, 1994).

Third, in educational practice, dissatisfaction with the traditional approach was a reason to reform statistics education. Even after following one or several statistics courses, many students continued to show misconceptions (e.g., Ben-Zvi & Garfield, 2004; Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007; Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2009; Shaughnessy, 2007) and negative attitudes toward this domain (e.g., Gal, Ginsburg, & Schau, 1997; Leong, 2006; see Section 2 of this chapter).

These developments have resulted in significant changes in the content, goals, and didactic approach of statistics education. Whereas in the past, the emphasis lay on probabilities, accumulation of statistical knowledge and learning to apply specific procedures, today the focus has moved to the teaching and learning of statistical reasoning and to a balanced introduction to the world of data analysis, data collection, and inference (Moore, 1997). New technology is used to visualize concepts such as sampling distributions (e.g., Vanhoof, Castro Sotos, Onghena, & Verschaffel, 2007) or to automate routine

operations to allow more emphasis on concepts and strategies. In the new approach, students should learn statistics by doing statistics; problem solving, active learning and group work became much more important (see, among others, Ben-Zvi & Garfield, 2004; Cobb & Moore, 1997; Moore, 1997). An exploratory study of the perceptions and the implementation of these reforms by secondary education teachers in Flanders can be found in März, Vanhoof, Kelchtermans, and Onghena (2010).

Within the described reform, non-cognitive factors such as statistics attitudes make up a very important area were substantial change was needed (Gal et al., 1997; Leong 2006; McLeod, 1992; Schau, 2003). In the traditional approach on statistics education, with its emphasis on a "passive", individual accumulation of knowledge and skills, there was little interest in the influence of statistics attitudes on learning statistics. However, if students are supposed to be active learners able to solve non-routine problems in a social environment, non-cognitive factors will play a more important role. For instance, students will develop positive or negative statistics attitudes as they encounter similar experiences with statistics repeatedly. It is believed that such attitudes may increase or decrease engagement and ability to solve statistics problems (McLeod, 1992). Negative statistics attitudes are often considered to be related to poor learning or low course grades. Positive attitudes are believed to go together with chances of students in developing useful statistical thinking skills (e.g., Gal et al., 1997; Murtonen, 2005; Tempelaar et al., 2007).

Despite increasing attention for affective aspects in statistics education, there remain several open questions regarding the structure and stability of statistics attitudes and regarding the relationship between statistics attitudes and achievement. In the present dissertation we present four studies investigating these open issues. More details concerning the definition and assessment of statistics attitudes and the specific aims and outline of this dissertation are presented in the remainder part of this introduction.

2 Statistics attitudes

2.1 Definition of statistics attitudes

Attitude is a central concept in educational psychology. Numerous studies on attitudes in different fields have resulted in various conceptualisations (e.g., Eccles & Wigfield, 2002; Op 't Eynde, De Corte, & Verschaffel, 2006). However, there seems to be

general agreement that an attitude represents "a latent disposition or tendency to respond with some degree of favourableness or unfavourableness to a psychological object" (Fishbein & Ajzen, 2010, p. 76; see also Ajzen, 2001; Eagly & Chaiken, 1993). Attitudes influence the way things are perceived, experienced, and thought about and are considered highly predictive of behaviour (Eagly & Chaiken, 1995; Fishbein & Ajzen, 2010).

In the context of mathematics education, McLeod (1992) distinguishes attitudes from emotions and beliefs (see also Gal et al., 1997). *Emotions* are fleeting positive and negative responses triggered by one's immediate experiences while studying mathematics. *Attitudes* are relatively stable, intense feelings that develop as repeated positive or negative emotional responses are automated over time. *Beliefs* are individually held ideas about mathematics, about oneself as a learner of mathematics, and about the social context of learning mathematics that together provide a context for mathematical experiences.

It is clear from these descriptions, that emotions, attitudes, and beliefs represent decreasing levels of affective involvement, increasing levels of cognitive involvement, decreasing levels of intensity of response, and decreasing levels of response stability (McLeod, 1992). As Tempelaar (2007) observes, the focus in statistics education research is more on beliefs and attitudes than on emotions, because emotions are unstable and difficult to measure appropriately.

There exists a wide variety of conceptualizations of statistics attitudes inconsistencies in the use of terminology. Especially the terms *attitudes* and *beliefs* have been frequently used, without explicit attention to the distinction between them (Gal et al., 1997; McLeod, 1992). Furthermore, the concept of attitude has been used interchangeably with other concepts such as *anxiety* (Nasser, 2004; Wisenbaker & Scott, 1997), *emotion* (Zembylas, 2007), *motivation* (Murphy & Alexander, 2000), or *self-efficacy* (Finney & Schraw, 2003). Therefore, the outcomes of a study can depend on the specific definition and theoretical frame used, the goals of the study, and the instrument used to measure statistics attitudes.

In the following section, by introducing the instruments used to assess attitudes, we present and frame how the concept statistics attitudes is used in this dissertation. Several attitude dimensions are used that fit into one or more theoretical frameworks of behaviour (e.g., Eccles & Wigfield, 2002; Fishbein & Ajzen, 2010). This operationalization of statistics attitudes used is rather broad. In terms of the distinction presented by McLeod (1992), some attitude dimensions include more affective involvement and are closely related to emotions,

while other attitude dimensions include more cognitive involvement and are closely related to beliefs.

2.2 Assessment instruments

Attitudes Toward Statistics (ATS; Wise 1985)

The Attitudes Toward Statistics scale (ATS; Wise, 1985) is a 29 item, Likert-type survey with five response possibilities ranging from “strongly disagree” to “strongly agree”. The survey includes both positively and negatively formulated items. The survey consists of two subscales – *Field* (20 items) and *Course* (9 items) – that respectively aim to measure attitudes toward the use of statistics in the students’ field of study (e.g., Educational Sciences or Physics) and attitudes toward the particular statistics course in which they are enrolled. As in research on mathematics education (McLeod, 1992), these subscales relate to the distinction between attitudes about the statistics domain (e.g., the value of statistics) and students’ attitudes about themselves as learners of statistics (e.g., self-efficacy regarding statistics).

As mentioned earlier, some items have a more affective loading (e.g., “I feel intimidated when I have to deal with mathematical formulas”), while others have more a cognitive loading (e.g., “Statistical analysis is best left to the "experts" and should not be part of a lay professional's job”).

Survey of Attitudes Toward Statistics (SATS-36; Schau et al., 1995)

The Survey of Attitudes Toward Statistics (SATS-36; Schau et al.,1995) has links with several theoretical frameworks of behaviour (e.g., see Hilton et al., 2004; Schau, 2003), but is mainly related to the expectancy-value model (e.g., Schau, 2003; Tempelaar et al. 2007). In this model (Eccles & Wigfield, 2002) *Expectancies for Success* and *Subjective Task Values* are assumed to directly influence motivation, achievement, persistence, and task choice. Two factors are distinguished within *Expectancies for Success*, namely (1) *Belief about one’s own ability in performing a task* and (2) *Perception of the task demand*. Subjective task value comprises four components that are described as follows (Eccles & Wigfield, 2002, p. 120): (1) *Intrinsic value*: The enjoyment the individual gets from performing the activity or the subjective interest the individual has in the subject; (2) *Utility value*: How well a task relates

to current and future goals, such as career goals; (3) *Attainment value*: Personal importance of doing well on the task; and (4) *Costs*: Negative aspects of engaging in the task, such as anxiety and fear of both failure and success, as well as the amount of effort needed to succeed and the lost opportunities that result from making one choice rather than another.

Schau et al. (1995) and Schau (2003) developed the SATS, containing several attitude subscales that were based on the dimensions of the expectancy-value theory. A first version of the SATS (SATS-28) consisted of four dimensions: (a) *Cognitive competence* (6 items): attitudes about intellectual knowledge and skills applied to statistics; and (b) *Difficulty* (7 items): attitudes about the difficulty of statistics as a subject, (c) *Value* (9 items): attitudes about the usefulness, relevance, and worth of statistics in personal and professional life, and (d) *Affect* (6 items): positive and negative feelings concerning statistics.

Later, two dimensions were added to the survey (SATS-36; Schau, 2003): *Interest* (4 items), students' level of individual interest in statistics and *Effort* (4 items), the amount of effort students expend on learning statistics. How the six factors of the SATS relate to the expectancy-value theory is shown in Figure 1.

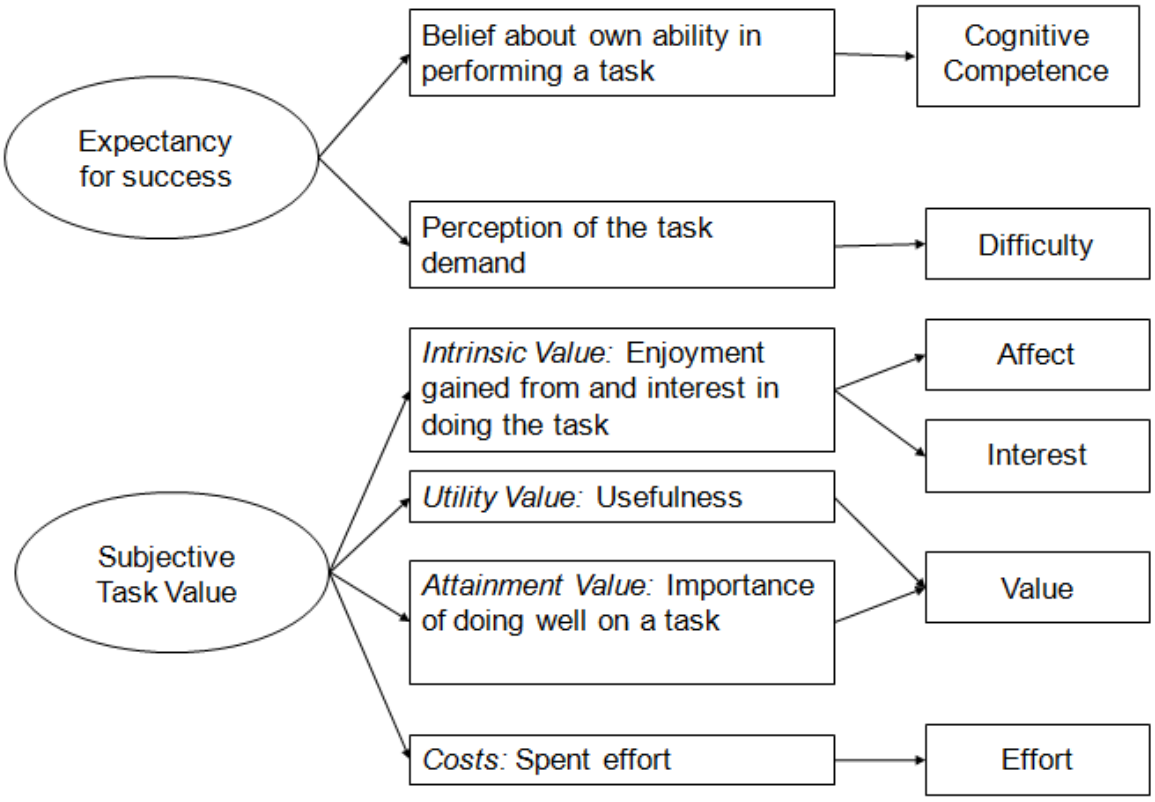

Figure 1. *Relationship between the components of the Expectancy Value Theory and the SATS-36 subscales*

Depending on the number of corresponding items, the developers labelled the survey as SATS-28 (four subscales) and SATS-36 (six subscales). Additionally, two versions of the SATS (SATS-pre and SATS-post) are available: one to administer before a statistics course and one to administer after. The difference between the two versions pertains to verb tense. A complete version of the SATS-36 and detailed scoring information can be consulted online via http://www.evaluationandstatistics.com/index.html. Because theoretical grounds and psychometric properties of the SAT are more elaborated than the ATS, the SATS was given more weight in this dissertation.

A translation of the ATS and SATS-36 from English into Dutch was made in August/September 2005. A report of this translation process is presented in Chapter 3. Full versions of the Dutch versions of the surveys (only the pretest version for the SATS-36) are enclosed in Appendix.

3 Aim and outline of this dissertation

With four longitudinal empirical studies, the present doctoral research aimed at contributing to the existing literature on statistics attitudes. As the title suggests, structure and stability of statistics attitudes and the relationship with achievement are the central focus. Chapters 2 to 6 present and discuss the background, specific research goals and results of the studies and their implications for statistics education research and practice. Because these studies are written down in self-contained manuscripts, some overlap exists, especially in the Methods sections. A brief overview of the chapters of this doctoral dissertation is presented in the following paragraphs.

Chapter 2 presents a study that used the ATS (Wise 1985) to describe students' statistics attitudes and the relationship of these attitudes with short- and long-term statistics exam results. Although studies already existed on the relationship between statistics attitudes and statistics achievement for introductory statistics courses, this study was the first to investigate this relationship in a longitudinal perspective. The central question was whether attitude measures at the beginning of the curriculum are equally predictive for long-term exam results as cognitive measures. The data for this study were pilot-data coming from another cohort than the participants of the other studies. Also, it is the only study making use of the ATS scale (Wise, 1985).

Chapter 3 focuses on several unsolved questions with regard to the structure and item functioning of SATS-36. Because earlier studies used the technique of item parceling to analyse the factor structure of this survey, individual item functioning had not been evaluated before. This longitudinal study contributed to the existing literature by addressing this remaining issue. Furthermore, it is explicitly investigated whether – as suggested by other researchers – the *Affect*, *Cognitive Competence*, and *Difficulty* subscales can be combined into one subscale without losing much information. In summary, the goal of the study was to detect specific strengths and flaws of the survey and to present researchers and statistics teachers practical guidelines for the use of the survey.

In Chapter 4 longitudinal measurement invariance of the SATS-36 is investigated in detail. Examination of invariance of factor loadings, intercepts, error variances was important in order to know whether the SATS has appropriate properties for longitudinal comparison. Investigation of invariance of factor variances and factor means revealed whether or not attitude means and variances are stable across time.

In Chapter 5 the directionality of the relationship between statistics attitudes and statistics achievement is investigated. Previously, not supported by appropriate empirical data and analyses, many researchers assumed a unidirectional effect from statistics attitudes to statistics achievement. However, alternative options regarding the direction of effects are possible: (1) The effect may go in the other direction, from achievement to statistics attitudes, or (2) there may be an effect in both directions, from attitudes to achievement and from achievement to attitudes. In this study, data collected according to our longitudinal design were analysed to provide empirical evidence on the directionality of effects.

Chapter 6, finally, summarizes the main findings of the studies presented in chapters 2 to 5 and discusses their implications for statistics education research. In addition, we propose some suggestions for the practice of statistics education.

Chapter 2

Attitudes toward statistics and their relationship with short- and long-term exam results[1]

Abstract

This study uses the Attitudes Toward Statistics (ATS) scale (Wise 1985) to investigate the attitudes toward statistics and the relationship of those attitudes with short- and long-term statistics exam results for university students taking statistics courses in a five year Educational Sciences curriculum. Compared to the findings from previous studies, the results indicate that the sample of undergraduate students have relatively negative attitudes toward the use of statistics in their field of study but relatively positive attitudes toward the course of statistics in which they are enrolled. Similar to other studies, we find a relationship between the attitudes toward the course and the results on the first year statistics exam. Additionally, we investigate the relationship between the attitudes and the long-term exam results. A positive relationship is found between students' attitudes toward the use of statistics in their field of study and the dissertation grade. This relationship does not differ systematically from the one between the first year statistics exam results and the dissertation grade in the fifth year. Thus, the affective and cognitive measures at the beginning of the curriculum are equally predictive for long-term exam results. Finally, this study reveals that the relationship between attitudes toward statistics and exam results is content-specific: We did not find a relationship between attitudes and general exam results, only between attitudes and results on statistics exams.

[1] Vanhoof, S., Castro Sotos, A. E., Onghena, P., Verschaffel, L., Van Dooren, W., & Van den Noortgate, W. (2006). Attitudes toward statistics and their relationship with short- and long-term exam results. *Journal of Statistics Education, 14*(3). Online: http://www.amstat.org/publications/jse/v14n3/vanhoof.html

1 Introduction

The importance of students' attitudes toward statistics when following an introductory statistics course is widely recognized. According to Gal, Ginsburg, and Schau (1997) such attitudes may affect the extent to which students will develop useful statistical thinking skills and apply what they have learned outside the classroom. Therefore, it is important to study thoroughly the attitudes students have toward statistics and the relationship of these attitudes with statistics achievement. A first step in accomplishing this goal is to develop and evaluate surveys to assess students' attitudes toward statistics; work that has already been initiated by a number of researchers (e.g., Roberts & Bilderback, 1980; Schau, Stevens, Dauphinee, & Del Vecchio, 1995; Shultz and Koshino, 1998; Waters, Martelli, Zakrajsek, & Popovich, 1988; Wise, 1985).

A widely used instrument is the Attitudes Toward Statistics instrument (ATS; Wise 1985). The ATS is a 29-item, Likert-type survey with five response possibilities ranging from "strongly disagree" to "strongly agree". The ATS includes both positively and negatively formulated items. The survey consists of two subscales – *Field* (20 items) and *Course* (9 items) – that respectively aim to measure attitudes toward the use of statistics in the students' fields of study and attitudes toward the particular statistics course in which they are enrolled. Example items include:

Field

I feel that statistics will be useful to me in my profession.

Studying statistics is a waste of time.

Course

The thought of being enrolled in a statistics course makes me nervous.

I get upset at the thought of enrolling in another statistics course.

The ATS scale can be used to give a general overview of the attitudes toward statistics of a group of students. Most of the previous studies using ATS (e.g., Elmore & Lewis, 1991, 1993; Waters et al., 1988; Wise, 1985) include an evaluation of the internal consistency, a description of the attitudes students have toward statistics before and after taking the statistics course, and an analysis of how these attitudes are related to their first

year statistics exam results (as an indication of their statistics achievement). Most of these studies therefore involve two administrations, one before and one after the statistics course.

The present study aims at extending the existing evidence on the relationship between attitudes toward statistics and achievement. This is done in three ways. First, the study provides new data and measures of reliability of the ATS by two administrations of the survey in an introductory statistics course for Flemish undergraduate students in Educational Sciences. Second, while previous investigations are limited to the relationship between attitudes and first year exam results, this study examines the relationship between the attitudes students have and their exam results not only at the beginning of the curriculum, but also in later years. Third, while the previous research only addresses the relationship between students' attitudes and their grades in a statistics course, the present study also investigates the relationship with their general exam results (short- and long-term).

We are aware that some authors caution against the indiscriminate use of paper-and-pencil Likert-type scales, like the ATS, to study attitudes (Gal & Ginsburg, 1994; Schau et al., 1995). For instance, it is difficult to imagine that students' attitudes toward statistics could be captured by two global ATS scores (Gal & Ginsburg, 1994). Furthermore, we have to take into account that there may be cultural differences in responding to such surveys, even at the level of subtle nuances in the translation and interpretation of the items. Therefore, we acknowledge that our study will only be one step toward a deeper understanding of the complex relationship between statistics attitudes and achievement.

2 Empirical background

Most of the previous studies use results from other investigations as a bench-mark. Therefore, we will also compare the data of the current study with data from previous studies (Aldogan & Aseeri, 2003; D'Andrea & Waters, 2002; Elmore & Lewis, 1991; Elmore, Lewis, & Bay, 1993; Mvududu, 2003; Rhoads & Hubele, 2000; Roberts & Reese, 1987; Shultz & Koshino, 1998; Waters et al., 1988; Wise, 1985). We first present a detailed overview of the results of these previous studies and emphasize the most important findings and trends that can be formulated based on these results. This overview will provide the reader with the necessary background to situate and interpret our new empirical data presented in Section 4.

The Appendix provides an overview of these studies with some additional information concerning the number of samples, administrations, and participants. It also includes the level of the course that is involved (undergraduate or graduate), the field of study (e.g. psychology, education, engineering) and some remarks. Most authors do not provide information on the specific content of the course (probability, descriptive statistics or inferential statistics). We acknowledge that differences in courses, fields of study, and other characteristics of the population and the specific statistics courses in the different studies can complicate the comparison. Yet, because most studies include an introductory statistics course in the field of human sciences (education, psychology), a prudent comparison seems justified.

In the following tables, we summarize the findings of these studies. Successively, we review (1) the internal consistency and test-retest reliability, (2) mean data (and standard deviations) for the *Course* and *Field* subscales (respectively for undergraduate and graduate students), and (3) the relationship with first year statistics exam results. Since not all investigations mention all measures, some tables contain only a subset of the studies involved in our comparative analysis.

Table 1 presents the observed internal consistency (Cronbach alphas). All studies yield coefficient alpha reliability estimates that are high for both subscales and for both administrations. In general, the estimates are between .77 and .93 for the *Course* subscale and between .83 and .96 for the *Field* subscale. Some studies (Elmore & Lewis, 1991; Elmore et al., 1993; Roberts & Reese, 1987) also mention the alpha estimate for the whole scale. Roberts and Reese (1987) find a whole scale alpha estimate of .91, Elmore and Lewis (1991) report for the first and the second administration an estimate of .92 and .93, respectively, and Elmore et al. (1993) .92 and .94.

Table 1

Internal Consistency (Cronbach alphas) for the two ATS subscales

Study	*N*	Course		Field	
		Adm 1	Adm 2	Adm 1	Adm 2
Aldogan and Aseeri 2003	178	0.92	-	0.90	-
Elmore and Lewis 1991	58	0.90	0.82	0.90	0.92
Elmore et al. 1993	289	0.90	0.90	0.90	0.93
Rhoads and Hubele 2000	63	0.77	0.85	0.89	0.90
Shultz and Koshino 1998(sample 1)	36	0.85	0.92	0.96	0.96
Shultz and Koshino 1998(sample 2)	38	0.93	0.89	0.90	0.92
Waters et al. 1988	302	0.90	0.90	0.83	0.86
Wise 1985	92	0.90	-	0.92	-

Note. "Adm." stands for "administration". Most studies include two administrations, namely one before (Adm 1) and one after (Adm 2) the statistics course. Shultz and Koshino (1998) include two samples. The first sample contains undergraduate students, the second sample graduate students (see the Appendix for more information).

Some authors also investigate the test-retest reliability for the *Course* and *Field* subscales. The reported correlations are respectively .91 and .82 (Wise 1985), .59 and .72 (undergraduates, Shultz & Koshino, 1998), and .71 and .76 (graduates, Shultz and Koshino 1998). For Wise (1985) there are only two weeks between the test and retest (as opposed to three months for Shulz & Koshino, 1998). Obviously, the time lapse between administrations can affect the reliability.

Table 2 presents the mean scores (and standard deviations) for the different studies. For all these data, if needed, item responses were reversed so that a higher score always refers to a more positive attitude. A distinction is made between undergraduate and graduate courses, since Shultz and Koshino (1998) predicted and found consistent differences in attitudes between these two groups when discussing their own and previous study results.

Since the ATS-items are scored on a Likert-type scale with five response possibilities, "strongly disagree" (score 1), "disagree" (score 2), "neutral" (score 3), "agree" (score 4) and "strongly agree" (score 5), 27 indicates an average neutral position for the whole *Course*

subscale, which contains 9 items. Similarly, because there are 20 *Field* subscale items, with each time "neutral (score 3)" as the neutral response possibility, 60 indicates an overall neutral position for the whole *Field* subscale.

Table 2

Mean scores (and standard deviations) for the two subscales of the Attitude Toward Statistics scale

Study	*N*	Course subscale		Field subscale	
Undergraduate		Adm. 1	Adm. 2	Adm. 1	Adm. 2
Elmore et al. 1993	289	24.1 (7.8)	22.1 (8.5)	79.4 (9.5)	80.2 (11.1)
Mvududu 2003 (sample 1)	120	34.9 (6.0)	-	79.5 (8.9)	-
Mvududu 2003 (sample 2)	95	28.9 (8.0)	-	74.0 (13.1)	-
Shultz & Koshino 1998 (sample 1)	36	23.3 (6.5)	24.0 (8.8)	74.5 (11.8)	74.3 (11.7)
Waters et al. 1988	212	28.3 (-)	30.2 (-)	-	-
Graduate					
Elmore & Lewis 1991	58	30.5 (7.4)	33.1 (6.3)	79.0 (9.8)	82.5 (10.9)
D'Andrea & Waters 2002	17	29.1 (9.0)	35.2 (5.7)	84.9 (9.2)	86.6 (6.7)
Shultz & Koshino 1998 (sample 2)	38	29.8 (8.9)	32.5 (7.1)	81.1 (9.2)	81.3 (9.6)

Note. Waters et al. (1988) do not provide standard deviations.

A comparison of the mean results for the undergraduate and graduate courses is in line with the conclusion of Shultz and Koshino (1998) that, in general, graduate students have higher scores than undergraduate students, for both the *Course* and *Field* subscale.

Table 3 shows the correlations between the attitude scores and the first year statistics exam results. In addition to the statistical significance of the correlations (which is discussed in all articles), we report effect sizes. Cohen (1988, 1992) provides a classification of effect sizes for correlations in terms of small (r = 0.1), medium (r = 0.3), and large (r = 0.5) effects as compared to the effects typically found in the social, educational and behavioural sciences. Except for Shultz and Koshino (1998), all studies demonstrate a statistically significant positive correlation between the first administration of the *Course* subscale scores and the exam results (first column). According to the guidelines of Cohen (1988, 1992), the corresponding correlations are small to medium. The correlations of the second administration (second column) are higher (effect sizes ranging from medium to large), and statistically significant for Shultz and Koshino (1998). None of the studies shows a statistically significant correlation between the *Field* subscale scores and the exam results for the first administration (third column). Two studies (Shultz & Koshino, 1998, first sample; Waters et al., 1988) show a statistically significant correlation for the second administration (fourth and sixth column), but for all studies in the table, the correlation at the second administration is smaller for the *Field* subscale than for as compared to the *Course* subscale.

Table 3

Correlations between ATS and first year exam results

Study	N	Course subscale		Field subscale	
		Adm. 1	Adm. 2	Adm. 1	Adm. 2
Shultz & Koshino 1998 (sample 1)	36	0.06	0.45*	0.16	0.43*
Shultz & Koshino 1998 (sample 2)	38	0.13	0.34*	0.13	0.08
Rhoads & Hubele 2000	63	0.29*	0.29*	ns	ns
Waters et al. 1988	302	0.20*	0.42*	0.07	0.17*
Wise 1985	70	0.27*	-	-0.04	-

Note. Rhoads and Hubele (2000) do not provide exact correlation values for the Field subscale. (ns stands for not significant)

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

These data are in line with the conclusion of Waters et al. (1988) that there exists a consistent positive relationship between students' attitudes toward statistics and their first year statistics exam results. They notice that especially the *Course* subscale scores are related to the statistics exam results, as also reported by Harvey et al. (1985, in Mvududu 2003). The latter authors suggested that a supportive atmosphere in the course can positively affect achievement, regardless of the attitudes toward the field of statistics.

3 Method

3.1 Participants

Participants of the present study are 264 students (218 female, 46 male) who took an introductory undergraduate statistics course in Autumn 1996 at the Department of Educational Sciences of the Katholieke Universiteit Leuven in the Flemish speaking part of Belgium. In general, most of the students who are enrolled in this academic program follow a curriculum with a considerable amount of mathematics in secondary school (four, six or eight hours every week).

The curriculum of Educational Sciences takes five years to complete. The introductory statistics course takes place in the first semester of the first year. In general, the course deals with some introductory methodology and statistical concepts (tables, figures, and descriptive statistics), but no formal probability or statistical inference. The mathematical background required for this course is limited.

3.2 Measures

Attitudes toward statistics are assessed with a Dutch translation of the ATS (Wise 1985). The ATS is administered twice. The first administration, in October 1996, at the beginning of the first year's introductory statistics course (n = 264) and the second, in October 1997, at the start of the same students' second year statistics course. In contrast to the studies mentioned in Section 2, the second administration takes place after the students know their exam result for the first year. About half of the students succeed in the first year. Therefore, the sample size of the second administration is much smaller (n = 127) and only includes students who succeed in their (overall) first year (eight of these 127 students had

not been successful in their statistics exam, but nevertheless got permission to pass to the second year).

To relate the attitude scores to statistics achievement, we record students' statistics exam results and their dissertation grades at the end of the five year program. For the statistics exam results, there are three results from obligatory statistics courses that students have to follow during their curriculum, namely in the first, the second, and the third year. For the first and the second statistics courses, the instructor is the same. For the third year's statistics course, the same teacher as in the two previous years teaches half of the course, and another teacher teaches half of it. It is important to notice that the third year results are somewhat atypical and more difficult to interpret, because the course is evaluated through group assessment. Students do not follow any statistics courses in the fourth and fifth year, but because of the major role of methodology and statistics in a student's dissertation, we consider this as a partial indication of long-term statistics achievement.

To relate the attitude scores to general achievement, we record students' general exam results for the five years of the curriculum. For the present study, we excluded the dissertation grade from the variable "general exam result", as it contributes 50% to that result.

All these measures together make it possible to relate the attitude scores of the two administrations at the beginning of the curriculum with (1) short- and long-term and (2) statistics and general exam results. As mentioned before, the conclusions of the relationship between the attitudes and long-term exam results only pertain to the students who actually pass the exams. Table 4 provides an overview of the different measures and of the sample sizes at each moment of data collection.

3.3 Statistical analyses

Reliability of the ATS is evaluated using both internal consistency (Cronbach alpha) and a test-retest reliability coefficient (correlations between ATS subscale scores on the first administration in October 1996 and the second administration in October 1997). Mean scores and standard deviations are calculated for both subscales.

Table 4

Overview of the different measures and sample sizes

Year	ATS	Statistics exam result	General exam result
96-97	1st administration (October 1996) (*N* = 264)	1st year (*N* = 234)	1st year (*N* = 234)
97-98	2nd administration (October 1997) (*N*= 127)	2nd year (*N* = 102)	2nd year (*N*= 102)
99-00		3rd year (group work) (*N* = 78)	3rd year (*N* = 78)
00-01		(no statistics course)	4th year (*N* = 74)
01-02		5th year (Dissertation grade) (*N* = 72)	5th year (Courses grade) (*N* = 72)

Note. The number of participants mentioned for the exam results refers to the participants who have a score on the first administration of the ATS as well as on the exams.

The relationships of the attitude scores with the short- and long-term exam results are examined by Pearson product-moment correlation coefficients, separately for statistics and general exam results. The relationships of the attitude scores with all these exam results are compared with the relationships of first year exam results with later exam results. In other words, cognitive and affective predictors of exam results are compared, again separately for statistics and general exam results.

4 Results

4.1 Reliability and mean scores

The alpha estimates are high for both administrations, namely respectively .89 and .91 (*Course* subscale) and .86 and .86 (*Field* subscale). The whole scale alpha estimates are respectively .91 and .89. These results of internal consistency are similar to those mentioned in Section 2. The test-retest reliability analyses show (considerably high) correlations of .62 for the *Field* subscale and .76 for the *Course* subscale. These figures are higher than those reported by Shultz and Koshino (1998) (although for our study the time lapse between the two administrations is longer), but lower than those reported by Wise (1985), where there were only two weeks in between the two measures.

The average *Course* subscale scores for the two administrations are respectively 28.5 (s = 6.4) and 30.7 (s = 6.5), indicating a rather positive attitude toward the statistics course (given that the neutral score is 27). Concerning the attitudes toward the course, our sample of undergraduate Flemish students is comparable with the (higher) graduate student scores observed in other studies. The average *Field* subscale scores, 66.9 (s = 7.6) and 68.0 (s = 6.7), respectively, are also positive (above the neutral score of 60), but compared to the other studies, these scores are low. Finally, the standard deviations of the *Field* and the *Course* subscale scores in our study are lower than in the other studies.

4.2 Relationship between attitudes toward statistics and statistics exam results/dissertation grade

Table 5 presents the correlations of the ATS subscale scores with all statistics exam results. In the last column of this table, we also mention the correlations between the first year statistics exam results and the other statistics exam results. Important to notice is that due to the partly different samples, comparisons must be made carefully. Therefore, we will concentrate on a comparison of correlations where the same students are involved. (The same analyses, restricted to the 72 students who have a measure on all variables, however, revealed the same trends in the data. On request, these data are available from the authors.)

Table 5

Correlations between ATS scores and statistics exam results/dissertation grade

Statistics exam	1st administration			2nd administration			Statistics exam	
	N	Course	Field	*N*	Course	Field	*N*	1st year
1st year	234	0.33***	0.15*	127	0.47***	0.20*	127	1
2nd year	102	0.23*	0.14	115	0.31***	0.20*	115	0.45***
3rd year	78	-0.03	-0.01	88	0.22*	0.07	88	0.26*
5th year (dissertation)	72	0.09	0.04	83	0.03	0.23*	83	0.19

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Because we are carrying out a large number of statistical tests on the same data, we have to take into account that the probability of committing at least one Type I error is substantially larger than the significance level set for each individual test. Multiple correlations are calculated and tested, the ones in Tables 5 and 6, and additional tests are performed to compare correlated correlation coefficients. To avoid potentially spurious results, we perform a Bonferroni correction on the overall significance level (.05). The resulting significance level for an individual test is .001, which means that a *p*-value must be smaller than .001 in order to conclude that the correlation differs from zero.

For the first year, the results show statistically significant (*p*-values <.001) positive correlations between the attitudes toward the course and the statistics exam results. Although the correlations for the *Field* subscale are not statistically significant after the Bonferroni correction, all effect sizes (Cohen 1988, 1992) range between small and medium. The *Course* subscale scores show the highest correlations for both administrations, meaning that for the included sample the attitudes toward the course are a slightly better predictor of the first and second year exam results than attitudes toward the field. The test for comparing correlated correlation coefficients provided by Meng, Rosenthal, and Rubin (1992) shows that the difference between the correlations (*Course* versus *Field*) is statistically significant for the second administration ($Z = 2.46$, $p = 0.01$ for the first administration and $Z = 3.39$, $p < .001$ for the second administration). These results are a compelling replication of the findings from the earlier studies summarized in Table 3.

For the second year, the trends are similar, but differ in terms of statistical significance. The *Course* subscale scores are the most highly related to the second year statistics exams scores. However, only for the second administration the correlation between *Course* and exam results is statistically significant. The test for comparing correlated correlation coefficients shows that the difference between the correlations (*Course* versus *Field*) is not statistically significant for the second year (Z = 1.02, p = 0.31 for the first administration and Z = 1.27, p = .20 for the second administration). The effect sizes (Cohen 1988, 1992) of the correlations still range between small and medium.

For the third year, the attitude scores do not show statistically significant correlations with the statistics exam results. However, recall that we have to be careful with the interpretation of the data from the third year statistics exam results, because they are based on group assessment (see Section 3.2).

In the fifth year, the attitudes scores do not correlate significantly with the dissertation grade, but when we take a closer look at the results, we see that the *Field* subscale scores for the second administration show a substantive (although not statistically significant) correlation with the dissertation grades in the fifth year (r = .23, p = .04).

Furthermore, in contrast to the correlation of the second administration with the first year statistics exam results, where *Course* was related highest to statistics exam results, in the long term, *Field* is more highly related to the dissertation grade than *Course* (test for comparing correlated correlation coefficients: Z = -1.93, p = .05).

Because this study is one of the first to explore the relation between attitudes toward statistics and long-term results, and because of the negative impact that Bonferroni corrections can have on the power of the tests, this correlation between the *Field* subscale and the dissertation grade – although no longer statistically significant after the Bonferroni correction – is worth mentioning and interesting to investigate further.

The last column of Table 5 shows the correlations between the first year statistics exam results and all following statistics exam results (including the dissertation grade). Because these data relate to the same students as those who participated in the second administration of the ATS, the relative predictive values of affective (ATS) and cognitive (first year statistics results) characteristics in predicting later exam results can be compared for that administration.

Not surprisingly, the second year statistics exam results are more highly correlated with the first year exam scores (r = .45, $p < .001$) than with the ATS scores ($r = .31$, $p < .001$ and $r = .20$, $p = .04$ respectively). The test for comparing correlated correlation coefficients shows that this difference between the correlations is most convincing (although not statistically significant after Bonferroni correction) for the *Field* subscale ($Z = -2.28$, $p = .02$).

In the long term, the observed correlation for *Field* ($r = 0.23$, $p = 0.04$) is higher than the correlation between the first year exam results and the dissertation grade ($r = .19$, $p = .08$). Thus for our sample, in the long-term, the *Field* score of the second administration is a better predictor of the dissertation grade than the first year statistics exam result. In other words, the observed affective measure shows a higher correlation with the dissertation grade than the cognitive measure, although the test for comparing correlated correlation coefficients provided by Meng et al. (1992) shows that this difference between the correlations is not statistically significant ($Z = .31$, $p = .76$).

4.3 Relationship between attitudes toward statistics and general exam results

Table 6 presents the correlations of the ATS subscale scores with all general exam results. An inspection of this table reveals that the important role of attitudes toward statistics is specific for statistics achievement (including the dissertation grade). There is no statistically significant correlation between the ATS scores and the short- and long-term general exam results. In our sample, the total grade in the first year is more highly correlated with the following general exam results than the attitude scales.

Table 6
Correlations between ATS scores and general exam results

General exam	1st administration			2nd administration			General exam	
	N	Course	Field	N	Course	Field	N	1st year
1st year	234	0.16*	0.07	127	0.17	0.12	127	1
2nd year	102	0.01	-0.03	115	0.09	0.03	115	0.42***
3rd year	78	-0.01	0.07	88	0.08	0.16	88	0.48***
4th year	74	0.13	0.17	84	0.04	0.13	84	-0.01
5th year (courses)	72	-0.01	0.01	83	-0.04	0.14	83	0.46***

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

5 Discussion

This study provides further insight into students' attitudes toward statistics and into the relationship between these attitudes and (short- and long-term) statistics and general exam results.

First, as in previous studies reported in the first part of this article (see Table 1), we find a high internal consistency for the Attitude Toward Statistics (ATS) scale (Wise 1985). The test-retest reliabilities are fairly high (.62 for the *Field* subscale and .76 for the *Course* subscale) and within the range of the reliabilities reported by previous studies.

Second, this study provides new descriptive data concerning students' attitudes toward statistics. These data are somewhat different from the trends mentioned in the literature. More specifically, the results on the *Course* subscale indicate that Flemish undergraduate students in Educational Sciences have an attitude toward the particular course in which they are enrolled that is more positive than the attitudes of undergraduate students elsewhere, but comparable to the attitudes of graduate students in other studies. However, the analysis of the *Field* subscale scores reveals a relatively negative attitude toward the use of statistics in the students' field of study as compared to the scores from graduate and undergraduate students from the "bench-mark" studies.

Third, the analysis of the relationship between the ATS scores and short-term statistics exam results complemented findings obtained by other authors, namely that especially attitudes toward the course are related to short-term exam results.

Fourth, (although at the border of statistical significance) an innovative element in our study is that it also yields findings concerning the analysis of the relationship between attitudes in the beginning of the curriculum and the dissertation grade. While for short-term exam results, attitudes toward the course are more highly related to statistics exam results than the attitudes toward the field, the latter are more highly related to the fifth-year dissertation grade than the attitudes toward the course. Our results suggest that students who recognize the importance of statistics for their field of study (in the case of the present study: educational sciences) will tend to obtain a better dissertation grade.

Fifth, this study also investigates the relative predictive value of affective (ATS) and cognitive (exam results) measures in predicting later exam results. The data show that the relationship between the attitudes toward the field after experiencing a statistics course (affective measure) and second year statistics exam results were smaller than between first

year exam results (cognitive measure) and those second year exam results. This finding is similar to the findings of Fienberg and Halperin (in Roberts & Bilderback, 1980), namely that a cognitive measure predicted statistics achievement with slightly higher accuracy than the measure of attitudes toward quantitative concepts. However, this difference between the affective and the cognitive measure as predictors is smaller for the relation with the long-term dissertation grade. In fact, the relationship for the affective measure is even slightly (but not significantly) higher than the relationship for the cognitive measure. These results are an important indicator of the essential role attitudes toward statistics (besides cognitive characteristics) play for the development of statistical competence.

Obviously, replications of this research on the relationship between attitudes toward statistics and long term statistics exam results are needed. For instance, a comparison of dissertation scores and other measures that can be used as indications of long-term statistics achievement, such as exam scores and/or scores on more traditional or achievement-based statistical tests problems, can provide a deeper insight into this relationship. Furthermore, it would be very interesting to follow up the non-successful students, under more to compare the attitudes and statistics achievement of these students with the students who did pass the exams.

Finally, results from this study reveal that the important relationship between attitudes toward statistics and statistics achievement is content-specific. Indeed, we found no relationship between the attitudes toward statistics and general exam results. Further research should investigate how the attitudes measured by the ATS differ from “general academic attitudes” and how different attitude scales are related to different kinds of achievement. Such research might reveal the importance of a separate assessment of attitudes toward studying specific fields of study, besides the assessment of “general academic attitudes”.

Appendix: Overview of the studies

Study	Sample	# Adm.	n	(Under)graduate	Field of study	Remarks
Wise 1985	1	1	92		Education	Original article ATS
	2	2	70 70		Education	
Roberts and Reese 1987	1	1	280	Undergraduate		Also administration of another scale to measure attitudes toward statistics, the Statistics Attitude Survey (SAS; Roberts and Bilderback 1980). ATS is treated as one scale in this study.
Waters et al. 1988	1	2	302	Undergraduate	Variety of majors (mainly psychology)	Also administration of SAS. Only 212 respondents were measures on both occasions
Elmore & Lewis 1991	1	2	58	Graduate	Variety of majors	
Elmore, Lewis, & Bay 1993	1	2	289	Undergraduate	Variety of majors	
Shultz & Koshino 1998	1	2	36	Undergraduate	Psychology	
	2	2	38	Graduate	Psychology	
Rhoads & Hubele 2000	1	2	63 61	Undergraduate	Engineering	Used to measure change in attitudes before and after a computer-integrated statistics course
D'Andrea & Waters 2002	1	2	32 17	Graduate	Education	Used to measure change in attitudes before and after a statistics course using 'short stories'
Aldogan & Aseeri 2003	1	1	178	Graduate	Variety of majors	Arabic version
Mvududu 2003	1	1	95	Undergraduate	Variety of majors (USA)	Cross-cultural study (USA and Zimbabwe)
	2	1	120	Undergraduate	Business, Accounting and Economics (Zimbabwe)	Used to measure the relationship between attitudes toward statistics and the use of constructivist strategies.

Chapter 3

Measuring statistics attitudes: Structure of the Survey of Attitudes Toward Statistics (SATS-36)[1]

Abstract

Because students' attitudes are considered very important for effective learning in statistics education, the availability of high-quality instruments to assess this concept is essential. Attitude assessment can provide valuable information to students and instructors, can be used in research on teaching and learning statistics, e.g., to evaluate the effectiveness of different curricula or didactical approaches. Although a number of instruments for assessing attitudes toward statistics have been developed, several questions with regard to the structure and item functioning remain unresolved. In this study, the structure of the Survey of Attitudes Toward Statistics (SATS-36), a widely used survey to measure six aspects of students' attitudes toward statistics, is investigated. Because earlier studies used the technique of item parceling to analyse the factor structure of this survey, individual item functioning was not evaluated before. This study contributes to the existing literature by addressing this remaining issue. Based on confirmatory factor analysis using individual items, the results suggest that the SATS-36 can be improved by removing some poorly functioning items. Furthermore, the results suggest that depending on the goals of a specific study either six subscales could be used or three of them (Affect, Cognitive Competence, and Difficulty) can be combined into one subscale without losing much information.

[1] Vanhoof, S., Kuppens, S., Castro Sotos, A. E., Verschaffel, L., & Onghena, P. (accepted pending revision). Measuring statistics attitudes: Structure of the Survey of Attitudes Toward Statistics. *Statistics Education Research Journal.*

1 Introduction

In recent years, attitudes toward statistics have received increasing attention in statistics education. In statistics education research, attitudes toward statistics are usually broadly defined as a multidimensional concept referring to distinct, but related dispositions pertaining to favourable or unfavourable responses with regard to statistics and statistics learning (Schau, Stevens, Dauphine, & Del Vecchio, 1995; Gal, Ginsburg, & Schau, 1997; Chiesi & Primi, 2009). The importance of attitudes in the context of introductory statistics courses is widely recognized (e.g., Gal et al., 1997; Leong, 2006). Students in the social and behavioural sciences tend to experience a course in statistics as intimidating and/or to feel insufficiently competent in order to acquire the necessary conceptual understanding (Finney & Schraw 2003). Such negative attitudes are often considered a major obstacle for effective learning (Waters, Martelli, Zakrajsek & Popovich, 1988).

In research and in practice it is important to assess dimensions of students' attitudes regarding statistics. Attitude assessment can be used to provide information to students and instructors, or to evaluate the effectiveness of different curricula or didactical approaches. For instance, researchers and teachers believe that if teachers choose challenging activities that promote investigation and are clearly related to everyday life, this can have a positive impact on students' beliefs and attitudes toward statistics (Keeler & Steinhorst, 1995; Shultz & Koshino, 1998; Mills, 2004; Suanpang, Petocz, & Kalceff, 2004; Leong, 2006; Carnell, 2008). Furthermore, attitude information can specifically be used to identify students who are potentially at risk for failing the statistics course. Such identification may be the primary step in assisting them to become successful (Roberts & Saxe, 1982; Cashin & Elmore, 2000).

Evaluation of attitudes toward statistics and their associations with other variables is only possible if proper assessment instruments are available. Such work has already been initiated by a number of researchers (e.g., Roberts & Bilderback, 1980; Schau et al., 1995; Wise, 1985). To improve teaching practice, however, constructing an instrument is not enough. In order to correctly evaluate students' attitudes toward statistics it is essential to evaluate the reliability, validity and possible pitfalls or flaws of an instrument. More specifically, evidence that the presupposed structure of the instrument demonstrates an acceptable fit to the data and that all items measure the underlying constructs of interest should be gathered (Hatcher, 1994). If such evidence is not available, results cannot be unambiguously interpreted.

Although some studies have already been conducted to investigate the structure of the SATS (Schau et al., 1995; Dauphinee, Schau & Stevens, 1997; Hilton et al., 2004; Cashin & Elmore, 2005; Tempelaar, van der Loeff, & Gijselaers, 2007; Chiesi & Primi, 2009), further clarification and investigation is needed. There is some disagreement on the number of factors that best represent the attitudes toward statistics assessed via the SATS. Furthermore, previous studies are limited because combinations of items (parcels; see Section 2) rather than individual items have been used. Therefore, these studies only provide partial insight into the underlying structure of the SATS and the functioning of individual SATS-items.

The goal of the present study is to address some of the above-mentioned remaining issues of previous research on the structure of the Survey of Attitudes Toward Statistics. Specifically, the structure will be investigated using information from individual items rather than parcels. Furthermore, the relative merits of two previously identified factor models (a four- and a six-factor model) will be compared.

This paper is organized in five sections. In Section 2 the SATS is presented and the research goals are introduced in detail and related to existing studies on the structure of the SATS. Advantages and disadvantages of using parcels and items for confirmatory factor analysis are also discussed. In Section 3 the methodology of this study is described. In Section 4, results are presented, followed, in Section 5, by a discussion.

2 Literature review

2.1 The Survey of Attitudes Toward Statistics (SATS)

The Survey of Attitudes Toward Statistics (SATS; Schau et al., 1995) was developed to assess students' attitudes toward statistics. The SATS is a Likert-type survey with seven response possibilities for each statement ranging from "strongly disagree" to "strongly agree". The first version of the SATS consisted of four subscales: (a) *Affect* (six items): positive and negative feelings concerning statistics; (b) *Cognitive Competence* (six items): attitudes about intellectual knowledge and skills applied to statistics; (c) *Difficulty* (seven items): attitudes about the difficulty of statistics as a subject, and (d) *Value* (nine items): attitudes about the usefulness, relevance, and worth of statistics in personal and professional life. Afterwards (Schau, 2003; Schau, personal communication, September 29,

2005), two subscales were added to the survey: *Interest* (four items), students' level of individual interest in statistics and *Effort* (four items), the amount of effort students spend on learning statistics. Depending on the number of subscales and corresponding items, the developers labelled the survey as SATS-28 and SATS-36. Additionally, two versions are available; one to administer before (SATS-28-pre/SATS-36-pre) and one to administer after a statistics course (SATS-28-post/ SATS-36-post). These two versions are identical, except for tense.

A complete version of the SATS-36 and detailed scoring information can be consulted online via http://www.evaluationandstatistics.com/index.html.

2.2 Structure of the SATS

In one study on the most recent version, namely SATS-36, the presupposed six-factor solution was supported by confirmatory factor analysis (Tempelaar et al., 2007). However, the factors *Affect* and *Cognitive Competence* appeared to be very strongly correlated, and the *Difficulty* scale was moderately to strongly correlated with these two subscales. These findings coincide with empirical studies on the SATS-28 (Chiesi & Primi, 2009; Dauphinee et al., 1997; Hilton et al., 2004; Cashin & Elmore, 2005; Schau et al., 1995). Correlations ranged between 0.80 and 0.94 for *Affect* and *Cognitive Competence*, between 0.57 and 0.73 for *Affect* and *Difficulty*, and between 0.46 and 0.64 for *Cognitive Competence* and *Difficulty*. In all studies except the one by Cashin and Elmore (2005) these related constructs were represented as three distinct latent factors.

Dauphinee et al. (1997) explicitly compared the original four-factor model of the SATS-28 to a three-factor model that combined *Affect* and *Cognitive Competence*. They concluded that the factors should remain distinct because: (1) the four-factor model fitted better, (2) the two factors operated differently in terms of course completion, and (3) it is important in statistics education to have a distinct construct (i.e., *Cognitive Competence*) that corresponds with Mathematics Self-Concept in the area of mathematics education.

Conversely, based on one exploratory factor analysis on the SATS-28 conducted by Cashin and Elmore (2005), a more parsimonious solution was suggested with *Affect*, *Cognitive Competence* and *Difficulty* combined into one factor. They argued that the SATS-28 may only pertain to two underlying factors, namely (1) the value of statistics as a tool in students' respective fields of study (*Value*) and (2) different aspects of how a student will

perform in his or her statistics course (measured by the *Affect*, *Cognitive Competence*, and *Difficulty* items). Interestingly, these two dimensions correspond to the two subscales (*Field* and *Course*) of the Attitudes Toward Statistics (ATS) scale (Wise, 1985), another instrument that measures students' attitudes toward statistics. As was the case in the other studies on the structure of the SATS, in the study of Cashin and Elmore (2005) especially the relationship between *Affect* and *Cognitive Competence* was pronounced and meaningful as both subscales relate to feelings concerning the specific course, skills, or personal capabilities to complete the coursework. In contrast to their prior expectations *Difficulty* did not form a separate factor in their exploratory factor analysis. Hence they suggested further research to investigate the factor structure of the SATS and the relationship between *Affect*, *Cognitive Competence* and *Difficulty* in more detail.

Theoretically, the more parsimonious *Course-Field* difference for the SATS-28 relates to the distinction often made in attitude research between students' attitudes about a specific domain (i.e., the value of statistics) and their attitudes about themselves as learners of a domain (i.e., affect, self-efficacy and perceived difficulty regarding statistics) (e.g., see Gal & Ginsburg, 1994; McLeod, 1992). For instance, students' attitudes about a specific domain are generally considered resistant to change, whereas their attitudes about themselves as learners of a domain are more likely to change depending on changing circumstances during the progress of students' curriculum (Gal & Ginsburg, 1994).

2.3 Parcel versus item-level confirmatory factor analysis

As already stated in the introduction, all available confirmatory factor analytic studies of the SATS have performed analyses on item parcels rather than individual items (Tempelaar et al., 2007; Chiesi & Primi, 2009; Dauphinee et al., 1997; Hilton et al., 2004; Schau et al., 1995). An item parcel refers to a simple sum or mean of several items from the same factor. It is assumed that the items of a parcel assess the same construct and that they are psychometrically unidimensional (Bandalos, 2002; Nasser & Wisenbaker, 2003; Hau & Marsh, 2004). In such analyses, parcels are treated as continuous indicators (Kline, 2005).

The main reasons for using the technique of item parceling in the context of SATS data are: "to improve reliability" (Schau et al., 1995, p. 872; Dauphinee et al., 1997, p. 133), "to avoid inherent non-normality associated with single item distributions" (Hilton et al., 2004, p. 97), and "to reduce the number of model parameters to achieve a more attractive

variable to sample size ratio, and to get more stable parameter estimates" (Tempelaar et al., 2007, p. 85).

Although the technique of item parceling has its advantages, it remains controversial. Parceling might be seen as "tricky" because modelled data should resemble the observed responses as much as possible (Little, Cunningham, Shahar & Widaman, 2002). In this sense, parceling introduces a potential source of subjective bias (Little et al., 2002), especially because several methods of parceling are available and the choice of the method can affect the results (Kline, 2005). Furthermore, the assumption of unidimensionality within parcels is often not investigated or even not met (Bandalos & Finney, 2001). If a set of items assigned to the same parcel is not unidimensional, analysis of the aggregate score across the items may be meaningless (Kline, 2005). Moreover, in some instances parceling can mask a multidimensional factor structure which may lead to a seriously misspecified CFA-model fitting the data reasonably well (Bandalos, 2002).

Besides these general disadvantages of the technique of item parceling, specific problems pertain to the parceling schemes used in previous research on the SATS. First, in the parceling scheme of Schau et al. (1995), Dauphinee et al. (1997), and Hilton et al. (2004) the *Affect*, *Cognitive Competence*, and *Difficulty* factors comprised only two parcels, while a minimum of three has been suggested (Nasser & Wisenbaker, 2003; Hau & Marsh, 2004). Second, in the parceling scheme of Tempelaar et al. (2007), there were exactly three parcels per factor, but some parcels necessarily contained only one item because there were only four *Interest* and *Effort* items.

As the technique of item parceling may jeopardize a good understanding of the true factor structure of the SATS-36 items (Bandalos, 2002; Bandalos & Finney, 2001), the present study performed a confirmatory factor analysis on individual items of the SATS-36 for the first time using statistical approaches for categorical item-indicators (see Section 3.3).

Two main research questions on the structure of the SATS-36 will be investigated using confirmatory factor analysis on the individual items:

1. Can the six-factor structure be confirmed for the SATS-36?
2. Is the six-factor structure preferable to a four-factor structure that represents *Affect*, *Cognitive Competence*, and *Difficulty* as a single factor?

The answers to these two questions are important to guide interpretation of the survey not only in research but also in teaching. For instance, if *Affect*, *Cognitive Competence*, and *Difficulty* measure the same construct or show similar correlations with other SATS-factors, a more parsimonious interpretation of the survey may be preferred.

3 Method

3.1 Participants

Participants are 514 first year Educational Sciences (321 female, 22 male) and Speech Pathology and Audiology (163 female, 8 male) students from two cohorts of an introductory undergraduate statistics course at the Department of Educational Sciences of the Katholieke Universiteit Leuven. As the numbers indicate, students in Educational Sciences and Speech Pathology and Audiology are mainly female. Most participants took limited to moderate mathematics-oriented programs in secondary education. Specifically, 154 participants followed programs including one, two or three weekly hours of mathematics, 334 participants followed programs including four, five or six weekly hours of mathematics and 26 students followed programs including seven or eight weekly hours of mathematics. 446 participants indicated that an introduction to statistics was part of their secondary school education.

The introductory statistics course was taught during the first semester of the students' first academic year. In general, the course dealt with some introductory methodological and statistical concepts (such as tables, figures, and descriptive statistics), but not with formal probability theory or statistical inference. The mathematical background required for following the course was limited.

The data were collected at the very beginning of this first year statistics course. For one cohort this occurred in September 2005, for the other cohort in September 2006. The instrument was completed voluntarily and handed in during class time. It was stressed that the data would be analysed anonymously.

3.2 Measures

Attitudes toward statistics were assessed with a Dutch translation of the pre-test version of the SATS-36 (Schau et al., 1995). Negatively formulated items were reversed to

assure that a high score equals a positive attitude. Like Tempelaar et al. (2007), the focus was on students' attitudes when entering university (for results from data: see Appendices).

The translation from English into Dutch took place in August/September 2005, using the following procedure. First, the survey was translated by the first author and by an expert translator with a Master's degree in German languages. Afterwards, both translations were compared and differences and possible ambiguities were discussed. This resulted in a Dutch version of the (pre-test version of the) SATS-36. Second, this translation was validated using a back-translation technique (Brislin, 1970). Statistics experts translated the items of the Dutch version back into English. Afterwards, the quality of the translation was judged by comparing this version with the original English version. Differences and possible ambiguities were discussed. Third, the Dutch version of the SATS-36 was administrated to a small number of people (n = 6) with a diverse statistical background. Participants were instructed to write down suggestions when filling out the survey. The comments did not point out any problems with the wording, merely some minor comments on the punctuation, so no changes to the translation of any item were made.

3.3 Statistical analyses

Confirmatory factor analysis

Confirmatory factor analyses (CFA) were performed using the software Lisrel 8.70 (Jöreskog & Sörbom, 2004) to test the proposed factor models for the SATS-36.

The ordinal nature of the SATS-36 items had to be respected and important assumptions (such as multivariate normality) of the techniques used had to be investigated. Two Robust Maximum Likelihood (RML) estimation techniques were performed: RML for ordinal data with polychoric correlations (Jöreskog, 1993) and RML with covariances (e.g., Kline, 2005). RML was preferred to weighted least squares estimation (WLS), because the latter requires a very large sample size (e.g., Flora & Curran, 2004; Kline, 2005; Finney & DiStefano, 2006) and more easily results in convergence problems or improper solutions as model complexity increases (Boomsma & Hoogland, 2001).

RML is preferred to standard ML for both polychoric correlations and covariances to correct for non-normality in the data. When deviation from multivariate normality is present and data are categorical, parameter estimates, chi- square statistics and standard errors

tend to be biased (Kline, 2005; Finney & DiStefano, 2006). If this is the case, RML, such as using a Satorra-Bentler adjustment of the chi-square statistic, is recommended to adjust estimations based on the degree of nonnormality (Satorra & Bentler, 1994; Kline, 2005; Finney & DiStefano, 2006).

First, polychoric correlations were analysed. A polychoric correlation estimates what the Pearson correlation coefficient between two ordinal variables would be if both were continuous and normally distributed in the population (Kline, 2005). The estimated underlying continuous population variables are then used in the confirmatory factor models, instead of the observed variables (Jöreskog, 2005).

Second, the covariance matrix was analysed, treating the observed variables as continuous. It has been shown that such an analysis combined with Satorra-Bentler adjustment produces sufficiently accurate parameter estimates for Likert scales with more than five response categories under conditions of non-normality (Finney & DiStefano, 2006). Because the SATS-36 is based on a seven-point Likert scale, this technique seems appropriate.

A major advantage of analyses based on covariances as compared to polychoric correlations is that covariance results allow interpretation of factor loadings and other estimates based on actual results or raw data. Therefore, discussion of the parameter estimates will be based on results from analysis of covariances when similar results are produced by both approaches (covariances and polychoric correlations). To cross-check results regarding the comparison of the six- and four-factor model, analyses on item parcels were also performed. The results of these analyses are shown in Appendix I, together with the results from polychoric correlations that corroborated results from covariances. The results shown in the manuscript are based on RML with covariances.

Details about data and assumptions

Before describing the decision rules and fit statistics used, details about the data and underlying assumptions of the techniques are investigated. When assumptions are violated, biased results may occur in terms of model fit, parameter estimates, and related significance tests (Schumacker & Lomax, 2004; Finney & DiStefano, 2006).

Summary statistics (means, standard deviations, kurtosis, skewness) for the items, polychoric correlations and covariance matrices are available upon request from the first

author. Tests of model assumptions (SSICentral, s.d.; Jöreskog, 1999) on the items showed deviations from multivariate normality (Skewness $Z = 21.749$, $p < 0.001$; Kurtosis $Z = 399.592$, $p < 0.001$; Skewness and Kurtosis $\chi^2 = 303.972$, $p < 0.001$). All items, except item 5 (*Cognitive Competence* 1), item 10 (*Value* 3), item 6 (*Difficulty* 1) and item 30 (*Difficulty* 5), show significant deviations from univariate normality. However, when inspecting the size of skewness and kurtosis, only item 27 (*Effort* 4) showed substantive deviation from normality (Skewness = -1.892; Kurtosis = 4.823). Because there were indications of multivariate non-normality, a Satorra-Bentler adjustment was performed (Satorra & Bentler, 1994).

The assumption of underlying bivariate normality is required to analyse polychoric correlations. Based on the RMSEA-values for population discrepancy (Jöreskog, 2005), no violations of this assumption were observed.

Global model fit

Satorra-Bentler-scaled chi-square statistics (*SBSχ^2*; Satorra & Bentler, 1994) were obtained during the CFA-analyses to assess the magnitude of discrepancy between the sample and fitted matrices (Hu & Bentler, 1999). As mentioned earlier, the Satorra-Bentler-scales chi-square statistic corrects for non-normality in the data. However, it is widely known that the chi-square-based statistics are very sensitive to sample size (e.g., Kline, 2005). This may result in the rejection of reasonable models because in the presence of large sample sizes, small degrees of lack of fit already result in small *p*-values (Byrne, 1989; Hu, Bentler, & Kano, 1992). For this reason additional goodness-of-fit indices were used to evaluate model fit: Root Mean Square Error of Approximation (*RMSEA*), Comparative Fit Index (*CFI*), and Non-normed fit index (*NNFI*). It has been suggested that a value of the *RMSEA* of less than .05 is an indication of a good fit whereas values between .05 and .08 still show a reasonable fit of the model. The indices *NNFI* and *CFI* normally range between zero and one, with higher values indicating a better fit. As a benchmark for good fit, the value 0.90 is often used (Hoyle & Panter, 1995; Hu & Bentler, 1999; Schumacker and Lomax, 2004).

Because one of the goals of this study was to compare the presupposed six-factor structure model of the SATS-36 to a four-factor model where *Affect*, *Cognitive Competence* and *Difficulty* are combined in one factor, additional fit statistics were inspected to assess the relative fit of these two nested models.

The six- and four-factor models were compared by means of the scaled chi-square difference test (scaled-$\chi^2\Delta$; Satorra & Bentler, 2001). However, because also this significance test is sensitive to relatively small deviations when sample size is large, the Akaike Information Criterion (*AIC*; Wang & Liu 2006) and the Bayesian Information Criterion (*BIC*; Wang & Liu 2006) were additionally used to compare models. The *AIC* and *BIC* take into account both model fit and model complexity. When comparing two models, the model with the lowest *AIC* and *BIC* is the preferred one (Jöreskog, 1993; Kline, 2005). *RMSEA*, *CFI* and *NNFI* values will be compared only exploratory because no formal guidelines for meaningful changes exist to the best of our knowledge. In the context of measurement invariance Cheung and Rensvold (2002) suggested that a difference in *CFI* equal to or greater than .01(in combination with other appropriate fit statistics) indicates a meaningful difference (see Chapter 4).

Local model fit

Because this is the first study that examined the factor structure of the SATS using confirmatory factor analysis on the individual items, we considered evaluating local model fit to be important. It is possible for a model to be misspecified in some parts but very well specified in other parts (Jöreskog, 1993). For the presented models, size and significance of the factor loadings, standardized residuals and modification indices are reviewed (e.g., Kline, 2005). Also correlations between latent factors are checked and discussed especially to investigate whether *Affect*, *Cognitive Competence* and *Difficulty* show similar or different correlations with the other SATS-36 factors.

Although standardized residuals show which relationships are not properly explained, they do not indicate how the model should be modified to fit the data better (e.g., Kline, 2005). Therefore, modification indices rather than standardized residuals were used to guide model modification. Clear guidelines or cut-off values regarding modification indices are not available. The best option is to initially consider the modification indices with the highest values. In addition, substantive and theoretical arguments were used to guide the modification of the models in order to avoid the risk of capitalizing on chance and building models that do not generalize to other samples or to the population (Jöreskog, 1993; Hatcher, 1994; Hoyle & Panter, 1995; Schumacker & Lomax, 2004; MacCallum, Roznowski & Necowitz, 1992).

Items with factor loadings below 0.40 (e.g., Hatcher, 1994) were considered for deletion from the SATS-36, because such items may not sufficiently relate to the expected underlying construct. As will be discussed in detail later, cross-validation of such modifications is needed.

4 Results

4.1 Six-factor solution

As presented in Table 1, adequate fit indices were obtained for the hypothesized six-factor SATS-36 model (Model 1). Associations among the latent factors for Model 1 are shown in Table 2. Note that the *Difficulty* factor should be interpreted as a Lack of *Difficulty*. As reported in Table 2, *Affect*, *Cognitive Competence*, and *Difficulty* were highly correlated (*r* between 0.844 en 0.888).

Table 1

Fit indices for the models tested based on the covariance matrix

Model	*SBS*χ^2	*df*	*RMSEA*	*NNFI*	*CFI*	*BIC*	*AIC*
6-factor original (1)	1607.30	579	0.059	0.94	0.95	2150.37	1781.30
6-factor modified (2)	1136.24	479	0.052	0.96	0.96	1648.10	1300.24
4-factor original (3)	1655.55	588	0.060	0.94	0.94	2142.44	1811.55
4-factor modified (4)	1209.47	488	0.054	0.96	0.96	1665.15	1355.47

Note. *SBS*χ^2= Satorra-Bentler-scaled Chi-square; *df* = Degrees of Freedom; *RMSEA* = Root Mean Square Error of Approximation; *NNFI* = Non-Normed Fit Index; *CFI* = Comparative Fit Index; *BIC* = Bayesian Information Criterion; *AIC* = Akaike Information Criterion

Table 2

Estimated latent factor correlations for the six-factor models

Model 1: six-factor original	Affect	Cognitive Competence	Difficulty	Value	Interest
Cognitive Competence	0.888***				
Difficulty	0.844***	0.855***			
Value	0.442***	0.431***	0.370***		
Interest	0.575***	0.484***	0.476***	0.715***	
Effort	-0.088	-0.120*	-0.221***	0.165**	0.200***
Model 2: Six-factor modified	**Affect**	**Cognitive Competence**	**Difficulty**	**Value**	**Interest**
Cognitive Competence	0.883***				
Difficulty	0.848***	0.860***			
Value	0.393***	0.432***	0.355***		
Interest	0.484***	0.483***	0.470***	0.715***	
Effort	-0.135*	-0.119*	-0.232***	0.164**	0.201***

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

To explore differences between *Affect*, *Cognitive Competence*, and *Difficulty*, their differential association with other SATS-factors were explored. *Affect*, *Cognitive Competence*, and *Difficulty* showed similar correlations with *Value* (resp. $r = 0.44$, $r = 0.43$, $r = 0.37$) and *Interest* (resp. $r = 0.58$, $r = 0.48$, $r = 0.48$). The correlations with *Effort* showed more diversity ($r = -0.09$, $r = -0.12$, $r = -0.22$), with *Difficulty* (perceived easiness) most negatively related to *Effort*. In other words, the more that students perceived statistics to be easy, the less *Effort* they expected to spend on learning statistics. Students' affect and competence seemed less associated with the expected amount of *Effort* compared to the associations with *Difficulty*, *Value* and *Interest*. Note that a relatively high correlation was estimated between *Interest* and *Value* ($r = 0.715$, $p < 0.001$).

Parameter loading estimates for all items are shown in the third column of Table 3. All factor loadings except for three *Difficulty* items were above 0.40 and significant. In a modified six-factor model the three items with factor loadings below 0.40 (item 22, item 34 and item 36) were deleted. Two of these items (item 22 and item 36) ask about most people's attitudes regarding the *Difficulty* of statistics, rather than – like the other *Difficulty*

items – students' own attitudes regarding the *Difficulty* of the field of statistics. This may explain the poor functioning of these two items.

Table 3

Parameter estimates of the models tested

		6-factor model	6-factor modified		4-factor model	4-factor modified
Affect	Item 3	0.650	0.569		0.619	0.584
	Item 4	0.585	0.616		0.583	0.595
	Item 15	0.685	0.714		0.670	0.678
	Item 18	0.659	0.690		0.633	0.641
	Item 19	0.624	0.532		0.577	0.535
	Item 28	0.749	0.777	**Integration**	0.738	0.742
Cognitive Competence	Item 5	0.705	0.711	**of**	0.676	0.686
	Item 11	0.544	0.542		0.523	0.522
	Item 26	0.546	0.547	***Affect*,**	0.541	0.546
	Item 31	0.477	0.478		0.448	0.450
	Item 32	0.608	0.609	***Cognitive***	0.579	0.582
	Item 35	0.827	0.823	***Competence***	0.794	0.796
Difficulty	Item 6	0.615	0.624	**and**	0.588	0.57
	Item 8	0.651	0.649		0.582	0.579
	Item 22	0.338	Deleted	***Difficulty***	0.277	Deleted
	Item 24	0.523	0.514		0.456	0.458
	Item 30	0.438	0.426		0.402	0.402
	Item 34	0.246	Deleted		0.248	Deleted
	Item 36	0.341	Deleted		0.320	Deleted
Value	Item 7	0.544	0.544	***Value***	0.544	0.544
	Item 9	0.570	0.571		0.568	0.568
	Item 10	0.512	0.513		0.510	0.510
	Item 13	0.501	0.501		0.500	0.500
	Item 16	0.620	0.619		0.621	0.621
	Item 17	0.567	0.567		0.570	0.570
	Item 21	0.424	0.424		0.424	0.424
	Item 25	0.519	0.519		0.517	0.517
	Item 33	0.617	0.617		0.618	0.618
Interest	Item 12	0.592	0.592	***Interest***	0.591	0.591
	Item 20	0.822	0.819		0.820	0.819
	Item 23	0.850	0.853		0.852	0.853
	Item 29	0.844	0.843		0.845	0.844
Effort	Item 1	0.720	0.718	***Effort***	0.720	0.719
	Item 2	0.749	0.750		0.747	0.747
	Item 14	0.783	0.783		0.784	0.784
	Item 27	0.700	0.699		0.700	0.700
Error covariance between item 3 and item 19			0.410			0.400

Besides dropping three *Difficulty* items with low factor loadings, inspection of the modification indices shows that another substantial justifiable improvement can be made. Including an error covariance between item 3 and item 19 (the first and fifth *Affect* items) resulted in a substantial decrease in chi-square. The existence of this error covariance was not surprising as both items were the only two positively formulated items of the *Affect* factor which also share closely related meanings (i.e., item 3 "I will like statistics"; item 19 "I will enjoy taking statistics courses"). In other words, there seems to be a unique association (method or content similarity) between these two items that is not accounted for by the latent factor.

The modified six-factor model (Table 1, Model 2) provided a considerable better fit to the data compared to the hypothesized six-factor model (Table 1, Model 1), as indicated by a significant scaled-chi-square difference test (scaled-$\Delta\chi^2(100) = 403.23$, $p < 0.001$). There was also a substantial decrease in *BIC* and *AIC* values in favour of the modified model (ΔBIC = 502.27, ΔAIC = 481.06). *RMSEA*, *CFI* and *NNFI* values were all in favour of the modified model, although the differences were small.

In the modified six-factor model all parameter estimates were above 0.40 (see fifth column of Table 3) and no additional substantial and theoretically plausible changes seemed reasonable based on the modification indices. Inspection of the correlation matrix between the latent factors for Model 2 (Table 2) shows – as was the case for Model 1 – high correlations between *Affect*, *Cognitive Competence* and *Difficulty* (all correlations greater than 0.84).

4.2 Four-factor solution

Table 1 also shows the fit indices of the four-factor model where *Affect*, *Cognitive Competence*, and *Difficulty* were combined into one factor (Model 3). As was the case for the six-factor model, the absolute model fit was adequate.

Again, the same three *Difficulty* items (item 22, item 34 and item 36) had factor loadings below 0.40 (Table 3) and modification indices indicated a substantial improvement in fit by allowing an error covariance between item 3 and item 19. In the modified four-factor model (Model 4), the three *Difficulty* items were deleted from the model and the error covariance between item 3 and item 19 was included.

The modified Model 4 performed significantly better than the hypothesized four-factor Model 3 (scaled-$\Delta\chi^2$ (100) = 391.65, $p < 0.001$) and there was a considerable impact on the *BIC* and *AIC* values in favour of Model 4 (ΔBIC = 477.29, ΔAIC = 345.77). The correlation structure of the latent factors for Model 3 and Model 4 are shown in Table 4.

Table 4

Estimated latent factor correlations for the four-factor models

Model 3: four-factor original	Integration of *Affect*, *Cognitive Competence*, and *Difficulty*	*Value*	*Interest*
Value	0.445***		
Interest	0.540***	0.715***	
Effort	-0.32**	0.164**	0.201***
Model 4: four-factor modified	**Integration of *Affect*, *Cognitive Competence*, and *Difficulty***	***Value***	***Interest***
Value	0.426***		
Interest	0.509***	0.715***	
Effort	-0.147**	0.164**	0.201***

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

4.3 Four- versus six-factor solution

A formal comparison of the Satorra-Bentler Chi-Square values of the two original models, Model 1 and Model 3, indicated that the six-factor model performed better (scaled-$\Delta\chi^2$ (9) = 49.15, $p < 0.001$). The difference in *BIC* was small (ΔBIC = 7.93, lower than 10; Raftery, 1995) and in favour of the four-factor model (Model 3). The difference in *AIC* was also small, but in favour of the six-factor model (Model 1) (ΔAIC = -30.25).

When the modified models are compared, the results are all in favour of the six-factor model (scaled-$\Delta\chi^2$ (9) = 70.85, $p < 0.001$; ΔBIC = -17.05; ΔAIC = -55.23).

The four-factor model was based on high correlations between three factors. To further investigate differences between *Affect*, *Cognitive Competence* and *Difficulty*, mean observed scores were compared. The mean scores were 3.63 (*Affect*), 4.22 (*Cognitive Competence*), and 3.36 (*Difficulty*). Although all differences were smaller than one point on

the Likert-scale, paired t-tests revealed that all differences were statistically significant ($p < 0.001$).

Although very high correlations are observed between *Affect*, *Cognitive Competence* and *Difficulty* in this study, a five-factor model with only *Affect* and *Cognitive Competence* combined was also tested. The reason to do this was that in earlier studies *Affect* and *Cognitive Competence* were clearly more related than *Affect* and *Difficulty* or *Cognitive Competence* and *Difficulty* (see Section 2.2). Because a very high correlation (r = 0.88) remained between the combined *Affect-Cognitive Competence* and *Difficulty scales*, this solution was not explored further.

5 Discussion

Results for the confirmatory factor analyses of the original six-factor model of the SATS-36 and the alternative four-factor model are discussed in the first section. The second section considers conditions when one model is preferred to the other. The last section presents some limitations of the present study and suggestions for further research.

5.1 Six- and four-factor model for the SATS-36

In an absolute sense, the predefined six-factor structure of the SATS-36 could not be falsified in this study, which corroborates Tempelaar et al.'s (2007) study using parcels. In addition, similar to Cashin and Elmore (2005), a four-factor structure which integrated *Affect*, *Cognitive Competence*, and *Difficulty* also showed adequate absolute properties. As such, an important conclusion of this study is that both the six- and four-factor models appropriately describe the observed interrelationships between SATS-36-items. However, a closer investigation of item functioning suggested that some model modifications are in order.

First, as reflected by low factor loadings, several *Difficulty* items (item 22, item 34 and item 36) should be deleted from the survey because they show less common variance with other *Difficulty* items. For item 22 and item 36 this may relate to the fact that both items refer to how statistics is perceived by most people, whereas other *Difficulty* items pertain to students' attitudes toward the field of statistics as such. The reason for the poor functioning of item 34 remains unclear and should be further investigated.

Second, item 3 and item 19 seemed to have more in common than represented by the *Affect* factor. This unique association likely refers to method and content similarity because they are the only two positively formulated items of the *Affect* factor and share closely related meanings. In this study – to take this unique association and the individual contribution of both items to *Affect* into account – an error covariance between the two items was added to the CFA model. A more practical solution would have been to delete one of these items without much loss of information because of the similarity in meaning (see for example John & Soto, 2007).

Based on these results we argue that – when assessing students' attitudes toward statistics – statistics teachers and researchers should take these improvements to the SATS-36 into account. However, caution is needed regarding this conclusion. Because this is the first study to analyse individual items, cross-validation of the modified models is essential to answer the question whether the suggested modifications generalize to other samples or to the population (MacCallum et al., 1992; Hoyle & Panter, 1995). Some cross-validation data are presented in Appendix II.

5.2 Comparison of six- and four-factor model

Results from the explicit comparison of both models were not straightforward. From a technical viewpoint, the original six-factor model demonstrated a better fit than the four-factor model, but the difference was very small when the extra model complexity of the six-factor modelled was penalized (cf. *BIC* fit index) or when *RMSEA*, *CFI* and *NNFI* were considered. From a more substantive viewpoint, several reasons can be formulated that favour the original six-factor version of the SATS-36 over the more parsimonious four-factor model. First, it will be easier to compare results to earlier studies. Second, in this study mainly correlations were considered, but it is still possible that mean scores for *Affect*, *Cognitive Competence*, and *Difficulty* differ considerably. In our study, all differences were significant, with a minor difference between *Affect* and *Difficulty* (3.63 versus 3.36) and a substantial difference between *Cognitive Competence* and *Difficulty* (4.22 versus 3.36). Furthermore, it can be important to situate individual students on all subscales; it is possible that an individual student has a relatively low score (for instance compared to the class average) on one of these subscales, and at the same time a relatively high score on another

subscale. Third, in line with previous research, differential associations emerged between these three factors and other SATS-factors.

In the absence of a univocal preference for one structure, the choice may depend on the goals of a specific study or specific educational setting. Researchers or teachers who require a more global description of students' attitudes toward statistics, may employ the more parsimonious subscale structure when using the SATS-36 (note that this does not mean that students have to answer less items). In such case, in line with the distinction formulated by McLeod (1992) and Gal and Ginsburg (1994) (see also Section 2.2), the combination of the *Affect*, *Cognitive Competence* and *Difficulty* subscales represent students' attitudes about themselves as learners of statistics (or as formulated by Cashin and Elmore (2005): different aspects of how a student will perform in his or her statistics course), as opposed to attitudes about the statistics field itself (i.e. *Value* and *Interest*).

In other instances, a more detailed picture of students' attitudes on statistics may be required. For example, when an examination of associations among the subscales or with others measures is at the forefront, it may be more informative to include the six constructs because a differential pattern may otherwise go unnoticed.

5.3 Limitations and suggestions for further research

The main limitations of the present study are the homogeneity of the sample (Educational Sciences and Speech Pathology and Audiology students; mainly female participants) and the fact that data from only one statistics course and one administration were considered. Future longitudinal studies including students from other academic fields and other statistics courses are needed to further validate the SATS-36 and generalize the present findings. We are already partly fulfilling this need by presenting some cross-validation data in Appendix II.

Specifically, the authors are looking forward to new studies on the differences in mean scores and on the differential relationships between *Affect*, *Cognitive Competence* and *Difficulty* on the one hand and other SATS factors or external variables on the other hand. Also further examination of the relatively high association between *Interest* and *Value* that was observed in this study would be interesting. Such studies would contribute to an understanding of the similarities and differences between the three SATS factors that are highly correlated. Despite these limitations, the results from the present study clearly

addressed important questions regarding the measurement of attitudes toward statistics and showed the additional value of analyzing individual items rather than item parcels.

Appendix I: Results based on item parceling and polychoric correlations

I.I Results based on item parceling

The parceling scheme employed by Tempelaar et al. (2007), which was guided by counterbalancing skewness, was adopted when analyzing item parcels. This resulted in ten parcels with three items, four parcels with two items, and four parcels with one item. Parcels with one item were created because there were only four *Interest* and *Effort* items in the SATS-36.

Tests of model assumptions of the parcelled data showed deviations from multivariate normality (Skewness and Kurtosis X^2 = 303.972, $p < 0.001$). One *Difficulty* item parcel exceeded the size of 1 for kurtosis and one *Effort* item parcel exceeded the size of 1 for both skewness and kurtosis. Because of these deviations from multivariate normality, asymptotic covariance matrices were used, which resulted in Satorra-Bentler-scaled chi-square values.

Table I shows the fit indices of the CFA-models based on item parceling. The hypothesized six-factor model of the SATS-36 fitted the data quite well as indicated by appropriate goodness-of-fit indices. The model that combined *Affect*, *Cognitive Competence*, and *Difficulty* into one latent factor performed significantly worse than the original six-factor model, as indicated by a significant scaled-chi-square difference test (scaled-$\Delta\chi^2(9) = 110.62$, $p < .001$) and by substantially lower *AIC* and *BIC* values for six-factor model (*ΔBIC* = 58.21, *ΔAIC* = 96.39). Differences in *RMSEA*, *NNFI* and *CFI* were small, however.

Table I

Fit Indices for the Six- and Four-Factor Model of the SATS-36 based on Item Parceling

Model	*SBS*χ^2	*Df*	*RMSEA*	*NNFI*	*CFI*	*BIC*	*AIC*
6-factor model	278.70	120	0.051	0.98	0.98	597.05	380.70
4-factor model	393.09	129	0.063	0.97	0.97	655.26	477.09

Note. *SBS*χ^2 = Satorra-Bentler-scaled Chi-square; *df* = Degrees of Freedom; *BIC* = Bayesian Information Criterion; *AIC* = Akaike Information Criterion; *RMSEA* = Root Mean Square Error of Approximation; *NNFI* = Non-Normed Fit Index; *CFI* = Comparative Fit Index.

The correlational structure of the latent factors for the six-factor model is shown in Table II. *Affect*, *Cognitive Competence*, and *Difficulty* were highly, although not extremely highly, correlated ($r_{Affect,CognitiveCompetence}$ = 0.85, $p < 0.05$; $r_{Affect,Difficulty}$ = 0.78, $p < 0.05$; $r_{CognitiveCompetence,Difficulty}$ = 0.78, $p < 0.05$).

Table II

Estimated latent factor correlations for Model 1

Model: 6-factor original	Affect	Cognitive Competence	Value	Difficulty	Interest	Effort
Affect	1.00					
Cognitive Competence	0.85*	1.00				
Value	0.42*	0.42*	1.00			
Difficulty	0.78*	0.78*	0.33*	1.00		
Interest	0.58*	0.49*	0.70*	0.43*	1.00	
Effort	-0.08	-0.09	0.17*	-0.22*	0.20*	1.00

* $p < 0.05$

I.II Results based on polychoric correlations

The assumption of underlying bivariate normality was required when using polychoric correlations. Based on the *RMSEA*-values for population discrepancy (Jöreskog, 2005), no violations of this assumption were observed.

As presented in Table III, adequate fit indices were obtained for the original six-factor model (Model I). However, in a modified version of this model (Model II) the error covariance between item 3 and item 19 (the first and fifth *Affect* items) was allowed, because it resulted in a substantial improvement of model fit. Model II provided a considerable better fit to the data compared to Model I (Table III), as indicated by a significant scaled-chi-square difference test (scaled-Δ-$\chi^2(1)$ = 268.62, $p < .001$ and a a substantial decrease in *BIC* and *AIC* values in favour of the modified model (ΔBIC = 262.38, ΔAIC = 266.65). Small improvements in terms of *RMSEA*, *NNFI* and *CFI* were observed for the modified model.

Table III

Fit indices for models based on the polychoric correlation matrix

Model	*SBS*χ^2	*df*	*RMSEA*	*NNFI*	*CFI*	*BIC*	*AIC*
6-factor original (I)	1495.95	579	0.056	0.96	0.96	2039.02	1669.95
6-factor modified (II)	1227.33	578	0.047	0.97	0.97	1776.65	1403.30
4-factor original (III)	1545.24	588	0.056	0.96	0.96	2032.13	1701.24
4-factor modified (IV)	1312.01	587	0.049	0.97	0.97	1805.15	1470.01

Note. SBSχ^2= Satorra-Bentler-scaled Chi-square; *df* = Degrees of Freedom; *BIC* = Bayesian Information Criterion; *AIC* = Akaike Information Criterion; *RMSEA* = Root Mean Square Error of Approximation; *NNFI* = Non-Normed Fit Index; *CFI* = Comparative Fit Index.

Table III also shows the fit indices of the four-factor model where *Affect*, *Cognitive Competence*, and *Difficulty* were joined into one factor (Model III). Again, the fit of this model substantially improved by allowing an error covariance between item 3 and item 19 (the first and fifth *Affect* items) (Model IV). Model IV performed significantly better than Model III (scaled-$\chi^2(1) = 233.23$, $p < .001$) and there was a considerable impact on the BIC and AIC values in favour of the Model IV (*ΔBIC* = 226.99, *ΔAIC* = 231.23).

A formal comparison of Model II and Model IV model indicated that Model II performed better (scaled-$\Delta\chi^2(9) = 84.68$, $p < .001$). The difference in BIC and AIC was clearly in favour of Model II (*ΔBIC* = 28.5, *ΔAIC* = 66.71). A small difference in RMSEA value was observed in favour of Model II. *CFI* and *NNFI* values were equal. As mentioned in the discussion, conclusions based on this comparison are not straightforward.

The latent factor correlations for Model 4 are shown in Table IV. Again, *Affect*, *Cognitive Competence*, and *Difficulty* were highly correlated ($r_{\text{Affect,CognitiveCompetence}} = 0.89$, $p < 0.05$; $r_{\text{Affect,Difficulty}} = 0.84$, $p < 0.05$; $r_{\text{CognitiveCompetence,Difficulty}} = 0.86$, $p < 0.05$).

Table IV

Estimated latent factor correlations for Model I

Model I: Six-factor original	Affect	Cognitive Competence	Value	Difficulty	Interest	Effort
Affect	1.00					
Cognitive Competence	0.89	1.00				
Value	0.39	0.43	1.00			
Difficulty	0.84	0.86	0.37	1.00		
Interest	0.46	0.47	0.71	0.46	1.00	
Effort	-0.16	-0.14	0.18	-0.24	0.22	1.00

Note. All correlations are significant (at least) at the 0.05 level.

Appendix II: Cross-validation of SATS-36 modifications across time and student groups

Cross-validation data for the modifications to the SATS-36 survey proposed in the manuscript are presented. First, it is examined whether the proposed modifications are replicated across time for the same sample of students (Educational sciences and Speech Pathology data) as the one analysed in the manuscript. Specifically, in the manuscript the first measurement time of five was presented and discussed. The five measurement times were: (1) just before the first year course (presented in the manuscript), (2) after the first year course but before the exam, (3) after the first year course, after the exam, (4) after the second year course, before the exam, and (5) at the beginning of the third year course. For a detailed description of the participants and procedure of this longitudinal data collection, we refer to chapter 3 or chapter 4.

Second, data from a mixture of students following diverse curricula were analysed to investigate whether results presented in the manuscript replicate to other groups of students (Table V). Data of these different groups of students are analysed together to attain an appropriate sample size for performing analyses on individual items. Students' attitudes toward statistics were assessed before and after an introductory statistics course (respectively Time 1 and Time 2), except for Applied Psychology students' that were assessed only after the statistics course.

Table V

Number of participants of different groups

Curriculum	Time	
	1	2
Archaeology or language studies	13	7
industrial engineering	352	233
business studies	161	84
Applied psychology	-	174
Mathematics, Physics or Computer Science	119	45
Total	645	543

Fit statistics for the original and modified 6-factor models are presented in Table VI. Standardized factor loadings estimates are shown in Table VII and VIII. Like the results

discussed in the manuscript, all results presented next are based on RML estimation with covariances.

Table VI

Fit indices for the cross-validation models

Educational sciences and Speech Pathology data									
Time	Model	*N*	*SBS*χ^2	*Df*	*RMSEA*	*NNFI*	*CFI*	*BIC*	*AIC*
1	6-factor original	514	1607.30	579	0.059	0.94	0.95	2150.37	1781.30
1	6-factor modified	514	1136.24	479	0.052	0.96	0.96	1648.10	1300.24
2	6-factor original	524	1957.58	579	0.067	0.95	0.95	2502.33	2131.58
2	6-factor modified	524	1328.13	448	0.061	0.96	0.97	1829.05	1488.13
3	6-factor original	622	2061.89	579	0.063	0.96	0.96	2621.56	2235.89
3	6-factor modified	622	1339.94	448	0.057	0.97	0.98	1854.58	1499.94
4	6-factor original	548	2382.11	579	0.071	0.93	0.94	2930.76	2556.11
4	6-factor modified	548	1347.85	448	0.061	0.96	0.96	1852.35	1507.85
5	6-factor original	519	1730.92	579	0.062	0.96	0.96	2274.84	1904.92
5	6-factor modified	519	1188.61	448	0.056	0.97	0.97	1688.76	1348.61
Data from several other curricula									
Time	Model	*N*	*SBS*χ^2	*Df*	*RMSEA*	*NNFI*	*CFI*	*BIC*	*AIC*
1	6-factor original	645	1973.59	579	0.061	0.93	0.93	2536.42	2147.59
1	6-factor modified	645	1311.72	448	0.055	0.95	0.96	1829.26	1471.72
2	6-factor original	543	2393.43	579	0.078	0.94	0.92	2941.28	2567.43
2	6-factor modified	543	1599.18	448	0.067	0.96	0.96	2102.95	1759.18

Note. *SBS*χ^2 = Satorra-Bentler-scaled Chi-square; *df* = Degrees of Freedom; *RMSEA* = Root Mean Square Error of Approximation; *NNFI* = Non-Normed Fit Index; *CFI* = Comparative Fit Index; *BIC* = Bayesian Information Criterion; *AIC* = Akaike Information Criterion

Table VII

Standardized factor loadings of the original 6-factor models

		Educational Sciences & Speech Pathology data					Mixed data	
		1	2	3	4	5	1	2
Affect	Item 3	0.65	0.81	0.80	0.73	0.77	0.42	0.74
	Item 4	0.59	0.62	0.70	0.67	0.73	0.65	0.62
	Item 15	0.69	0.74	0.77	0.72	0.76	0.73	0.69
	Item 18	0.66	0.70	0.75	0.65	0.69	0.58	0.65
	Item 19	0.62	0.78	*0.76*	0.69	0.71	0.41	0.74
	Item 28	0.75	0.78	0.84	0.74	0.81	0.72	0.70
Cognitive Competence	Item 5	0.71	0.75	0.80	0.75	0.80	0.73	0.74
	Item 11	0.54	0.62	0.62	0.55	0.56	0.58	0.74
	Item 26	0.55	0.53	0.57	0.42	0.50	0.44	0.49
	Item 31	0.48	0.62	0.66	0.56	0.57	0.46	0.67
	Item 32	0.61	0.83	0.77	0.76	0.77	0.57	0.82
	Item 35	0.83	0.81	0.89	0.78	0.83	0.77	0.79
Difficulty	Item 6	0.62	0.78	0.75	0.66	0.72	0.50	0.70
	Item 8	0.65	0.76	0.78	0.67	0.73	0.69	0.76
	Item 22	0.34	0.43	0.44	0.30	0.32	0.37	0.45
	Item 24	0.52	0.59	0.55	0.46	0.56	0.50	0.65
	Item 30	0.44	0.48	0.46	0.47	0.48	0.39	0.40
	Item 34	0.25	0.29	0.29	0.36	0.35	0.36	0.27
	Item 36	0.34	0.41	0.37	0.39	0.33	0.43	0.37
Value	Item 7	0.54	0.64	0.67	0.61	0.60	0.62	0.66
	Item 9	0.57	0.63	0.65	0.68	0.66	0.49	0.70
	Item 10	0.51	0.60	0.62	0.62	0.60	0.47	0.53
	Item 13	0.50	0.59	0.57	0.42	0.50	0.54	0.59
	Item 16	0.62	0.66	0.69	0.69	0.68	0.63	0.61
	Item 17	0.57	0.55	0.56	0.52	0.57	0.52	0.55
	Item 21	0.42	0.43	0.38	0.31	0.34	0.32	0.48
	Item 25	0.52	0.57	0.63	0.58	0.56	0.49	0.59
	Item 33	0.62	0.74	0.78	0.75	0.73	0.69	0.66
Interest	Item 12	0.59	0.65	0.69	0.66	0.63	0.60	0.64
	Item 20	0.82	0.91	0.88	0.87	0.82	0.83	0.91
	Item 23	0.85	0.84	0.84	0.80	0.80	0.74	0.83
	Item 29	0.84	0.88	0.89	0.86	0.86	0.85	0.87
Effort	Item 1	0.72	0.68	0.76	0.77	0.67	0.52	0.68
	Item 2	0.75	0.85	0.89	0.84	0.85	0.83	0.90
	Item 14	0.79	0.63	0.82	0.60	0.79	0.80	0.67
	Item 27	0.70	0.36	0.49	0.46	0.63	0.60	0.47

Note. Factor loadings in bold and underlined have values < 0.40. All coefficients are statistically significant at 0.01-level.

Table VIII

Standardized factor loadings of the modified 6-factor models

		Educational Sciences & Speech Pathology data					Mixed data	
		1	2	3	4	5	1	2
Affect	Item 3	0.57	0.72	0.74	0.64	0.73	0.34	0.62
	Item 4	0.62	0.68	0.73	0.71	0.75	0.67	0.66
	Item 15	0.71	0.76	0.79	0.75	0.77	0.73	0.74
	Item 18	0.69	0.74	0.78	0.70	0.71	0.61	0.72
	Item 19	0.53	0.68	0.69	0.59	0.66	0.33	0.62
	Item 28	0.78	0.82	0.86	0.79	0.83	0.74	0.77
Cognitive Competence	Item 5	0.71	0.75	0.80	0.75	0.80	0.73	0.75
	Item 11	0.54	0.62	0.62	0.55	0.56	0.57	0.74
	Item 26	0.55	0.54	0.57	0.42	0.50	0.44	0.50
	Item 31	0.48	0.62	0.66	0.57	0.57	0.47	0.67
	Item 32	0.61	0.83	0.77	0.76	0.77	0.57	0.82
	Item 35	0.82	0.81	0.88	0.78	0.83	0.77	0.79
Difficulty	Item 6	0.62	0.78	0.74	0.69	0.73	0.51	0.71
	Item 8	0.65	0.76	0.77	0.64	0.71	0.66	0.76
	Item 22	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted
	Item 24	0.51	0.58	0.53	0.43	0.55	0.46	0.62
	Item 30	0.43	0.48	0.45	0.44	0.46	0.40	0.41
	Item 34	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted
	Item 36	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted
Value	Item 7	0.54	0.65	0.67	0.62	0.60	0.62	0.67
	Item 9	0.57	0.65	0.66	0.69	0.67	0.50	0.71
	Item 10	0.51	0.61	0.62	0.62	0.61	0.47	0.56
	Item 13	0.50	0.59	0.57	0.42	0.50	0.54	0.59
	Item 16	0.62	0.65	0.68	0.68	0.67	0.63	0.59
	Item 17	0.57	0.54	0.56	0.52	0.56	0.52	0.53
	Item 21	0.42	Deleted	Deleted	Deleted	Deleted	Deleted	Deleted
	Item 25	0.52	0.56	0.63	0.57	0.55	0.48	0.59
	Item 33	0.62	0.73	0.77	0.74	0.72	0.68	0.65
Interest	Item 12	0.59	0.65	0.69	0.66	0.63	0.60	0.64
	Item 20	0.82	0.90	0.88	0.87	0.81	0.83	0.91
	Item 23	0.85	0.85	0.84	0.81	0.81	0.74	0.84
	Item 29	0.84	0.88	0.89	0.86	0.85	0.85	0.87
Effort	Item 1	0.72	0.68	0.76	0.76	0.67	0.52	0.69
	Item 2	0.75	0.85	0.89	0.84	0.85	0.83	0.90
	Item 14	0.78	0.63	0.82	0.60	0.79	0.80	0.67
	Item 27	0.70	0.46	0.49	0.46	0.63	0.60	0.47

Note. Factor loadings in bold and underlined have values < 0.40. All coefficients are statistically significant at 0.01-level.

For all additional administrations (across time and groups) the modifications suggested in the manuscript were replicated (deleting three *Difficulty* items with low factor loadings and adding an error covariance between item 3 and item 19). Additionally, because four administrations showed factor loadings < .40 for item 21 (*Value* subscale; "Statistics conclusions are rarely presented in everyday life") this item was also deleted in the modified models of the additional administrations. The modified models were substantially better fitting in terms of *AIC* and *BIC*. Small increases in *RMSEA*, *CFI* and *NNFI* were observed.

Although these cross-validation data increase confidence in the conclusions described in the manuscript (MacCallum et al., 1992; Hoyle & Panter, 1995), additional samples should be investigated. Especially concerning generalization across groups some limitations can be formulated. The data consist of a mixture of different kinds of students that followed different statistics courses, there was considerable drop out, and all administrations were performed in a Flemish context. Nevertheless, the modifications based on the (pretest) sample analysed in the manuscript appear not to be idiosyncratic.

Chapter 4

Longitudinal measurement invariance of the Survey of Attitudes Toward Statistics (SATS-36)[1]

Abstract

The goal of the present study was to examine longitudinal measurement invariance of the Survey of Attitudes Toward Statistics (SATS-36). By performing structural equation modeling on data from a five-wave longitudinal design with 785 university students, increasingly restrictive invariance tests (invariance of factor configuration, factor loadings, indicator intercepts, error variances, factor variances and factor means) were tested. Because valid assessment was not guaranteed for the Effort subscale, this subscale was discarded from analyses. Evidence of weak invariance and partial strong invariance was found for all other SATS-36 subscales, providing support for the SATS-36 as a useful survey for comparing statistics attitudes across time. Within-time variability of the subscales did not noticeably differ across time. Latent attitude means about the statistics domain remained stable over time, while latent mean differences emerged for students' attitudes about themselves as learners of statistics.

[1] Vanhoof, S., Kuppens, S., Ceulemans, E., Onghena, P., & Verschaffel, L. (2010). *Longitudinal measurement invariance of the Survey of Attitudes Toward Statistics (SATS-36).* Manuscript submitted for publication.

1 Introduction

The Survey of Attitudes Toward Statistics (SATS; Schau, Stevens, Dauphinee, & Del Vecchio, 1995) is an increasingly used survey to assess students' statistics attitudes. Despite the availability of some reliability and validity data on this survey, such as factorial validity, Cronbach's alpha reliability, convergent validity (e.g., see Cashin & Elmore, 2005; Dauphinee et al., 1997; Hilton et al., 2004; Tempelaar, van der Loeff, & Gijselaers, 2007), empirical evidence on longitudinal measurement invariance is very scarce for the SATS-28, an earlier version of the survey (Chiesi & Primi, 2009; Hilton et al., 2004; see Section 3), and – to the best of our knowledge – non-existent for the current SATS-36 version. However, longitudinal measurement invariance is a necessary prerequisite to evaluate temporal change, which is often the goal in statistics attitude research (e.g., Carnell, 2008; Leong, 2006). If longitudinal measurement invariance is not demonstrated, it cannot be determined whether observed temporal change is due to true change in the underlying construct or to changes in measurement of the construct over time (e.g., Bollen & Curran, 2006; Brown, 2006; Byrne, Shavelson, & Muthén, 1989; Little, Preacher, Selig & Card, 2007; Vandenberg & Lance, 2000). The main goal of the present study is to address this gap by testing longitudinal measurement invariance of the SATS-36 using a five-wave longitudinal study in a three-year instructional sequence.

1.1 Survey of Attitudes Toward Statistics

The SATS (Schau et al., 1995) has been developed to assess different aspects of students' statistics attitudes. A first version of the SATS (SATS-28) consisted of four subscales: (a) *Affect* (six items): positive and negative feelings concerning statistics; (b) *Cognitive Competence* (six items): attitudes about intellectual knowledge and skills applied to statistics; (c) *Difficulty* (seven items): attitudes about the difficulty of statistics as a subject, and (d) *Value* (nine items): attitudes about the usefulness, relevance, and worth of statistics in personal and professional life. Afterwards (Schau, 2003), two subscales were added to the survey (SATS-36): *Interest* (four items), students' level of individual interest in statistics and *Effort* (four items), the amount of effort students spend on learning statistics. Depending on the number of subscales and items involved, the developers labelled the survey as SATS-28 and SATS-36. Additionally, two versions of the SATS (SATS-pre and SATS-

post) are available, one to administer before a statistics course and one to administer after. The only difference between the two versions pertains to the tense of the verbs being used (respectively, future and past). A complete version of the SATS-36 and detailed scoring information can be consulted online via http://www.evaluationandstatistics.com/index.html.

1.2 Longitudinal measurement invariance of the SATS-28 and SATS-36

As mentioned in the introduction, only two studies have provided some evidence of longitudinal measurement invariance of the SATS-28. In the study of Hilton et al. (2004) invariance of factor loadings, factor variances and factor correlations across time, gender and gender*time was investigated. Factor loadings and factor correlations were invariant, whereas invariance of factor variances was not established across administration time. Except for the *Cognitive Competence* factor, all post-course variances were larger than their corresponding pre-course variances. The results of the invariance tests for factor loadings and factor variances/covariances of Chiesi and Primi (2009) were similar. For their Italian version of the SATS-28, factor loadings and factor covariances were invariant except for the association between *Affect* and *Difficulty*. Specifically, in the post-course measurement a stronger association between *Affect* and *Difficulty* was observed. Tests of latent means invariance revealed that all four factor means significantly increased during the course. Small effects were observed for the *Value* (d = .2) and *Affect* (d = .2) factors, a medium effect for *Difficulty* (d = .5) and a large effect for *Cognitive Competence* (d = .9).

Although these two studies already yielded some empirical evidence on longitudinal measurement invariance of the SATS-28, there are several reasons why further investigation is needed. First, no evidence is reported on the invariance of indicator intercepts (prerequisite to investigate latent mean differences) or error variances (prerequisite to investigate observed mean differences). Second, in both studies (and other SATS studies, such as Chiesi & Primi, in press; Dauphinee et al., 1997; Hilton et al., 2004; Schau et al., 1995) analyses were performed on item parcels rather than individual items. Although the technique of item parceling has its advantages, it remains controversial (e.g., Kline, 2005; Little, Cunningham, Shahar & Widaman, 2002). A major disadvantage of the use of item parcels in the context of longitudinal measurement invariance, is that no information on individual items is provided. As a consequence, item specific flaws regarding longitudinal

measurement invariance cannot be detected. A third reason for further investigation is that both studies (Chiesi & Primi, 2009; Hilton et al., 2004) investigated the first version of the SATS (SATS-28; see earlier) that not yet included the *Interest* en *Effort* subscales. No studies have been conducted on the most recent six-factor SATS-36 version. Finally, both studies were limited to only two measurement times, providing limited insight in the stability of the SATS-28 factor structure across time and the evolution of attitudes during the course of their curriculum.

The goal of the present longitudinal study was to extend the available research on measurement invariance of the SATS-36 by addressing the aforementioned limitations of earlier research. Besides tests of the longitudinal invariance of the relationships between the measured variables and latent constructs (factor configuration, factor loadings, intercepts, and error variances), tests concerning the latent constructs themselves (factor variances and factor means) were also performed. Because individual items rather than parcels were investigated, specific strengths and flaws of the survey could be detected.

2 Method

2.1 Participants

Participants were 785 Educational Sciences (496 female, 27 male) and Speech Pathology and Audiology (254 female, 8 male) students from three cohorts (203 students from Cohort 2004, 321 students from Cohort 2005, and 261 students from Cohort 2006) of an introductory undergraduate statistics course at the Department of Educational Sciences of the Katholieke Universiteit Leuven. Each cohort includes students that started their curriculum (consisting of three bachelor and two master years) in that specific year. As the numbers indicate, students in Educational Sciences and Speech Pathology and Audiology are mainly female.

The students followed three mandatory statistics courses, one in each of the three bachelor years. The introductory statistics course took place in the first semester of the first year. This course dealed with some introductory methodology and statistical tools (tables, figures and descriptive statistics). The second statistics course was in the first semester of the second year and covers design and sampling, probability and sampling distributions, as well as an introduction to statistical inference. In the second semester of the third year,

more advanced methodology and statistical techniques were covered, such as regression analysis and analysis of variance. For each statistics course, there was a two-hour theoretical class each week taught by the statistics professor and every two weeks students had to attend an additional two-hour practicum taught by a teaching assistant.

The goals and instructional approach of these three statistics courses were inspired by the textbook "Introduction to the Practice of Statistics" (Moore & McCabe, 2006). As can be read in the instructor's guide accompanying the textbook, this means that there was an emphasis on conceptual understanding of statistical notions, such as confidence interval, *p*-value, and power. The required mathematical background to follow these statistics courses was limited.

2.2 Measures

Attitudes toward statistics were assessed with a Dutch translation of the SATS-36 scale (Schau, 2003; Schau et al., 1995). Prior to analysis, scores on negatively formulated items were reversed to assure that a high score always indicates a positive attitude.

2.3 Procedure

Statistics attitudes were measured at five measurement times (Figure 1): (1) just before the first year course, (2) after the first year course but before the exam, (3) after the first year course, after the exam and after students knowing their exam results, (4) after the second year course, before the exam, and (5) after the second year course, after the exam and after students knowing their exam results (i.e. at the beginning of the third year course). For the first and second year course, this administration schedule yielded pre- and post-course as well as pre- and post-exams data, with a minimum of administration times. Statistics exams for the first and second year took place one month after the second and fourth SATS-36 administration, respectively.

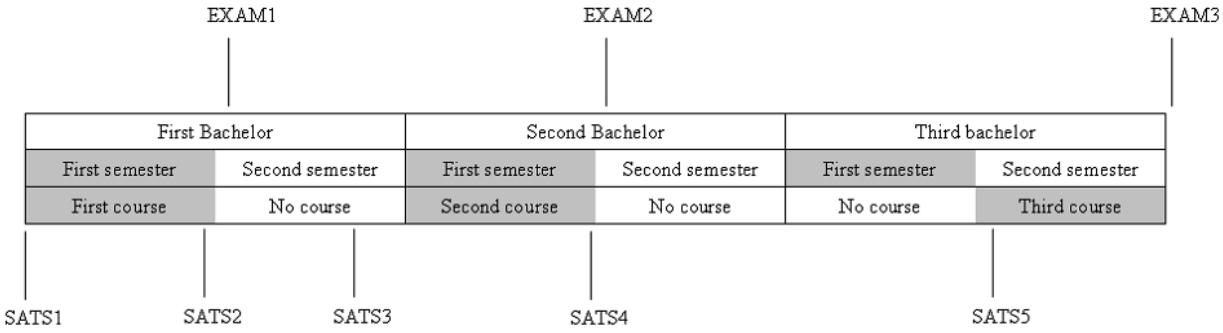

Figure 1. *Administration schedule of the SATS and Exams*

Students from Cohort 2005 and 2006 were involved from the start of their first statistics course (all five SATS-36 administrations), while the Cohort 2004 students only started participating in the study from the third administration time on. The SATS surveys were voluntarily completed and handed in during class time by most students present in the classes or were returned by (e)mail (see further). It was stressed that the data would be analysed anonymously.

A possible threat in longitudinal studies is dropout (Shadish, Cook, & Campbell, 2002). In the present study, special efforts were made to avoid dropout of participants. Students were contacted by telephone or e-mail each time they were not present when an administration took place. If they were willing to participate, they returned the SATS surveys electronically or by mail. By doing so, we managed to keep track of students who did not complete the survey during classes (364 SATS surveys over the five administration times) and even students who stopped their studies in Educational Sciences or Speech Pathology and Audiology (94 students).

Of all students 76.6% (*N* = 601) completed all surveys in their cohort or only missed one administration. The total number of completed surveys on the successive administrations were *N* = 555 (Time 1), *N* =524 (Time 2), *N* =622 (Time 3), *N* =594 (Time 4) and *N* =567 (Time 5).

A comparison based on *t*-tests of the attitudes scores on all measurement times of students with complete and incomplete data revealed no differences between these two groups regarding statistics attitudes.

2.4 Statistical Analyses

Invariance models

Using the LISREL 8.7 software (Jöreskog & Sörbom, 2004), structural equation modeling was performed on the individual SATS-36 items to test the longitudinal measurement invariance of the SATS-36. In order to analyse individual items, invariance tests were conducted separately for the six attitudes subscales of the SATS-36. Analyzing all subscales simultaneously would result in an inadequate subject-to-parameter ratio (Kline, 2005). To fully utilize all available information from the data, all models were estimated by Full Information Maximum Likelihood (FIML) estimation using the means and covariance matrix.

As commonly recommended (e.g., Bollen & Curran, 2006; Brown, 2006; Gregorich, 2006; Little et al., 2007; Vandenberg & Lance, 2000), longitudinal measurement invariance was tested by specifying a sequence of increasingly restrictive models (i.e., forcing elements in the factor model to be invariant across time) after evaluating confirmatory factor solutions separately at each measurement time. The first model tested *configural invariance* across measurement times. This least demanding test of invariance requires that each latent factor is associated with identical item clusters across time, without invariance constraints being imposed on any parameter estimates across time. Configural invariance is a necessary, but not sufficient condition for longitudinal measurement invariance. When this level of invariance is not achieved, the factor configuration has changed across time and no interpretable comparisons of attitudes over time are possible.

Second, assuming configural invariance, *weak invariance* was tested by constraining factor loadings to be equal across time. The central question here is whether the same latent factors are being measured across time. If weak invariance holds, quantitative comparisons of estimated *factor* variances and covariances are justified (Gregorich, 2006).

Third, assuming weak invariance, *strong invariance* was tested by additionally constraining indicator intercepts to be equal across time. When strong measurement invariance has been demonstrated, a meaningful investigation of factor mean differences across time is possible, as the test reveals whether mean shifts of an indicator reflect real changes or only/also changes in the intercept (Bollen & Curran 2006).

Fourth, *strict invariance* was tested by additionally constraining residual variances to be equal across time. In order to meaningfully compare both observed mean and variance scores across time, evidence of strict invariance should be obtained.

Fifth, if at least strong invariance was demonstrated, invariance tests of factor variances and means were conducted. Because in this study each SATS-36 subscale was analysed separately, invariance of factor covariances could not be investigated. Invariance of factor variances and means was tested by respectively constraining factor variances and means to be equal across time. This allowed determining whether the amount of variability in attitude subscales differed across time and whether latent mean differences emerged across time. Because attitudes are generally considered to be stable and resistant to change, especially after high school (e.g., Leong, 2006; McLeod, 1992), no or small differences in the latent constructs themselves (factor variances and means) were expected. In line with the distinction made between students' attitudes about a specific domain (i.e., *Value*) and their attitudes about themselves as learners of a domain (i.e., *Affect*, *Cognitive Competence*) (e.g., see McLeod, 1992), it was expected that especially the latter were more likely to change depending on changing circumstances during the progress of students' curriculum (Gal & Ginsburg, 1994).

Error covariances of the same indicators were allowed over time and reflect method effects associated with repeated administrations of the same measure (Brown, 2006). The factor loading of the first indicator was set to 1 and the first indicator's intercept to 0 to set the scale. This scaling provides the latent factors with a mean and scale similar to the reference indicator. As recommended by Bollen and Curran (2006), sensitivity checks were performed by choosing other scaling indicators because the scaling indicator can be the source of invariance problems (see also Vandenberg & Lance, 2000).

Model evaluation

Absolute model fit is usually evaluated using the chi-square statistic, which indicates an adequate fit if the obtained value is nonsignificant. However, it is widely known that the chi-square-based statistics are very sensitive to sample size (e.g., Chueng & Rensvold, 2002; Kline, 2005). This may result in the rejection of reasonable models because in the presence of large sample sizes small degrees of lack of fit already result in small *p*-values (Byrne, 1989; Hu, Bentler, & Kano, 1992). For this reason other goodness-of-fit indices were used to

evaluate absolute model fit: Root Mean Square Error of Approximation (*RMSEA*), Comparative Fit Index (*CFI*), and Non-normed fit index (*NNFI*). It has been suggested that a value of the *RMSEA* of less than .05 is an indication of a good fit whereas values between .05 and .08 still show a reasonable fit of the model (e.g., Kline, 2005). The *NNFI* and *CFI* indices normally range between zero and one, with higher values indicating a better fit. As a benchmark for good fit, the value 0.90 was used (e.g., Kline, 2005).

Because the chi-square difference test for comparing nested models also depends on sample size, as recommended by Cheung and Rensvold (2002), differences in comparative fit indices (*ΔCFI*) were used to evaluate invariance constraints. *CFI* differences greater than -.01 indicate that the null hypothesis of invariance should be rejected.

Partial invariance

If invariance at a particular level (weak, strong or strict invariance) was not reached, the models were explored in more detail by examining partial invariance (Bollen & Curran, 2006; Byrne, et al., 1989; Vandenberg & Lance, 2000). As suggested by Chueng and Rensvold (2002), the presence of a small portion of noninvariant items in the model will not affect comparisons to a meaningful degree. Partial invariance was tested using a backwards method, i.e. by removing the invariance constraints primarily associated with misfit (based on highest modification indices) until the difference in CFI with the previous invariance model was smaller than or equal to -.01. Because this post hoc practice involves a risk to capitalize on chance, a substantive explanation for the invariant items was considered important (Vandenberg & Lance, 2000). Furthermore, parameter differences (e.g., intercepts) of the invariant items were investigated, because if only minor absolute differences were observed, this would increase confidence in the models (Bollen & Curran, 2006).

3 Results

Because of the large number and size of the mean and covariance matrices analysed, they are not reported here. The mean and covariance matrices are available on request from the first author, as well as summary statistics (means, standard deviations, kurtosis, and skewness) for the items investigated.

3.1 Factor structure

First, separately for all SATS-36 subscales, confirmatory factor analytic models were fitted at each measurement time. Inspection of the fit indices revealed that the initial models were not adequate for *Affect*, *Cognitive Competence*, *Value, Difficulty,* and *Effort*, but that – except for *Effort* – minor modifications would result in good fitting models.

For *Affect*, *Cognitive Competence*, and *Value*, adding an error covariance between two items led to adequately fitting measurement models at each measurement time. For *Affect* the error covariance between items 3 and 19 ("I will like statistics." and "I will enjoy taking statistics courses.") referred to the only two positively formulated items which have closely related meanings. For the *Cognitive Competence* subscale, the error covariance between items 31 and 32 ("I can learn statistics." and "I will understand statistics equations.") referred to the only two positively formulated items that followed immediately after each other in the survey. The error covariance between items 16 and 17 ("Statistical thinking is not applicable in my life outside my job." and "I use statistics in my everyday life.") of the *Value* subscale referred also to items that immediately followed after each other.

For *Difficulty*, removing three items with factor loadings below .40 (item 22: "Statistics is a subject quickly learned by most people.", item 34: "Statistics is highly technical.", and item 36: "Most people have to learn a new way of thinking to do statistics.") also resulted in an adequate model fit. Item 22 and 36 pertain to "most people's" attitudes regarding the *Difficulty* of statistics, rather than students' own attitudes. Apparently, the term "technical" (item 34) does not refer to clear difficulty aspects such as "easy" (item 6) or "complicated" (item 8). Not unlikely, students can consider statistics to be technical, but not difficult.

Because there were no justifiable modifications possible for *Effort* that yielded a satisfactory fit, this subscale was not taken into account for further analyses. One item of this subscale (item 27: "I plan to attend every statistics class session") showed a factor loading below .40. This item also showed the most apparent deviation from normality at each administration time with skewness values ranging between -1.92 and -2.16 and kurtosis values between 4.26 and 5.02. With mean observed values ranging between 6.23 and 6.49 on a scale from 1 to 7, the deviations from normality are likely caused by ceiling effects.

Table 1 summarizes fit statistics for all final measurement models. All standardized factor loadings exceeded .40. Cronbach's alpha reliability values on the five measurement times for the final subscales ranged between 0.82 and 0.89 (*Affect*), 0.78 and 0.84 (*Cognitive Competence*), 0.65 and 0.72 (*Difficulty*), 0.78 and 0.83 (*Value*), 0.85 and 0.88 (*Interest*).

Table 1

Fit indices for the final separate measurement models

Affect	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	9.59	8	0.019	0.998	0.997
Time 2	17.85	8	0.048	0.994	0.988
Time 3	21.13	8	0.051	0.991	0.984
Time 4	15.58	8	0.04	0.995	0.991
Time 5	26.03	8	0.063	0.988	0.978
Cognitive Competence	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	7.76	8	<0.0001	1.00	1.00
Time 2	32.39	8	0.076	0.978	0.983
Time 3	19.30	8	0.048	0.995	0.996
Time 4	18.88	8	0.048	0.993	0.995
Time 5	49.63	8	0.096	0.934	0.951
Difficulty	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	3.77	2	0.040	0.993	0.98
Time 2	1.60	2	< 0.0001	1.00	1.00
Time 3	8.17	2	0.070	0.986	0.957
Time 4	7.48	2	0.068	0.981	0.944
Time 5	0.18	2	< 0.0001	1.00	1.01
Value	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	96.29	26	0.070	0.930	0.903
Time 2	139.59	26	0.091	0.904	0.866
Time 3	95.29	26	0.065	0.954	0.937
Time 4	112.89	26	0.075	0.945	0.924
Time 5	135.22	26	0.086	0.922	0.892
Interest	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	1.77	2	< 0.0001	1.00	1.00
Time 2	6.49	2	0.065	0.996	0.988
Time 3	6.15	2	0.058	0.997	0.990
Time 4	14.96	2	0.104	0.990	0.970
Time 5	21.55	2	0.131	0.983	0.950

Note. *FIML*χ^2 = Full Information Maximum Likelihood Chi-Square, *df* = degrees of freedom, *RMSEA* = Root Mean Square Error of Approximation, *CFI* = Comparative Fit Index, *NNFI* = Non-Normed Fit Index. Because poor fit indices for the separate measurement models for *Effort* emerged, this subscale was not considered for further analysis.

3.2 Longitudinal measurement invariance

Table 2 shows the fit statistics for the invariance models for all investigated SATS-36 subscales. For all subscales, the configural invariance models provided an adequate fit to the data, as indicated by appropriate fit indices (for all subscales *RMSEA* < .05, *CFI* > .90, *NNFI* > .90). Furthermore, for all subscales, *RMSEA* and *NNFI* values for the configural and weak invariance models were very similar and the difference in CFI values never exceeded the critical value of -.01. As such, all SATS-36 subscales demonstrated weak invariance across time.

Table 2

Fit statistics for the longitudinal invariance models

Affect	*df*	*FIML*χ^2	*RMSEA*	*CFI*	*ΔCFI*	*NNFI*
Configural invariance	330	944.81	.0487	.946	-	.929
Weak invariance	350	993.89	.0484	.943	-.003	.930
Strong invariance	370	1270.83	.0557	.921	-.022	.907
Partial strong invariance	362	1104.44	.0531	.935	-.008	.921
Strict invariance	386	1189.21	.0514	.929	-.006	.919
Invariance of factor variances	390	1215.07	.0519	.927	-.002	.918
Invariance of factor means	394	1506.84	.0600	.902	-.025	.892
Cognitive Competence	*df*	*FIMLχ2*	*RMSEA*	*CFI*	*ΔCFI*	*NNFI*
Configural invariance	335	709.10	.0377	.957	-	.944
Weak invariance	355	759.02	.0381	.953	-.004	.943
Strong invariance	375	1069.70	.0486	.920	-.033	.907
Partial strong invariance	367	824.60	.0399	.947	-.006	.938
Strict invariance	391	907.52	.0410	.940	-.007	.934
Invariance of factor variances	395	943.29	.0421	.937	-.003	.930
Invariance of factor means	399	1102.26	.0474	.919	-.018	.912
Difficulty	*df*	*FIMLχ2*	*RMSEA*	*CFI*	*ΔCFI*	*NNFI*
Configural invariance	120	177.32	.0247	.985	-	.977
Weak invariance	132	196.09	.0249	.983	-.002	.976
Strong invariance	144	306.17	.0379	.958	-.025	.945
Partial strong invariance	140	239.94	.0301	.974	-.009	.965
Strict invariance	156	255.20	.0285	.974	-.000	.969
Invariance of factor variances	160	293.25	.0326	.966	-.008	.959
Invariance of factor means	164	405.93	.0433	.938	-.028	.928

Value	*df*	*FIMLχ2*	*RMSEA*	*CFI*	*ΔCFI*	*NNFI*
Configural invariance	650	1215.33	.0332	.937	-	.924
Weak invariance	678	1268.42	.0331	.934	-.003	.924
Strong invariance	706	1683.20	.0420	.891	-.043	.879
Partial strong invariance	694	1357.74	.0349	.926	-.008	.916
Strict invariance	726	1466.26	.0360	.917	-.009	.911
Invariance of factor variances	730	1487.40	.0364	.915	-.002	.909
Invariance of factor means	734	1531.01	.0371	.911	-.004	.905
Interest	*df*	*FIMLχ2*	*RMSEA*	*CFI*	*ΔCFI*	*NNFI*
Configural invariance	120	172.86	.0237	.993	-	.989
Weak invariance	132	197.10	.0251	.991	-.002	.987
Strong invariance	144	512.01	,0571	.951	-.040	.935
Partial strong invariance	140	210.40	.0253	.991	-.000	.987
Strict invariance	156	232.98	.0251	.990	-.001	.987
Invariance of factor variances	160	234.98	.0244	.990	-.000	.988
Invariance of factor means	164	299.22	.0324	.982	-.008	.979

Note1. The *CFI* of partial strong invariance models is compared to the CFI of the weak invariance models.

In the following step, intercepts were constrained to be equal across time to test for strong invariance. For all subscales, the difference in *CFI* from the weak to the strong invariance model exceeded the critical value -.01, implying that strong invariance was not supported. Partial invariance of the intercepts was examined by freeing the intercepts constraints that showed highest modification indices successively until the difference in *CFI* was smaller than or equal to -.01. This resulted in models with one, two or three noninvariant item intercepts depending on the subscale. For *Affect*, intercept constraints on item 18 (“I will be under stress during statistics class.”) and item 19 (“I will enjoy taking statistics courses.”) interfered with model fit; for *Cognitive Competence* item 11 (“I will have no idea of what's going on in this statistics course.) and item 26 (“I will make a lot of math errors in statistics.”); for *Difficulty* item 30 (“Statistics involves massive computations.”); for *Value* item 9 (“Statistics should be a required part of my professional training.”), item 10 (“Statistical skills will make me more employable.”) and item 25 (“I will have no application for statistics in my profession.”); and for *Interest* item 12 (“I am interested in being able to communicate statistical information to others.”). An inspection of the intercepts of the noninvariant items revealed relatively minor absolute differences in intercepts. Possible

interpretations of these noninvariant intercept items will be presented in the Discussion Section.

Next, building on the models with partially invariant intercepts, residual variances were constrained to be equal across time to test strict invariance. Strict invariance was established for all subscales as indicated by appropriate fit statistics and a *ΔCFI*-value smaller than -.01.

Because at least partial strong invariance was supported, latent factor variances and means could be compared (Gregorich, 2006). Comparisons of factor variances are based on the strict measurement invariance models, taking noninvariant intercepts into account. For all investigated attitude subscales, adding these constraints did not yield a significantly poorer fit to the data compared to the less restrictive model. As such, the factor variances appeared to be equal across time.

Finally, factor means were restricted to be equal across time. Because the difference in CFI was smaller than .-01, factor means for *Value* and *Interest* were invariant across time. On the other hand, factor means for *Affect*, *Cognitive Competence*, and *Difficulty* showed significant changes across time (ΔCFI < -.01). T-tests showed that not all differences were statistically significant (see Table 3). Specifically, after knowing the exam results of the first course increased the mean attitudes were observed; subsequently, during the second year course the mean attitudes decreased, but they increased again after exam results were provided. The magnitude of the mean differences expressed as Cohen's *d* (Cohen, 1988) was small to medium.

Table 3

Latent factor means, confidence intervals, and Cohen's d effect sizes for Affect, Cognitive Competence, Difficulty and Effort

Affect	Mean	95% CI Mean		Cohen's *d*
		Lower	Upper	
Time 1	3.59	3.50	3.67	-
Time 2	3.89 +	3.81	3.98	0.36
Time 3	4.14 +	4.06	4.23	0.29
Time 4	3.83 -	3.74	3.91	0.37
Time 5	4.24 +	4.16	4.33	0.48
Cognitive Competence	Mean	95% CI Mean		Cohen's *d*
		Lower	Upper	
Time 1	4.11	4.01	4.21	-
Time 2	4.22	4.12	4.31	0.10
Time 3	4.47 +	4.38	4.56	0.25
Time 4	4.18 -	4.09	4.28	0.28
Time 5	4.59 +	4.50	4.69	0.40
Difficulty	Mean	95% CI Mean		Cohen's *d*
		Lower	Upper	
Time 1	3.40	3.32	3.49	-
Time 2	3.55	3.46	3.63	0.18
Time 3	3.72	3.63	3.80	0.22
Time 4	3.50 -	3.41	3.58	0.28
Time 5	3.80 +	3.71	3.88	0.38

Note. Attitudes at one point in time are compared to the previous point in time. "+" indicates a significant increase, "-" indicates a significant decrease.

4 Discussion

4.1 Longitudinal measurement invariance

This study examined longitudinal measurement invariance of the SATS-36 (Schau et al., 1997), a survey to assess students' statistics attitudes before and after statistics courses.

Mean and covariance structure analyses were used to assess longitudinal measurement invariance across five measurement moments during students' curriculum.

In a first phase, confirmatory factor models were used to investigate the factor structure of the subscales at the five measurement times separately. Because for the *Effort* subscale valid assessment was not guaranteed, this subscale was discarded from further analyses. Possibly, ceiling effects caused the lack of fit for this subscale. For the other subscales, item specific and substantively motivated suggestions could be formulated to improve the factor structure, because analyses were performed on individual items – and not on item parcels as was the case in earlier studies (e.g., Chiesi & Primi , 2009; Tempelaar et al.,2007).

Concerning measurement invariance, for all investigated SATS-36 subscales at least *weak* invariance was attained, implying that the meaning of the factors or the conceptual frame is invariant across time (Gregorich, 2006; Vandenberg & Lance, 2000). As a result the SATS-36 subscales can be appropriately used to compare factor variances and relative stability (e.g., an individual's rank order or relative position in the group; Finkel, 1995) across time.

Strong invariance across time did not hold for the SATS-36 subscales. *Partial* invariance of the intercepts, however, was supported for all subscales. As a consequence, (prudent) comparison of factor means across time is justified using the SATS-36 subscales (e.g., Bollen & Curran, 2006; Vandenberg & Lance, 2000). Some tentative hypotheses can be formulated regarding three types of noninvariant items. First, several noninvariant items (items 11, 18, and 19) pertain to the students' own statistics class or course, whereas the other items refer to the statistics field in general. Perhaps students mix field and course when answering these items across different administration times. Second, two noninvariant items share formulations closely related to mathematics, such as *math errors* and *computations* (items 26 and 30). It might be possible that especially at the first administration students' attitudes towards statistics are mainly based on their earlier experience with elementary and secondary school mathematics in general rather than by their (rare) experiences with statistics lessons in particular (see e.g., Gal et al., 1997 for a similar remark). Third, the noninvariant *Value* items (item 9, 10, and 25) refer to students' attitudes towards the *possible application of statistics in their professional live*. Perhaps the meaning of items regarding the value of statistics in general remains constant across time,

but when it concerns students' value of statistics for their own professional life, the meaning might depend on the specific time of measurement. Of course, these interpretations are tentative and cross-validation of the noninvariant items is necessary (Vandenberg & Lance, 2000). Finally, for all investigated SATS-36 subscales invariance of error variances was established. This implies that both observed means and variances can be compared (Gregorich, 2006).

Invariance tests of the factor variances revealed that within-time variability of the factors did not appreciably differ across time. This result differs from Hilton et al.'s (2004) finding that post-course variances were larger than their corresponding pre-course variances for all four investigated SATS-28 subscales.

Comparison of latent attitude means across students' curriculum, as expected, revealed that the mean attitude level was stable across time for attitudes about the usefulness, relevance, and worth of statistics in personal and professional life (*Value*), and students' level of individual interest in statistics (*Interest*). Significant latent mean differences were observed for students' attitudes about intellectual knowledge and skills applied to statistics (*Cognitive Competence*)**,** students' positive and negative feelings concerning statistics (*Affect*), and their attitudes about the difficulty of statistics as a subject (*Difficulty*). These results relate to the distinction often made in attitude research between students' attitudes about a specific domain (i.e., the value of statistics) and their attitudes about themselves as learners of a domain (i.e., affect, self-efficacy and perceived difficulty regarding statistics) (e.g., see McLeod, 1992; Gal & Ginsburg, 1994). On the one hand, attitudes about the statistics domain appear to be quite stable during students' curriculum. Apparently, these attitudes are formed at the beginning of the curriculum and are not substantially influenced by statistics courses or exam results. On the other hand, students' attitudes about themselves as learners of statistics (i.e., affect, self-efficacy and perceived difficulty of statistics) are more susceptible to changes. With one exception, more positive attitudes of students about themselves as learners of statistics were observed, with the largest increases (although still relatively small) not observed from the beginning to the end of a statistics course, but from before to after students' knowing their exam results. This result is in line with theories such as the Expectancy-Value theory (e.g., Eccles & Wigfield, 2002) that assume effects from achievement on attitudes. The reasoning in these theories is that good or bad achievement can result in a change in students' attitudes about their

competence to study statistics or about the difficulty of statistics as a subject, certainly when results are inconsistent with their attitudes.

Decreasing attitudes were observed from the beginning to the end of the second statistics course (before the exam). This course dealt with design and sampling, probability and sampling distributions – as well as an introduction to statistical inference – concepts that are traditionally considered to be difficult for students (e.g., Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007).

4.2 Limitations

Although this study is the first to investigate longitudinal measurement invariance of the SATS-36, limitations of the study can be formulated concerning the generalizability of the findings to other samples and contexts. Although the data collection was extensive, it was limited to one group of typically female Educational Sciences and Speech Pathology and Audiology students at one university. Results might be idiosyncratic and may not necessarily generalize to other educational contexts or Educational Sciences and Speech Pathology groups at other countries/universities. Furthermore, some post hoc modifications were made to the separate CFA models and regarding partial strong invariance, which always involves the risk to capitalize on chance (MacCallum et al., 1992). Cross-validation is needed. Another limitation is that analyses were performed on separate SATS-36 subscales. This allowed the examination of longitudinal measurement invariance by analyzing individual items, but it made investigation of the invariance of covariation between the SATS-36 subscales across time impossible.

4.3 Implications and suggestions for further research

Some implications for the assessment of attitudes and suggestions for further research can be drawn from the present study. First, statistics teachers and researchers are encouraged to take formulated improvements to the SATS-36 into account. For instance, some SATS-36 items were not deemed appropriate indicators for the attitude aspects they are supposed to measure and – based on the results of this study – should be deleted from the survey.

Second, although this study provides empirical evidence regarding the suitability of the SATS-36 for assessing temporal change in statistics attitudes, researchers are

encouraged to investigate longitudinal measurement invariance in their specific context before assessing evolutions in statistics attitudes. If the factor structure of an instrument is not stable over time, then longitudinal comparisons are not meaningful (Chueng & Rensvold, 2002). Too often, observed attitude scores are compared without testing these necessary assumptions. At the same time, it cannot be expected that statistics teachers would investigate measurement invariance before assessing attitude change. Because strict invariance was reached (hopefully supported by future results from other studies), teachers can become more and more confident that observed attitude (mean or variance) changes (for instance from the beginning to the end of their statistics course) can be meaningfully interpreted. Although the size of absolute intercept differences for the noninvariant items was small in this study, we especially suggest researchers to investigate the invariance of SATS-36 item intercepts in further studies.

Third, when designing studies to assess evolutions of statistics attitudes, researchers are encouraged not only to assess attitudes from the beginning to the end of a statistics course, but also before and after students knowing their exam results. Knowledge of exam results seems an important factor, but in earlier studies assessing attitudes before and after a statistics course, the exact measurement time of the second administration (before or after the exams) was often not considered important.

Chapter 5

The directionality of the relationship between statistics attitudes and achievement: Evidence from a longitudinal study with university students[1]

Abstract

The goal of the present study was to investigate the directionality of the relationship between statistics attitudes and statistics achievement. Previously, not supported by appropriate empirical data, many researchers assumed a unidirectional effect from statistics attitudes to statistics achievement. In this study, using structural equation modeling, data collected according to a longitudinal design with 785 university students were analysed to provide empirical evidence on the directionality of effects. A comparison of alternative plausible models showed results that were opposite from the common view: A unidirectional model with effects from statistics achievement to statistics attitudes was found for students' attitudes about themselves as learners of statistics. Regarding attitudes about the domain of statistics, no effects over and above the stability effect of attitudes and achievement were present during the progress of the students' curriculum. Based on these results, it is suggested that rather than fostering positive attitudes because of their effect on achievement, improving students' achievement in statistics is a strategy for eliciting positive statistics attitudes about themselves as learners.

[1] Vanhoof, S., Kuppens, S., Ceulemans, E., Timmerman, M., Verschaffel, L., & Onghena, P. (2010). *The directionality of the relationship between statistics attitudes and achievement: evidence from a longitudinal study with university students.* Manuscript submitted for publication.

1 Introduction

The importance of taking students' attitudes toward statistics into account when studying statistics achievement is widely recognized (Gal, Ginsburg, & Schau, 1997). A central idea in a number of educational theories (e.g., Expectancy-Value Theory) is that affective factors have an effect on achievement, possibly mediated by cognitive (e.g., attention demands) or motivational mechanisms (Eccles & Wigfield, 2002; Garfield, Hogg, Schau, & Whittinghill, 2002; Pekrun, 1992; Schau, 2003). In line with this idea, several authors interpreted small to moderate associations they observed between statistics attitudes and statistics achievement as unidirectional, with effects from attitudes to achievement (e.g., Cashin & Elmore, 2005; Waters, Martelli, Zakrajsek, & Popovich, 1988; Tempelaar, van der Loeff, & Gijselaers, 2007). However, in most empirical studies on the relationship between statistics attitudes and achievement designs were used that do not permit drawing such conclusions regarding the directionality of effects. Furthermore, although less central, most educational theories also incorporate the possibility of reverse effects from achievement to attitudes (Eccles & Wigfield, 2002; Garfield et al., 2002; Pekrun, 1992; Schau, 2003).

Three alternative interpretations of directionality of the effects between statistics attitudes and statistics achievement are possible: (1) The effect is unidirectional from attitudes to achievement, (2) The effect may go in the other direction, from achievement to statistics attitudes, or (3) there may be an effect in both directions, with attitudes and achievement mutually influencing each other. This question on the directionality of effects is not only theoretically interesting but has also practical implications. For instance, it has implications for the importance placed on statistics attitude improvement interventions as a means to facilitating statistics achievement.

Appropriate research designs such as longitudinal studies are needed if we want to gain insight into the exact nature of the relationship between statistics attitudes and achievement (Gal et al., 1997). The present study aims at examining the directionality of effects on the basis of a longitudinal study in a three-year instructional sequence, using structural equation modeling. As will become clear (see Section 2), necessary preparatory analyses for the investigation of directionality of effects, will also yield interesting psychometric properties of the Survey of Attitudes Toward Statistics (SATS-36; Schau,

Stevens, Dauphinee, & Del Vecchio, 1995; Schau, 2003), such as factor structure, item functioning, factorial stability (e.g., see also Chapter 4) and information regarding the relative stability of statistics attitudes.

1.1 Statistics Attitudes

Attitude is a central concept in educational psychology. An enormous number of studies on attitudes in different fields have resulted in a wide variety of conceptualisations (e.g., Eccles & Wigfield, 2002; Op 't Eynde, De Corte, & Verschaffel, 2006). However, there seems to be general agreement that an attitude represents "a latent disposition or tendency to respond with some degree of favourableness or unfavourableness to a psychological object" (Fishbein & Ajzen, 2010, p. 76; see also Ajzen, 2001; Eagly & Chaiken, 1993). Attitudes are considered generally stable, but less stable than personality traits and more stable than emotions (McLeod, 1992; Graham et al., 2007).

In statistics education research, attitudes toward statistics are often defined as multidimensional: including the degree of affective (emotions and the motivation related to the classes and examinations), cognitive (beliefs and knowledge about the ability requested to learn statistics and about the discipline) and behavioural (action tendencies in studying and the achievement in examinations) aspects (Chiesi & Primi, 2009; Fishbein & Ajzen, 2010; Gal et al., 1997; Olson & Zanna, 1993). A number of instruments have already been developed and evaluated to assess students' attitudes toward statistics, like the Attitudes Toward Statistics (ATS) scale (Wise, 1985) and the Survey of Attitudes Toward Statistics (SATS-36; Schau, Stevens, Dauphinee, & Del Vecchio, 1995; Schau, 2003). As in research on mathematics education (McLeod, 1992), these surveys distinguish between attitudes about the statistics domain (e.g., the value of statistics) and about themselves as learners of statistics (e.g., self-efficacy regarding statistics).

1.2 Relationship between statistics attitudes and statistics achievement

Several studies have reported small to moderate positive associations between statistics attitudes and statistics achievement, where the latter is commonly operationalized by exam results (e.g., Kottke, 2000; Lalonde & Gardner, 1993; Nasser, 1999, 2004; Roberts & Bilderback, 1980; Vanhoof et al., 2006; Wise, 1985; Wisenbaker, Nasser, & Scott 1998). Interesting to note is that a similar relationship between affective and cognitive factors has

been observed regarding statistics anxiety and achievement in statistics classes (for a review see Onwuegbuzie & Wilson, 2003), regarding attitude toward mathematics and achievement in mathematics (Ma & Xu, 2004), regarding writing attitudes and writing achievement (Graham et al., 2007) and regarding academic self-concept and achievement (Green, Nelson, Martin, & Marsh, 2006).

A repeated observation regarding the relationship between statistics attitudes and statistics achievement is that especially students' attitudes about themselves as learners of statistics (e.g., self-efficacy regarding statistics) were related to achievement, rather than attitudes about the domain of statistics (Carmona, 2002; Chiesi & Primi, 2009; Rhoads & Hubele, 2000; Schutz, Drogosz, White, & Distefano, 1998; Sorge & Schau, 2002; Waters et al., 1988; Wise, 1985).

1.3 Directionality of effects

So far, in the field of statistics education, the most popular conception of observed associations between statistics attitudes and statistics achievement is that unidirectional attitude effects exist, with statistics attitudes preceding exam results (e.g., Cashin & Elmore, 2005; Waters et al., 1988; Tempelaar et al., 2007). In most research designs used, this view is implicitly present because attitudes measured at the beginning or end of a course are used to predict later achievement (e.g., Cashin & Elmore, 2005; Chiesi & Primi, 2009; Rhoads & Hubele, 2000; Roberts & Saxe, 1982; Waters et al., 1988; Waters, Martelli, Zakrajsek, & Popovich, 1989; Wise, 1985). In other studies the impact of didactic approaches on attitudes is studied by comparing pre- and post-course attitudes toward statistics. Without explicitly including achievement, in these studies it is believed that approaches that foster positive attitudes indirectly will improve achievement (D'Andrea & Waters, 2002; Mvududu, 2003; Suanpang, Petocz, & Kalceff, 2004).

To explain attitude effects on achievement, some authors refer to models such as the Expectancy-Value model (Eccles & Wigfield, 2002) in which affective factors are assumed to influence achievement (e.g., Schau, 2003; Tempelaar et al., 2007). Possible explanatory mechanisms are that negative attitudes need more cognitive resources than positive attitudes, which results in less attention for completing a task, or that negative attitudes influence achievement because of their negative effect on intrinsic motivation (Graham et al., 2007; Heckhausen, 1991; Pekrun, 1992). For instance, in the context of writing

achievement, Graham et al. (2007) assume that children with negative attitudes may try to avoid writing tasks or invest little effort when they are required to write.

Despite the fact that reverse effects from achievement to attitudes are theorized in earlier mentioned models such as the Expectancy-Value model (Eccles & Wigfield, 2002) or Pekrun's model about the impact of emotions on learning and achievement (Pekrun, 1992), this reverse effect is much less commonly mentioned with regard to statistics achievement and attitudes. Nevertheless some authors do bring up the possibility of a reverse effect from achievement to statistics attitudes, often referring to the effect of earlier high school mathematics achievement on university students' statistics attitudes (Carmona, 2002; Gal et al., 1997; Harlow, Burkholder, Morrow, & Morrow, 2002; Onwuegbuzie & Wilson, 2003; Sorge & Schau, 2002; Schau, 2003; Wisenbaker & Scott, 1995). The reasoning is that good or bad achievement can result in a change in students' attitudes about their competence to study statistics or about the difficulty of statistics as a subject, certainly when results are inconsistent with their attitudes.

Notwithstanding these different views, the directionality of effects between statistics attitudes and achievement remains empirically unexplored. In most studies statements regarding the directionality of effects were not supported by appropriate empirical data and analyses. With the present study, we aimed at filling this gap in the empirical research by examining the directionality of relationships between statistics attitudes and statistics achievement on the basis of a longitudinal design. To this end, we assessed statistics attitudes and statistics achievement at various time points for the same students.

The three alternative interpretations regarding the directionality of effects mentioned in the introduction were compared. Consistent with results from similar studies in related fields such as mathematics and writing education, a bidirectional relationship was expected with attitudes and achievement positively influencing each other (Calsyn & Kenny, 1977; Graham et al., 2007; Ma & Xu, 2004; Marsh & Yeung, 1997; Minato & Kamada, 1996). The majority of those studies did not observe a predominance of one of the effects, but in the study by Ma and Xu (2004), predominance of achievement effects was found. As this study was most similar to ours in terms of design and number of waves, a similar result was anticipated.

Because in earlier studies a stronger relationship between achievement and students' attitudes about themselves as learners of statistics was observed, as compared to the

relationship between achievement and attitudes about the domain of statistics (see Section 1.2), we expected to observe especially effects (in either direction) for students' attitudes about themselves as learners.

2 Method

2.1 Participants

Participants were 785 Educational Sciences (496 female, 27 male) and Speech Pathology and Audiology (254 female, 8 male) students from three cohorts (203 students from Cohort 2004, 321 students from Cohort 2005, and 261 students from Cohort 2006) of an introductory undergraduate statistics course at the Department of Educational Sciences of the Katholieke Universiteit Leuven Each cohort includes students that started their curriculum (consisting of three bachelor and two master years) in that specific year. As the numbers suggest, students in Educational Sciences and Speech Pathology and Audiology were mainly female.

The students had to follow three compulsory statistics courses, one in each of the three bachelor years. The introductory statistics course took place in the first semester of the first year. This course dealt with some introductory methodology and statistical tools (tables, figures and descriptive statistics). The second statistics course was in the first semester of the second year and covered design and sampling, probability and sampling distributions, as well as an introduction to statistical inference. In the second semester of the third year, more advanced methodology and statistical techniques were covered, such as regression analysis and analysis of variance. For each statistics course, there was a two-hour theoretical class each week taught by the statistics professor and every two weeks students had to attend a two-hour exercise class taught by a teaching assistant.

The instructional approach of the statistics courses follows the philosophy and goals of the textbook "Introduction to the Practice of Statistics" (Moore & McCabe, 2006). As can be read in the instructor's guide accompanying the textbook, this means that there was an emphasis a conceptual understanding of statistical notions, such as confidence interval, *p*-value, and power. The mathematical background required to follow the statistics courses was limited. Furthermore, a substantial amount of time was devoted to research methodology and sampling.

2.2 Measures

Attitudes toward statistics were assessed with a Dutch translation of the SATS-36 (Schau et al., 1995; Schau, 2003). The SATS-36 is a 36-item Likert-type survey with seven response possibilities for each statement ranging from “strongly disagree” to “strongly agree”. Two versions of the SATS-36 (SATS-36-pre and SATS-36-post) are available, one to administer before a statistics course and one to administer after. The difference between the two versions pertains to verb tense. Prior to analysis, scores on negatively formulated items were reversed to assure that a high score indicates a positive attitude.

The SATS-36 consists of six components: (a) *Affect* (6 items): positive and negative feelings concerning statistics; (b) *Cognitive competence* (6 items): attitudes about intellectual knowledge and skills applied to statistics; (c) *Difficulty* (7 items): attitudes about the difficulty of statistics as a subject; (d) *Value* (9 items): attitudes about the usefulness, relevance, and worth of statistics in personal and professional life; (e) *Interest* (4 items), students’ level of individual interest in statistics, and (f) *Effort* (4 items), the amount of effort students expend on learning statistics. *Affect*, *Cognitive Competence*, and *Interest* measure aspects of students’ attitudes about themselves as learners of statistics, whereas *Value* and *Difficulty* ask about students’ attitudes regarding the domain of statistics (cf. Section 1.2). In earlier studies (based on a confirmatory factor analysis on item parcels) the factor structure of the SATS was confirmed and appropriate Cronbach’s alpha reliability values were observed (e.g., Chiesi & Primi, 2009; Hilton et al., 2004; Schau et al., 1995; Schau, 2003; Tempelaar, 2007). In a confirmatory factor analytic study based on individual items, some minor changes to the surveys were proposed (see Chapter 4).

Items in the statistics exams were in line with the philosophy and instructional approach of the statistics courses and focused on methodology and students’ mastery of statistical knowledge and application and understanding of statistical methods. All exams resulted in a score between 0 and 20. The first and second year exams were closed book exams, but students were allowed to use an extensive sheet with formulae. The third year exam was an open book exam. In the analyses, for each year, statistics achievement was operationalized as the total exam result.

2.3 Procedure

The attitudes toward statistics were measured at five measurement times (Figure 1) was guided by the goal of the study, and taken as (1) just before the first year course, (2) after the first year course but before the exam, (3) after the first year course, after the exam, (4) after the second year course, before the exam, and (5) at the beginning of the third year course. This administration schedule enabled us to collect data from the students before and after each course (except after the third year course; Figure 1) as well as before and after the exams, with a minimum of administration times.

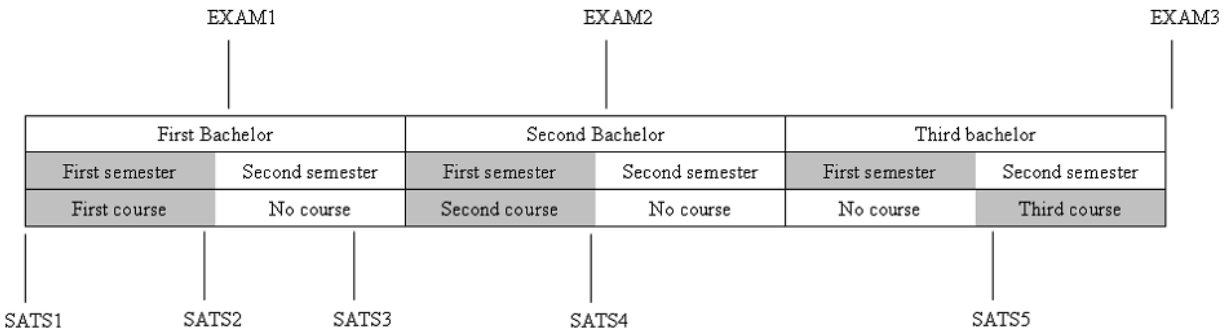

Figure 1. *Administration schedule of the SATS and Exams*

Note. For example, for Cohort 2005 the SATS-36 was administered September 2005, December 2005, May 2006, December 2006, and April 2008. Exams took place in January 2006, January 2007 and June 2008.

Students from Cohort 2005 and 2006 were involved from the start of their first statistics course (all five SATS-36 administrations), while the Cohort 2004 students only started participating in the study from the third administration time on. The surveys were voluntarily completed and handed in during class time, by mail or electronically (see further). It was stressed that the data would be analysed anonymously.

A possible threat in longitudinal studies is dropout (Shadish, Cook, & Campbell, 2002). In the present study, special efforts were made to avoid dropout of participants. Students were contacted by telephone or e-mail each time they were not present when an administration took place. If they were willing to participate, they returned the surveys electronically or by mail. By doing so, we managed to keep track of students who were

absent during classes (364 surveys over the five administration times) and even students who stopped their studies (94 students) in Educational Sciences or Speech Pathology and Audiology.

Of the Cohort 2004 students (who could in principle participate three times) 92.7% (*N* = 188) participated two or three times. Of Cohort 2005 and 2006 (with maximally five participations), 84.4% (*N* = 491) participated at least three times. 76.6% (N = 601) of all students completed all surveys possible in their cohort or only missed one administration. The total numbers of filled out surveys on the successive administrations were *N* = 555 (Time 1), *N* =524 (Time 2), *N* =622 (Time 3), *N* =594 (Time 4) and *N* =567 (Time 5). Of all 2862 completed versions of the SATS-36, 84.0% (*N* = 2404) were filled out during class time and 16.0% (*N* = 458) were obtained after contacting students.

A comparison based on *t*-tests of the attitudes scores on all measurement times of students with complete and incomplete data revealed no evidence that our data would suffer from selection problems regarding attitudes. However, not surprisingly, dropouts did perform significantly worse on the first year exams than the non-dropouts. The obtained structural equation modeling results that will be discussed next were similar for analyses including and excluding drop-outs. Therefore, results from models for all available data are reported.

2.4 Statistical analyses

Structural equation modeling using LISREL software (Jöreskog & Sörbom, 2004) was performed to analyse the data as it allows simultaneously investigating measurement properties of the surveys and the directionality of effects. Because the six subscales consist of four to nine items, simultaneous consideration of all SATS-36-subscales would result in a subject/parameter ratio that is insufficient; therefore, separate analyses per subscale were performed. Missing values on item-level (total percentage missing values = 0.0009%) were imputed by means of the EM (expectation-maximization) algorithm (Schafer & Graham, 2002). Missing survey-level data were treated using a full information maximum likelihood estimation procedure (FIML).

A two-step modeling approach, including a measurement and a structural phase, was employed to analyse the data. As advocated by others (e.g., Finkel, 1995) several intermediate steps were taken in the measurement phase. First, measurement models for

the attitude data were tested separately at the five measurement times for all individual subscales. Second, for all subscales, a joint measurement model, including the measures on attitudes and exam results, was fitted. Exam results were represented as single-indicator latent variables (with error variances set to zero) (Figure 2). The latent factors were permitted to correlate. Moreover, error covariances of the same indicators were allowed over time to reduce bias in the structural paths (see further). Third, as the assumption of measuring the same construct over time is essential in longitudinal models (Chiesi & Primi, 2009; Gregorich, 2006; Hilton et al., 2004), the invariance of factor loadings was tested for the statistics attitudes constructs. The null hypothesis, pertaining to the equality of factor loading across the five measurement waves, was tested in terms of relative model fit, by comparing an unconstrained model to a model that constrained the factor loadings to be equal over time.

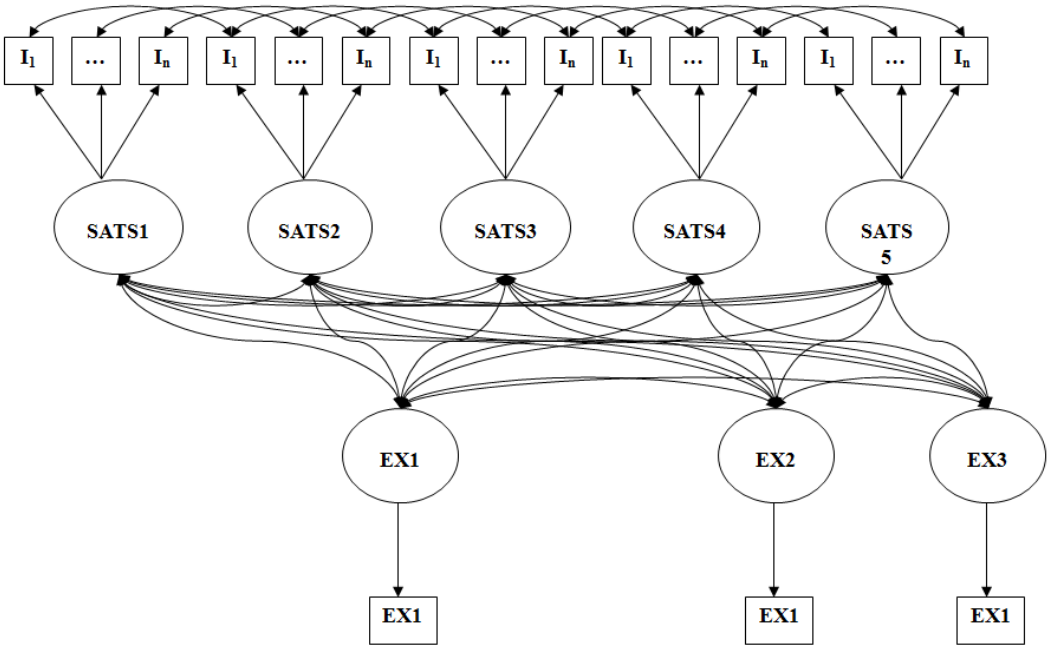

Figure 2. *Joint measurement models*

In the structural phase, four models reflecting different hypotheses about the directionality of effects between statistics attitudes and exam results were tested and compared. Figure 3 shows the path diagrams of the structural models tested. First, an autoregressive model (M1) was fitted; this model functions as the baseline model. The model only included autoregressive effects, which means that each variable is only

influenced by its value at the previous measurement time, and the initial covariance between the first attitude and exam measure. This model reflects the stability of construct across time. Without taking this stability (i.e., autoregressive effects) into account, apparently strong, but spurious relationships between variables of interest are likely to be found (MacCallum & Austin, 2000). Therefore, in the following models we examined cross-lagged effects, while taking stability across time into account.

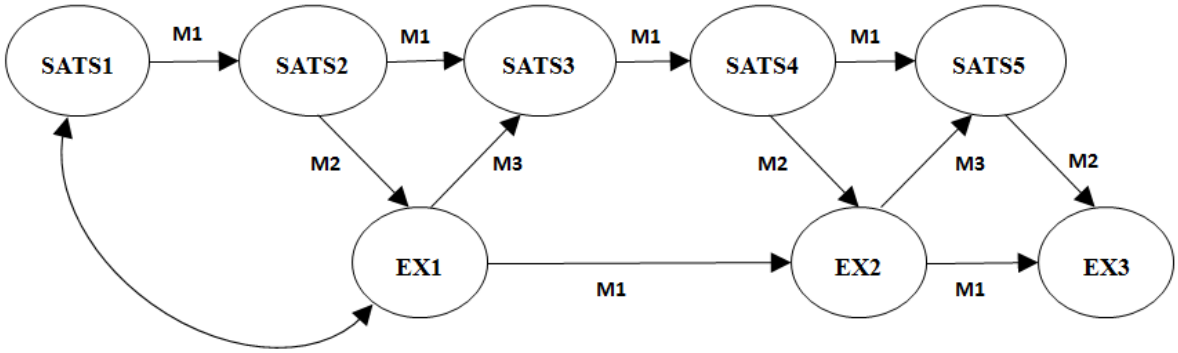

Figure 3. *Structural models representing autoregressive, unidirectional, and bidirectional relationships between statistics attitudes and exam results*

Notes.

Initial covariance is included for all models.

Autoregressive model = M1

Unidirectional model SATS to EXAM = M2 is added to the autoregressive model

Unidirectional model EXAM to SATS = M3 is added to the autoregressive model

Bidirectional model = M2 and M3 are added to the autoregressive model

The second and third models are the "standard" (M2) and the "reverse" (M3) unidirectional models, respectively. M2 was specified by extending the autoregressive model with cross-lagged structural paths from statistics attitudes to exam results. Analogously, M3 was specified by adding paths from exam results to statistics attitudes. The two unidirectional models were compared to the autoregressive model and to each other.

The best fitting unidirectional model ("standard" or "reverse") was compared to the bidirectional model (M4) with effects going from exam results to statistics attitudes and vice versa. This model reflects the idea that statistics attitudes and exam results mutually

influence each other. The cross-lagged paths from attitudes to exam results represent attitude effects and the cross-lagged paths from exam results to attitudes represent exam effects.

In the final models, the equality of attitude and/or exam effects across the measurement times was also tested by comparing an unconstrained model with a model constraining the particular effects to be equal across the measurement waves (Finkel, 1995). When the equality constraint is tenable, the cross-lagged effects are assumed to be consistent across the measurement waves.

Absolute model fit is usually evaluated using the chi-square statistic, which indicates an adequate fit if the obtained value is non-significant. However, since chi-square fit statistics are highly sensitive to sample size and may lead to the rejection of acceptable models in larger samples, several alternative fit indices were examined, namely the Root Mean Square Error of Approximation (*RMSEA*), the Comparative Fit Index (*CFI*), and the Non-Normed Fit Index (*NNFI*) (Kline, 2005). A value of the *RMSEA* less than .05 is considered indicative of a good fit to the data, whereas values between .05 and .10 are considered to represent a reasonable fit. As a benchmark for good fit of *CFI* and *NNFI,* the value .90 or larger was used (Kline, 2005).

Relative fit of the competing models was assessed using the Akaike Information Criterion (*AIC)* and the Bayesian Information Criterion (*BIC)* in conjunction with other alternative fit indices (Wang & Liu, 2006). When paths were added to a model and did not show a substantial improvement to the more parsimonious model, the more parsimonious model was preferred. Although traditionally a chi-square difference test is used to compare nested models, it has the same undesirable properties as the standard chi-square statistic, such as systematically rejecting reasonable parsimonious models when sample size is large (Marsh, 1994). As recommended by Cheung and Rensvold (2002), differences in comparative fit indices (*ΔCFI*) were used to evaluate invariance constraints, with a value for $|\Delta CFI|$ equal to or smaller than .01 indicating that the null hypothesis of invariance should not be rejected.

3 Results

Because of the number (six or one for each subscale) and size of the covariance matrices analysed, they are not reported here. The covariance matrices are available on request from the first author.

3.1 Measurement models

In the measurement phase, separately for all SATS-36 components, confirmatory factor analytic models were first fitted within each measurement time. Inspection of the fit indices revealed that the initial models were not adequate for *Affect*, *Cognitive Competence*, *Value, Difficulty,* and *Effort*.

For *Affect*, *Cognitive Competence*, and *Value*, adding an error covariance between two items led to adequately fitting measurement models at each measurement time: between items 3 and 19 ("I will like statistics." and "I will enjoy taking statistics courses.") that were the only two positively formulated items of the *Affect* factor and have closely related meanings, between items 31 and 32 ("I can learn statistics." and "I will understand statistics equations.") that were the only two positively formulated items of the *Cognitive Competence* subscale and follow immediately after each other in the survey, and between items 16 and 17 ("Statistical thinking is not applicable in my life outside my job." and "I use statistics in my everyday life.") of the *Value* subscale that also follow immediately after each other.

For *Difficulty*, removing three items with factor loadings below .40 (item 22: "Statistics is a subject quickly learned by most people.", item 34: "Statistics is highly technical." and item 36: "Most people have to learn a new way of thinking to do statistics.") also yielded an adequate fit. Item 22 and 36 ask about "most people's" attitudes regarding the *Difficulty* of statistics, rather than students' own attitudes.

Because there were no justifiable modifications possible for *Effort* that yielded a satisfactorily fit, this subscale was not taken into account for further analyses. Table 1 summarizes fit statistics for all final measurement models. All standardized factor loadings exceeded .40. Cronbach's alpha reliability values on the five measurement times for the final subscales ranged between 0.82 and 0.89 (*Affect*), 0.78 and 0.84 (*Cognitive Competence*), 0.65 and 0.72 (*Difficulty*), 0.78 and 0.83 (*Value*), 0.85 and 0.88 (*Interest*).

In the next step, for all subscales, the five joint measurement models were tested. They provided an adequate fit to the data, as indicated by appropriate fit indices (for all subscales *RMSEA* < .05, *CFI* > .93, *NNFI* > .92). In the final measurement step, the invariance of the factor loadings across the measurement waves was tested by comparing an unconstrained model with a model constraining the factor loadings of the joint measurement model to be equal across the five measurement times. For all subscales, *RMSEA* and *NNFI* values for the two models were very similar and differences in comparative fit indices ($|\Delta CFI| \leq .004$) never exceeded the proposed benchmark of .01. These results suggested that the null hypothesis of equality of factor loadings across the measurement waves was tenable, which means that our models showed measurement invariance, a necessary condition to continue testing the structural models (Cheung & Rensvold, 2002). The fit statistics for the final joint measurement models are also shown in Table 1.

3.2 Structural models

Table 2 depicts the fit statistics for all structural models. For all subscales, a baseline model (M1) incorporating autoregressive coefficients fits the data adequately with *RMSEA-values* around or under the .05 threshold and *CFI* and *NNFI-values* around or over .90.

Table 1

Fit indices for the final separate and joint measurement models

Affect	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	9.59	8	0.019	0.998	0.997
Time 2	17.85	8	0.048	0.994	0.988
Time 3	21.13	8	0.051	0.991	0.984
Time 4	15.58	8	0.04	0.995	0.991
Time 5	26.03	8	0.063	0.988	0.978
Joint Model	1161.65	425	0.047	0.938	0.924
Cognitive Competence	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	7.76	8	<0.0001	1.00	1.00
Time 2	32.39	8	0.076	0.978	0.983
Time 3	19.30	8	0.048	0.995	0.996
Time 4	18.88	8	0.048	0.993	0.995
Time 5	49.63	8	0.096	0.934	0.951
Joint Model	737.80	425	0.031	0.966	0.958
Difficulty	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	3.77	2	0.040	0.993	0.98
Time 2	1.60	2	< 0.0001	1.00	1.00
Time 3	8.17	2	0.070	0.986	0.957
Time 4	7.48	2	0.068	0.981	0.944
Time 5	0.18	2	< 0.0001	1.00	1.01
Joint Model	262.37	177	0.025	0.980	0.971
Value	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	96.29	26	0.070	0.930	0.903
Time 2	139.59	26	0.091	0.904	0.866
Time 3	95.29	26	0.065	0.954	0.937
Time 4	112.89	26	0.075	0.945	0.924
Time 5	135.22	26	0.086	0.922	0.892
Joint Model	1672.95	992	0.030	0.933	0.923
Interest	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*
Time 1	1.77	2	< 0.0001	1.00	1.00
Time 2	6.49	2	0.065	0.996	0.988
Time 3	6.15	2	0.058	0.997	0.990
Time 4	14.96	2	0.104	0.990	0.970
Time 5	21.55	2	0.131	0.983	0.950
Joint model	276.22	177	0.027	0.987	0.982

Note. FIMLχ^2 = Full Information Maximum Likelihood Chi-Square, *df* = degrees of freedom, *RMSEA* = Root Mean Square Error of Approximation, *CFI* = Comparative Fit Index, *NNFI* = Non-Normed Fit Index. Factor loadings for the joint measurement models were set equal over time.

Table 2

Fit indices for the structural models

Affect	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*	*AIC*	*BIC*
Autoregressive model (M1)	1557.37	446	0.056	0.907	0.890	1787.37	2323.92
Unidirectional model Affect to ER (M2)	1492.50	443	0.055	0.912	0.896	1728.50	2279.05
Unidirectional model ER to Affect (M3)	1361.09	444	0.051	0.923	0.909	1595.09	2140.98
Equality Constraints (M5)	1363.57	445	0.051	0.923	0.909	1595.57	2136.79
Bidirectional model (M4)	1338.75	441	0.051	0.925	0.911	1578.75	2138.63
Cognitive Competence (CC)	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*	*AIC*	*BIC*
Autoregressive model (M1)	1024.41	446	0.041	0.937	0.926	1254.41	1790.96
Unidirectional model CC to ER (M2)	966.44	443	0.039	0.943	0.932	1095.30	1641.19
Unidirectional model ER to CC (M3)	861.30	444	0.035	0.955	0.946	1095.3	1641.19
Equality Constraints (M5)	866.92	445	0.035	0.954	0.946	1098.92	1640.14
Bidirectional model (M4)	844.66	441	0.034	0.956	0.948	1084.66	1644.54
Difficulty (Diff)	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*	*AIC*	*BIC*
Autoregressive model (M1)	425.98	198	0.038	0.946	0.931	581.98	945.90
Unidirectional model Diff to ER (M2)	377.65	195	0.035	0.957	0.944	539.65	917.57
Unidirectional model ER to Diff (M3)	355.46	196	0.032	0.962	0.951	515.46	888.71
Equality Constraints (M5)	359.66	197	0.032	0.962	0.951	517.66	886.25
Bidirectional model (M4)	337.30	193	0.031	0.966	0.956	503.30	890.55
Value	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*	*AIC*	*BIC*
Autoregressive model (M1)	1782.23	163	0.031	0.924	0.915	2108.23	2868.74
Unidirectional model Value to ER (M2)	1770.49	166	0.031	0.925	0.926	2096.49	2857.00
Unidirectional model ER to Value (M3)	1743.10	165	0.030	0.928	0.919	2073.10	2842.94
Equality Constraints (M5)	1746.43	164	0.030	0.927	0.919	2074.43	2839.60
Bidirectional model (M4)	1739.38	168	0.030	0.928	0.919	2075.38	2859.22

Interest	*FIML*χ^2	*df*	*RMSEA*	*CFI*	*NNFI*	*AIC*	*BIC*
Autoregressive model (M1)	450.26	198	0.040	0.968	0.959	606.26	970.18
Unidirectional model Interest to ER (M2)	421.50	195	0.038	0.971	0.963	583.50	961.42
Unidirectional model ER to Interest (M3)	382.27	196	0.035	0.976	0.969	542.27	915.52
Equality Constraints (M5)	386.65	197	0.035	0.976	0.969	544.65	913.24
Bidirectional model (M4)	369.01	193	0.034	0.978	0.971	535.01	922.26

Note. FIMLχ^2 = Full Information Maximum Likelihood Chi-Square, *df* = degrees of freedom, *RMSEA* = Root Mean Square Error of Approximation, *CFI* = Comparative Fit Index, *NNFI* = Non-Normed Fit Index. ER = Exam Result.

Next, the baseline model was compared to the unidirectional "standard" (M2) and "reverse" (M3) models by respectively adding paths from statistics attitudes to exam results and vice versa. For all subscales except *Value*, as indicated by an improvement in the alternative fit indices, both unidirectional models provided a better fit to the data compared to the autoregressive model. However, a direct comparison of both unidirectional models revealed that the reverse model (M3) yielded a better fit to the data as indicated by lower AIC and BIC values.

For *Value,* the fit-values for both unidirectional models were not substantially better than those for the autoregressive model. In other words, for this subscale, adding the unidirectional paths did not seem to improve the model over the stability paths.

Subsequently, for all studied subscales except *Value*, the unidirectional model with exam effects (M3) was compared to a bidirectional model (M4) by adding paths from statistics attitudes to exam results. For all subscales, fit indices remained fairly similar, suggesting that the bidirectional model did not substantially improve upon the selected unidirectional model (Table 2).

Finally, comparing the fit-values of the unidirectional "reverse" model (M3) to those of a model in which the cross-lagged effects are constrained to be equal across measurement waves (M5), indicated that the null hypothesis of consistent cross-lagged effects was tenable. Specifically, *RMSEA* and *NNFI* values for the two models were very similar and differences in comparative fit indices never exceeded the proposed benchmark ($\Delta CFI \leq .01$). Hence, for all studied subscales except *Value*, a unidirectional model with

invariant exam result effects best represented the relationships between student attitudes and exam results (Table 2).

An inspection of the final models (Table 3) revealed that standardized autoregressive coefficients for attitudes ranged between .66 (R^2 = .44, p < .001) and .83 (R^2 = .69, p < .001). This suggests considerable stability of attitudes across time, but still leaves 17% to 34% of the variance to be accounted for by other factors, such as exam results. Exam results were less stable across time. Standardized path coefficients for all subscales were equal to .49 (R^2 = .24, p < .001) for the effect from Exam 1 to Exam 2 and .39 (R^2 = .15, p < .001) for the effect from Exam 2 to Exam 3. Over and above the autoregressive effects, small to medium positive effects ranging between .16 (R^2 = .03, p < .001) and .33 (R^2 = .11, p < .001) going from exam results to statistics attitudes were obtained. As expected, better exam results resulted in higher levels of subsequent attitudes.

Table 3

Standardized Coefficients for the Structural Paths of the Final Models

	Affect	Cognitive Competence	Difficulty	Value	Interest
Autoregressive coefficients					
SATS: Time 1 to Time 2	0.72	0.73	0.66	0.69	0.61
SATS: Time 2 to Time 3	0.76	0.78	0.83	0.80	0.76
SATS: Time 3 to Time 4	0.79	0.83	0.83	0.80	0.79
SATS: Time 4 to Time 5	0.76	0.80	0.82	0.78	0.74
ER: 1st year to second year	0.49	0.49	0.49	0.49	0.49
ER: 2nd year to third year	0.39	0.39	0.39	0.39	0.39
Unidirectional effects					
ER 1st year to SATS Time 3	0.33	0.30	0.23	-	0.21
ER 2nd year to SATS Time 5	0.28	0.25	0.18	-	0.16

Note. All coefficients are statistically significant at 0.001-level. Unidirectional effects are constrained to be equal within each subscale except *Value*. *Value* shows no unidirectional effects, because for this subscale the final model was the autoregressive model.

4 Discussion

Although in several studies a small to modest relationship between statistics attitudes and statistics achievement was observed (see Section 1.1), the directionality of this relationship remained largely unexplored. Many researchers assumed a unidirectional effect from statistics attitudes to achievement. We tested this assumption by comparing several models regarding the directionality of relationships between statistics attitudes and statistics achievement using a longitudinal design: (1) a unidirectional "standard" model representing the commonly mentioned hypothesis with effects from statistics attitudes to statistics achievement, (2) a unidirectional "reverse" model representing the opposite hypothesis with effects from statistics achievement to statistics attitudes, and (3) a bidirectional model representing the hypothesis that effects in both directions are present. Statistics attitudes were operationalized by the SATS-36 (Schau et al., 1995) and statistics achievement by statistics exam results. Because preparatory analyses yielded important results regarding the factorial stability of the SATS-36 and the relative stability of attitudes across time, before answering our main research question on the directionality of effects, these results will be discussed first.

4.1 Factorial stability of the SATS-36 across time

Tests of the measurement models for each subscale resulted in information on the factorial stability of the SATS-36, which is a critical requirement for making inferences about changes over time (Pitts, West, & Tein, 1996; see also Chapter 4). Because for the *Effort* subscale valid assessment was not guaranteed, this subscale was discarded from further analyses.

The tests of factorial stability (i.e., equal factor loadings, implying that the same construct is measured over time) showed that the hypothesis of equal factor loadings across five measurement waves was clearly justifiable. This result provides evidence that the SATS-36 factors have the same meaning across administrations. As a result, comparisons of relationships across time are meaningful because differences across time are not contaminated by differences in residual variation (Gregorich, 2006). Until now, factorial stability of the SATS across time was only demonstrated in two studies both involving only two administrations (Chiesi & Primi, 2009; Hilton et al., 2004).

4.2 Relative stability of attitudes and achievement across time

Students attitudes were very stable. That is, the primary predictor of statistics attitudes at Time 2 or later turned out to be statistics attitudes at previous times. Although students maintained a fairly similar rank order concerning attitudes toward statistics across the five measurements, there remained a substantive amount of variation to explain by other factors, such as statistics achievement. Similar observations regarding the stability of attitudes are made in other studies inspecting pretest-posttest correlations (e.g., Harlow et al., 2002; Rhoads & Hubele, 2000; Wise, 1985). Our result seems in line with the description of attitude as a generally stable construct (See Section 1.1). As mentioned in the introduction, attitudes are considered less stable than personality traits and beliefs, but more stable than emotions, which are reaction to a specific situation.

Relative positions regarding statistics achievement appeared to be less stable than statistics attitudes. This result is in line with Harlow et al. (2002), but opposite to the observation of Ma and Xu (2004) that stability effects were stronger for achievement than for attitudes. For all separate subscales that we investigated, 24% of the variation in the second year result was explained by the first year result, and 15% of the variation in the third year result was explained by the second year result. Because dropouts of the study performed significantly worse on their first year exam than non-dropouts, restriction of range might be a reason for the relatively low observed stability coefficients of statistics achievement.

4.3 Directionality of effects

An empirical comparison of different directionality models (autoregressive, “standard” unidirectional, “reverse” unidirectional, bidirectional) failed to support the common unidirectional interpretation of the relationship between statistics attitudes and achievement in which only effects from attitudes to achievement are assumed. Although we expected a bidirectional model in line with most studies in related fields (Calsyn & Kenny, 1977; Graham et al., 2007; Ma & Xu, 2004; Marsh & Yeung, 1997; Minato & Kamada, 1996), for the factors *Affect*, *Cognitive Competence*, *Difficulty,* and *Interest*, this study surprisingly supported a “reverse” unidirectional model with only exam result effects and no attitude effects. In other words, this study shows that statistics exam results were related to

subsequent attitudes rather than the other way around. Apparently, if university students succeed in their statistics exams and are able to make the right attributions about success, this explains their statistics attitudes about themselves as learners (Eccles & Wigfield, 2002).

For the subscales mentioned, the exam effects were consistent across measurement waves. Although achievement only explained a small to medium amount of variance of the subsequent attitudes, this result is remarkable because there was respectively a time interval of 3 and 14 months between the first and second year, and second and third year exams. Our result is in disagreement with the result of Harlow et al. (2002), who observed no effect from pre-course quantitative skills to post-course quantitative attitudes, but in line with Ma and Xu (2004), who found that mathematics achievement demonstrated predominance over mathematics attitudes. In terms of design and number of waves, the study of Ma and Xu (2004) is most similar to the present one. Other studies mostly only included maximum two waves.

For *Value* (attitudes about the usefulness, relevance, and worth of statistics in personal and professional life), taking into account exam result and/or attitude effects did not result in a substantial improvement as compared to the baseline, autoregressive model. For this subscale, during students' first three years of their curriculum, there were no cross-lagged effects between statistics attitudes and exam results in either direction. This deviating result for *Value* is in line with the expectations. In several earlier studies it was observed that attitudes regarding the usefulness of the domain of statistics are to a smaller extent related to achievement than attitudes regarding the self (such as the *Affect*, *Cognitive Competence*, and *Difficulty* subscales of the SATS-36) (Carmona, 2002; Chiesi & Primi, 2009; Rhoads & Hubele, 2000; Schutz et al., 1998; Sorge & Schau, 2002; Waters et al., 1988; Wise, 1985; see also Introduction). Achievement seems to affect how students' perceive themselves in relation to statistics education, but not how they value statistics as a subject. For instance, students that perform badly are more likely to have a low self-esteem in relation to statistics, but can still consider statistics to be very useful. Although the items of the *Difficulty* subscale are formulated in such a way that attitudes regarding the difficulty of statistics as a subject are assessed, apparently students' provide answers regarding *their own perception* of the difficulty, because results of this subscale are in line with other attitudes concerning self-perception in relation to statistics (*Affect* and *Cognitive Competence*).

In sum, all attitudes dimensions were quite stable in the present study. On the other hand, developments that did occur in attitudes regarding the self could be partly attributed to changes in achievement. Other attributes than achievement seem responsible for changes regarding the value of the domain of statistics. For all attitudes dimensions changes in achievement could not be attributed to changes in attitudes. Because no exam results data were available for students who dropped out, one can wonder whether restriction of range in later waves could also be a reason for the absence of any attitude effects in this study. Of course, although the design and analyses in this quasi-experimental study permit drawing conclusions regarding the directionality of effects, all causal claims when interpreting the effects stay tentative (Shadish, Cook, & Campbell, 2002). If the observed effects would be caused by other variables, no direct effect from better or worse achievement to attitudes would have been observed.

4.4 Strengths, limitations and suggestions for further research

The main merit of this study is the use of a longitudinal design that permits to extensively explore and compare alternative directionality hypotheses, which was impossible in earlier studies. This does not allow making claims about the causality of effects because it is possible that a third variable may account for the relationship between attitudes and achievement. While a major threat in longitudinal studies is dropout of participants, which often results – in educational studies – in restricting the sample to successful students, we have made a major effort to keep unsuccessful students in the sample too.

A limitation of this study was the nature of the sample. Although the data collection was extensive, it was limited to one – mainly female – group of students Educational Sciences and Speech Pathology and Audiology at one university. Results might be idiosyncratic and may not necessarily generalize to other educational contexts or Educational Sciences and Speech Pathology groups at other countries/universities. Replication and further research is needed.

4.5 Practical implications

The observed predominance of achievement effects on statistics attitudes has practical implications for the assessment of attitudes and achievement both in research and

in the classroom, more specifically regarding the sequence of assessment in predictive studies. To date, most earlier studies assessed prior attitudes to predict later achievement (see Section 1.3). The findings of this study show that the logic would be to, at least, include an assessment of attitudes also after achievement to allow an investigation of directionality.

Although the effects that emerge from exam results to statistics attitudes are small to modest, they can be viewed as substantively important because they identify a possible mechanism to modify otherwise relatively stable attitudes. Besides the fact that statistics achievement is the most important goal in statistics courses, this study suggests that throughout the whole course of a curriculum improving students' achievement in statistics also results in positive statistics attitudes beliefs about themselves as learners. Policy makers could take this observation into account when designing curriculum and evaluation policy. For teachers, it may be important to view and treat low achievers as an "at risk" group for a decline in attitudes related to the students' personal engagement and capacities for learning statistics.

Chapter 6

General conclusion and discussion

1 Summary and discussion of main results

In this final chapter, we first present an overview and discussion of the main findings from the present doctoral dissertation. Results regarding the two major domains of contributions are successively discussed: (1) psychometric properties of the SATS-36, and (2) the (directionality of the) relationship between statistics attitudes and statistics achievement. Because throughout the dissertation also different aspects of stability of attitudes were investigated, these results are separately reviewed. Afterwards, limitations of the research, future research suggestions, and recommendations for practice are provided.

1.1 Psychometric properties of the SATS-36

In the present doctoral dissertation we investigated psychometric properties of our Dutch translation of the Survey of Attitudes Toward Statistics (SATS-36; Schau, et al., 1995; Schau, 2003; Chapter 2-5), an increasingly used survey designed to measure students' statistics attitudes before and after statistics courses. The results from the five-wave longitudinal design of the study clearly add to the existing evidence, because the design of the study allowed us to test psychometric properties of the SATS-36 over time. Within our sample of 785 Educational Sciences and Speech Pathology and Audiology students from the Katholieke Universiteit Leuven three major psychometric aspects of the SATS-36 were investigated.

First, the **hypothesized six-dimensional structure** of the SATS-36 was investigated in more detail than previously (Chapter 3). Because earlier studies tested the structure by analyzing item parcels, individual item functioning and specific strengths or flaws of the survey could not be detected by those studies. Furthermore, although the technique of item parceling has its advantages, it remains controversial. Therefore, in the study presented in Chapter 3 individual items were analysed to examine the factor structure of the SATS-36. Although the hypothesized structure showed adequate fit to the data in an absolute sense, some theoretically sound model modifications (i.e., changes to the SATS-36 survey) resulted in substantive improvement of fit. Specifically, it was concluded that three *Difficulty* items should be deleted from the survey because they show less common variance with other *Difficulty* items and that two *Affect* items have more in common than represented by the *Affect* factor (therefore error covariance should be taken into account). These suggested modifications to the survey were consistently found on data from the same students across time and on data from a mixture of other students following various educational programs. Nevertheless, because this is the first study to analyse individual items, and because all data were gathered in a Flemish context and with a Flemish translation of the surveys, cross-validation of the modified models is essential to answer the question whether the suggested modifications generalize to other educational contexts or students from other countries.

Second, the presupposed six-dimensional structure was compared to a **four-dimensional structure** in which *Affect*, *Cognitive Competence*, and *Difficulty* were combined into one factor. Based on an exploratory factor analysis, Cashin and Elmore (2005) suggested that a more parsimonious solution was possible by integrating these three factors. According to them these three factors are measuring “aspects of how a student will perform in his or her statistics course” (p. 521), which is distinguished theoretically from “the value of statistics as a tool in students’ respective fields of study” (p. 522) (as in the *Value* subscale). As in research on mathematics education (McLeod, 1992), this relates to the distinction between students’ attitudes about the statistics domain (e.g., the value of statistics) and students’ attitudes about themselves as learners of statistics (e.g., self-efficacy regarding statistics). An explicit empirical comparison of the four- and six-dimensional models was never performed before. Conclusions regarding the comparison were nuanced. From a pure technical viewpoint, the original six-factor model demonstrated only a slightly better fit than the four-factor model, but from a more substantive viewpoint several reasons were

formulated to favour the original six-factor model. Because arguments for both solutions could be formulated, it was concluded that the choice may depend on the goals of a specific scientific study or specific educational setting (e.g., whether or not a more global description of attitudes suffices).

Third, **longitudinal measurement invariance** of the SATS-36 was examined (Chapter 4) on data from our five-wave longitudinal design. Both measurement aspects (invariance of factor configuration, factor loadings, intercepts, error variances) and aspects with respect to the latent variables (invariance of latent factor variances and means) were investigated. Longitudinal measurement invariance is essential for the evaluation of temporal change, which is often the goal in statistics attitude research. Without longitudinal measurement invariance being established, it cannot be determined whether observed temporal change is due to true change in the latent construct or to changes in measurement of the construct over time. Despite the importance of longitudinal measurement invariance, in the context of statistics attitudes the present doctoral dissertation was the first to study this topic in such detail. Evidence of weak invariance and (partial) strong invariance was found for all separate SATS-36 subscales, indicating that the six dimensions were measured in the same way across measurement times, and providing support for the SATS-36 as a useful survey for comparing statistics attitudes across time.

1.2 (Directionality of the) relationship between statistics attitudes and statistics achievement

Research has repeatedly shown a relationship between statistics attitudes and statistics achievement. The present doctoral dissertation also aimed to contribute to the existing research on (the directionality of) this relationship and by investigating it in a longitudinal context. In Chapter 2, the Attitudes Toward Statistics scale (ATS; Wise 1985) was used to investigate the **relationship between statistics attitudes and short- and long-term statistics exam results**. The findings confirmed the connection between statistics attitudes (before and after the introductory statistics course) and short-term statistics achievement. While for short-term exam results, attitudes toward the course were more highly related to statistics exam results than the attitudes toward the field (see also among others Carmona, 2002; Chiesi & Primi, 2009), the latter were higher related to the fifth-year dissertation grade

(used as an indication of long-term statistics achievement) than the attitudes toward the course.

More important than the observation of the mere correlations, the present dissertation extended the existing research on the relationship between statistics attitudes and statistics achievement by investigating the **directionality of effects** (Chapter 5). Such examination is crucial to understand the complex processes between affective and cognitive factors in the context of statistics classes. Most researchers up to now, not supported by appropriate (longitudinal) empirical data, assumed a unidirectional effect from statistics attitudes to statistics achievement. By gathering information on the same students over time, the present dissertation provided a unique opportunity to test several hypotheses regarding the directionality of the effects between attitudes and achievement in the context of statistics education.

Results were opposite to the common view: A unidirectional model with effects from statistics achievement to statistics attitudes was found for students' attitudes about themselves as learners of statistics. Although achievement only explained a small to medium amount of variance of the subsequent attitudes, in line with the theory of Eccles and Wigfield (2002), this seems to imply that if students succeed in their statistics exams and are able to make the right attributions about success, this explains their statistics attitudes about themselves as learners (Eccles & Wigfield, 2002). Effects in the other direction, from attitudes to achievement, were not observed in this doctoral dissertation.

Regarding attitudes about the domain of statistics, no effects over and above the stability effect of attitudes and achievement were present during students' progress through their curriculum. So, achievement seems to affect how students perceive themselves in relation to statistics education, but not how they value statistics as a domain. For instance, students who perform badly on statistics exams are more likely to have a low self-esteem in relation to statistics, but may still consider statistics to be very useful.

1.3 Stability of attitudes

Throughout the dissertation, interesting results regarding different aspects of stability of attitudes across the five measurement times have been reported. In this paragraph, these results are summarized. First, as already mentioned earlier in this chapter, **factorial stability** of the SATS-36 was investigated (Chapter 4 and 5). The SATS-36 subscales

proved to be stable across time (invariance of factor loadings, intercepts, error variances). This result provided evidence that the SATS-36 factors have the same meaning across administrations.

Second, in Chapter 5, high **relative stability** of statistics attitudes was observed. The high stability coefficients indicated that the primary predictor of statistics attitudes at Time 2 or later turned out to be statistics attitudes at previous times. In other words, students maintained a fairly similar rank order concerning statistics attitudes across the five measurements. Similar observations regarding the stability of statistics attitudes are made in other studies inspecting pretest-posttest correlations (e.g., Harlow et al., 2002; Rhoads & Hubele, 2000; Wise, 1985).

Third, based on an investigation of **mean stability** – in line with expectations described by Gal and Ginsburg (1994) – attitudes about the statistics domain appeared to be stable during students' curriculum. Students' attitudes about themselves as learners of statistics on the other hand were more susceptible to changes, but observed changes were rather small. Largest mean increases were not observed from the beginning to the end of a statistics courses, but from before to after students' knowing of their exam results (although still relatively small). This result was in line with educational theories such as the Expectancy-Value theory (e.g., Eccles & Wigfield, 2002; see Chapter 4). Decreasing attitude scores were observed from the beginning to the end of the second statistics course. This course dealt with design and sampling, probability and sampling distributions, as well as an introduction to statistical inference, concepts that are traditionally considered to be difficult for students (e.g., Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007).

In sum, this doctoral dissertation taught us that statistics attitudes are relatively stable, also over a longer period of time. These results seemed in line with the description of attitude as a generally stable construct (e.g., McLeod, 1992) and paralleled previous research that indicated that attitudes are resistant to change, especially after high school (e.g., Leong, 2006). Of course, the results on relative and mean stability at the group level presented in this doctoral dissertation do not preclude the possibility that individuals' behaviour may change with time.

2 Limitations and suggestions for further research

Although the current study extended the research domain on statistics attitudes in several ways, it is important to acknowledge its limitations. The present paragraph addresses some limitations and presents guidelines for further research.

A first limitation is that the study only relied on **Likert-type surveys**. Although they have the clear advantage that information can be gathered cost-effectively and time-efficiently, several drawbacks can be mentioned, such as the fact that they do not allow for investigation into the processes, experiences, and interpretations behind attitudes (Gal, Ginsburg, & Schau, 1997). The use of qualitative methods in future studies, like interviews, observations, open-ended surveys, and focus groups can provide deeper insights into the relationship between statistics attitudes and achievement in statistics (e.g., Leong, 2006).

A second limitation pertains to the **sample** of the present study. Although the data collection was extensive, it was mainly (not completely; see Chapter 3, Appendix II) limited to one (typically largely female) group of students Educational Sciences and Speech Pathology and Audiology at one Flemish university. Perhaps, results have to be understood within the specific context of the students and courses under investigation, and may not necessarily generalize to other educational contexts, to Educational Sciences and Speech Pathology groups at other countries/universities. Additional samples should be investigated to increase confidence in the conclusions described in this doctoral dissertation.

Third, we choose to isolate attitudes and achievement from the broader educational context in which they are embedded, to be able to study the properties of the SATS-36 and the directionality of associations between these two variables in detail. However, a wide array of factors can and should be considered when studying the role of affective factors in statistics education, such as student characteristics (e.g., gender, age, number of mathematics hours in secondary education, previous statistics experience), teacher characteristics (e.g., experience, didactical approach) and characteristics of the classroom or institution in which these factors interact. Unfortunately, because of time constraints, we could not make full advantage of the **design** of our study, in which information on many of these variables (and others) is available. Furthermore, at three measurement times, information on the evolution of statistical reasoning during students' progress in their curriculum is available by administrations of the Statistical Reasoning Assessment (Garfield, 2003). We plan to incorporate these aspects in future studies.

Future research can also focus on other aspects of the studies presented that received relatively little attention. Examples include an inspection of statistics attitudes of students who stopped their studies untimely or an investigation of students' individual growth trajectories of attitudes and/or achievement.

Finally, based on our studies, some specific recommendations for statistics education research (designs) are formulated throughout this doctoral dissertation, such as taking suggested modifications to the SATS-36 into account (Chapter 3), analyzing both individual items and item parcels to profit from advantages of both approaches (Chapter 3), including attitude assessments before and after exams and students' knowing their exam results (Chapter 4 and 5), and establishing measurement invariance before investigating attitude change (Chapter 4).

3 Recommendations for practice

In this final section, the main practical recommendations of our findings that follow from this doctoral dissertation are summarized. Throughout the chapters, further support was found that the SATS-36 proves to be an appropriate survey to efficiently assess students' statistics attitudes. However, when assessing students' attitudes, based on the results from this doctoral dissertation, statistics teachers should take the suggested improvements to the SATS-36 into account (e.g., deleting poor functioning items). Possibly, when our results will be replicated, it will be necessary to develop a new, partly adapted version of the SATS-36 that takes these suggested modifications into account.

Our results further made clear that, depending on the goals of a specific teacher or educational setting, a more parsimonious version of the SATS-36 can be used. Teachers who require a more global description of students' attitudes toward statistics, may without much loss of information integrate the *Affect*, *Cognitive Competence* and *Difficulty* scales in one global scale. In such case, in line with the distinction formulated by McLeod (1992) and Gal and Ginsburg (1994), the combination of the *Affect*, *Cognitive Competence* and *Difficulty* subscales represents students' attitudes about themselves as learners of statistics (or as formulated by Cashin and Elmore (2005): different aspects of how a student will perform in his or her statistics course), as opposed to attitudes about the statistics field itself.

Clearly, it cannot be expected from statistics teachers to investigate measurement invariance before assessing attitude change. Based on our results (and hopefully future

results from other studies), teachers can be confident that observed attitude changes (for instance from the beginning to the end of their statistics course) can be meaningfully interpreted.

As was the case for statistics researchers, statistics teachers interested in the evolution of attitudes and the interplay with achievement in their specific educational setting are encouraged to administer the SATS-36 not only before and after their statistics courses, but also before and after exams and knowledge of exam results. After all, based on our results on the directionality of effects, we suggested that rather than fostering positive attitudes because of their effect on achievement, improving students' achievement in statistics is a strategy for eliciting positive statistics attitudes about themselves as learners. Therefore, although the effects that emerge from exam results to statistics attitudes are small to modest, they can be viewed as substantively important because they identify a possible mechanism to modify otherwise relatively stable attitudes.

References

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology, 52*, 27-58.

Aldogan, A., & Aseeri, A. (2003). Psychometric characteristics of the Attitude Towards Statistics scale. *Umm Al-Qura University Journal of Educational and Social Sciences and Humanities, 15*(2), 99-114.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102.

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269-296), Hillsdale, NJ: Lawrence Erlbaum Associates.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2*, 127-155.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning and thinking: Goals, definitions and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Biehler, R. (1993). Software tools and mathematics education: The case of statistics. In C. Keitel & K. Ruthven (Eds.), Learning from computers: *Mathematics education and technology* (pp. 68-100). Berlin: Springer.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach.* Hoboken, NJ: Wiley.

Bontempo, D. E., & Mackinnon, A. (2006, July). *Measurement equivalence/invariance of the Developmental Behavior Checklist: Factorial invariance of categorical factor models.* Presentation at the 19th biannual meeting of the International Society for the Study for Behavioral Development. Melbourne, Australia.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago: Scientific Software International.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185-216.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: The Guilford Press.

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models*. New York: Springer.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial invariance. *Psychological Bulletin, 105*(3), 456-466.

Callaert, H. (2004). Statistiek in de nieuwe leerplannen van het secundair onderwijs: Een moderne aanpak [Statistics in the new secondary curriculum: A modern approach]. *Wiskunde en Onderwijs, 30*, 202-207.

Calsyn, R., & Kenny, D. (1977). Self-concept of ability and perceived evaluations by others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69*, 136-145.

Carmona, J., Martinez, R. J., & Sanchez, M. (2005). Mathematical background and attitudes toward statistics in a sample of undergraduate students. *Psychological Reports, 97*, 53-62.

Carnell, L. J. (2008). The effect of a student-designed data collection project on attitudes toward statistics. *Journal of Statistics Education, 16*(1). Retrieved from http://www.amstat.org/publications/jse/v16n1/carnell.html

Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics scale: A construct validity study. *Educational and Psychological Measurement, 65,* 1-16.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98-113.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2009). How confident are students in their misconceptions about hypotheses tests? *Journal of Statistics Education, 17* (2). Retrieved from http://www.amstat.org/publications/jse/v17n2/castrosotos.html

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Chiesi, F., & Primi, C. (2009). Assessing statistics attitudes among college students: Psychometric properties of the Italian version of the Survey of Attitudes Toward Statistics (SATS). *Learning and Individual Differences, 19*(2), 309-313.

Chiesi, F., & Primi, C. (in press). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ9(1)_Chiesi_Primi.pdf

Cobb, G. (2005). Foreword. In J. B. Garfield (Ed.), *Innovations in teaching statistics, MAA Notes*, Vol. 65 (pp. vii-viii). Washington DC: Mathematical Association of America.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 133,* 155-159.

D'Andrea, L., & Waters, C. (2002, July). *Teaching statistics using short stories: Reducing anxiety and changing attitudes.* Paper presented at the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Dauphinee, T. L., Schau, C., & Stevens, J. J. (1997). Survey of Attitudes Toward Statistics:Factor structure and factorial invariance for women and men. *Structural Equation Modeling*, *4*(2), 129-141.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. New York: Harcourt College Publishers.

Eagly, A. H., & Chaiken, S. (1995). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 413-432). Mahwah, NJ: Erlbaum.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.

Elmore, P. B., & Lewis, E. L. (1991, April). *Statistics and computer attitudes and achievement of students enrolled in applied statistics: Effect of a computer laboratory.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Elmore, P. B., Lewis, E. L., & Bay, M. L. G. (1993, April). *Statistics achievement: A function of attitudes and related experience.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Finkel, S. E. (1995). *Causal analysis with panel data.* Thousand Oaks, CA: Sage.

Finney, S. J., & DiStefano, C. (2006). Dealing with nonnormal and categorical data in structural equation models. In G.R. Hancock & R.O. Mueller (Eds.). *A second course in structural equation modeling* (pp. 269-313). Greenwich, CT: Information Age.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology, 28*, 161-186.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach.* New York: Psychology Press.

Flora, D. B., & Curran P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Gal, I., & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education, 2*(2). Retrieved from http://www.amstat.org/publications/jse/v2n2/gal.html

Gal, I., Ginsburg, L., & Schau, C. (1997). Monitoring attitudes and beliefs in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 37-51). The Netherlands: IOS Press.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22-38. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(1).pdf

Garfield, J. B., Hogg, B., Schau, C, & Whittinghill, D. (2002, July). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education*, *10*(2). Retrieved from http://www.amstat.org/publications/jse/v10n2/garfield.html

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2008). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest, 8*, 53-96.

Graham, S., Berninger, V., & Fran, W. (2007). The structural relationship between writing attitude and writing achievement in first and third grade students. *Contemporary Educational Psychology, 32*(3), 516-536.

Green, J., Nelson, G., Martin, A. J., & Marsh, H. (2006). The causal ordering of self-concept and academic motivation and its effect on academic achievement. *International Education Journal, 7*(4), 534-546.

Greer, B. (2000). Statistical thinking and learning. *Mathematical Thinking and Learning, 2*, 1-9.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*, 78-94. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1808350

Harlow, L. L., Burkholder, G. J., Morrow, J. A., & Morrow, J. A. (2002). Evaluating attitudes, skill, and performance in a learning-enhanced quantitative methods course: A structural equation modeling approach. *Structural Equation Modeling, 9*(3), 413-430.

Hatcher, L. (1994). *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling.* Cary, NC: SAS Institute

Hau, K. T., & 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.

Heckhausen, H. (1991). *Motivation and action.* Berlin: Springer-Verlag.

Hilton, S. C., Schau, C., & 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.

Hodgson, T., & Burke, M. (2000). On simulation and the teaching of statistics. *Teaching Statistics, 22*, 91-96.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158-176). Thousand Oaks: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1)*, 1-55.

Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351-362.

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R.W. Robins, R.C. Fraley, & R.F. Krueger (Eds.), *Handbook of research methods in personality psychology (*pp. 461-494). New York: Guilford.

Jöreskog, K. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, NJ: Sage.

Jöreskog, K. G. (1999). *Formulas for skewness and kurtosis. Unpublished manuscript.* Retrieved from http://www.ssicentral.com/lisrel/techdocs/kurtosis.pdf

Jöreskog, K. G. (2005). *Structural equation modeling with ordinal variables using LISREL*. Unpublished paper. Retrieved from http://ssicentral.com/lisrel/corner.htm

Jöreskog, K. G., & Sörbom, D. (2004). *LISREL 8.7 for Windows [Computer Software]*. Lincolnwood, IL: Scientific Software International, Inc.

Keeler, C. M., & Steinhorst, R. K. (1995). Using small groups to promote active learning in the introductory statistics course: A report from the field. *Journal of Statistics Education, 3*(2). Retrieved from http://www.amstat.org/publications/jse/v3n2/keeler.html

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: The Guilford Press.

Kottke, J. L. (2000). Mathematical proficiency, statistics knowledge, attitudes toward statistics, and measurement course performance. *College Student Journal, 34*, 344-347.

Lalonde, R. N., & Gardner, R. C. (1993). Statistics as a second language? A model for predicting performance in psychology students. *Canadian Journal of Behavioural Science, 25*, 108-125.

Leong, J. L. (2006). High school students' attitudes and beliefs regarding statistics in a service-learning based statistics course. *Unpublished doctoral dissertation.* USA, Georgia State University.

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.

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. International *Journal of Behavioral Development, 31*, 357-365.

Ma, X., & Xu, J. (2004). Determining the causal ordering between attitude toward mathematics and achievement in mathematics. *American Journal of Education, 110*(3), 256-280.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201-226.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

Marsh, H.W., & Yeung, A.S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology, 89*, 41-54.

März, V., Vanhoof, S., Kelchtermans, G., & Onghena, P. (2010). De vernieuwing van het statistiekonderwijs in Vlaanderen: Percepties en betekenisgeving in het implementatieproces. *Pedagogische Studiën, 87*, 134-151.

März, V., Vanhoof, S., & Onghena, P. (2010). De plaats van statistiek in het onderwijs: in of uit de wiskundeles?. *Impuls voor Onderwijsbegeleiding, 40*, 170-177.

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 575-596). New York: Macmillan.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172-175.

Mills, J. (2004). Students' attitudes toward statistics: Implications for the future. *College Student Journal, 38*, 349-362.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10*(1), retrieved from www.amstat.org/publications/jse/v10n1/mills.html

Minato, S., & Kamada, T. (1996). Results of research studies on causal predominance between achievement and attitude in junior high school mathematics of Japan. *Journal for Research in Mathematics Education, 27*, 96-99.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*, 123-165.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the Practice of Statistics* (5th ed.). New York: Freeman.

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology, 25*(1), 3-53.

Murtonen, M. (2005). *Learning of quantitative research methods - Univeristy students' views, motivation, and difficulties in learning.* Annales Universitatis Turkuensis, Ser. B., Tom. 287.

Mvududu, N. (2003). A cross-cultural study of the connection between students' attitudes toward statistics and the use of constructivist strategies in the course. *Journal of Statistics Education,* 11(3). Retrieved from http://www.amstat.org/publications/jse/v11n3/mvududu.html

Nasser F., & Wisenbaker J. (2003). A Monte Carlo study investigating the impact of item Parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement, 63,* 729–757.

Nasser, F. (1999). Prediction of statistics achievement. *Proceedings of the International Statistical Institute 52nd Conference*, Helsinki, Finland, pp. 7-8.

Nasser, F. (2004). Structural model and effects of cognitive and affective factors on the achievement of Arabic-speaking pre-service teachers in introductory statistics. *Journal of Statistics Education, 12*(1). Retrieved from http://www.amstat.org/publications/jse/v12n1/nasser.html

Olson, J. M., & Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology, 44*, 117-154.

Onwuegbuzie, A., & Wilson, V. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments – a comprehensive review of the literature. *Teaching in Higher Education, 8*, 195–209.

Op ‘t Eynde, P., De Corte, E., & Verschaffel, L. (2006) “Accepting emotional complexity”: A socio-constructivist perspective on the role of emotions in the mathematics classroom. *Educational Studies in Mathematics, 63*, 193-207.

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology: An International Review, 41*, 359–376.

Pitts, S. C., West, S. G., & Tein, J. Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning, 19*, 333-350.

Raftery, A. E. (1995). Bayesian model selection in social research. In A. E. Raftery (Ed.), *Sociological Methodology* (pp. 111-164). Oxford: Blackwell.

Rhoads, T. R., & Hubele, N. F. (2000). Student attitudes toward statistics before and after a computer-integrated introductory statistics course. *IEEE Transactions on Education, 43*(2), 182-187.

Roberts, D. M., & Saxe, J. E. (1982). Validity of a statistics attitude survey: A follow-up study. *Educational and Psychological Measurement, 42*, 907-912.

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and validity of a statistics attitude survey. *Educational and Psychological Measurement, 40,* 235-238.

Roberts, D. M., & Reese, C. (1987). A comparison of two scales measuring attitudes towards statistics. *Educational and Psychological Measurement, 47*, 759-764.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Schau, C. (2003). *Students' attitudes: The "other" important outcome in statistics education*. Paper presented at the Joint Statistical Meetings, San Francisco, CA. Retrieved from http://evaluationandstatistics.com/JSM2003.pdf

Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement, 55*, 868-875.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginners' guide to Structural Equation Modelling.* Mahwah, NJ: Lawrence Erlbaum.

Schutz, P. A., Drogosz, L. M., White, V. E., & Distefano, C. (1998). Prior knowledge, attitude, and strategy use in an introduction to statistics course. *Learning & Individual Differences, 10*, 291-309.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957-1009). Greenwich, CT: Information Age Publishing, Inc.

Shultz, K. S., & Koshino, H. (1998). Evidence of reliability and validity for Wise's Attitude Toward Statistics scale. *Psychological Reports, 82,* 27-31.

Simon, J. L. (1994). What some puzzling problems teach about the theory of simulation and the use of resampling. *American Statistician, 48,* 290-293.

Sorge, C., & Schau, C. (2002). *Impact of engineering students' attitudes on achievement in statistics.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

SSICentral (s.d.). *The diagnosis and treatment of non-normality.* Unpublished manuscript. Available online at http://www.ssicentral.com/lisrel/techdocs/Session4.pdf

Suanpang, P., Petocz, P., & Kalceff, W. (2004). Student attitudes to learning business statistics: Comparison of online and traditional Methods. *Educational Technology & Society, 7*(3), 9-20.

Tempelaar, D. T., van der Loeff, S. 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. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2).pdf

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Vanhoof, S., Castro Sotos, A. E., Onghena, P., & Verschaffel, L. (2007). Students' reasoning about sampling distributions before and after the Sampling Distribution Activity. In International Statistical Institute (Ed.), *Proceedings of the 56th Session of the International Statistical Institute*. Lisbon, Portugal.

Vanhoof, S., Castro Sotos, A. E., Onghena, P., Verschaffel, L., Van Dooren, W., & Van den Noortgate, W. (2006, November). Attitudes toward statistics and their relationship with short- and long-term exam results. *Journal of Statistics Education, 14(3)*. Retrieved from http://www.amstat.org/publications/jse/v14n3/vanhoof.html

Wang, Y., & Liu, Q. (2006). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research, 77*, 220–225.

Waters, L. K., Martelli, T. A., Zakrajsek, T., & Popovich, P. M. (1988). Attitudes toward statistics: An evaluation of multiple measures. *Educational and Psychological Measurement, 48,* 513-516.

Waters, L. K., Martelli, T. A., Zakrajsek, T., & Popovich, P. M. (1989). Measuring attitudes toward statistics in an introductory course on statistics. *Psychological Reports, 64*, 113-114.

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement, 45*, 401-405.

Wisenbaker, J. M., & Scott, J. S. (1995, April). *Attitude about statistics and achievement in introductory statistics course.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wisenbaker, J. M., & Scott, J. S. (1997, March). *Modeling aspects of students' attitude and achievement in introductory statistics course.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Wisenbaker, J., Nasser, F., & Scott, J. (1998, June). *A multicultural exploration of the interrelation among attitude about and achievement in introductory statistics.* Paper presented at the Annual Meeting of the International Conference of Teaching Statistics, Singapore.

Zembylas, M. (2007a). Theory and methodology in researching emotions in education. *International Journal of Research & Method in Education, 30*(1), 57-72.

Zieffler, A. S. (2006). A longitudinal investigation of the development of college students' reasoning about bivariate data during an introductory statistics course. *Unpublished doctoral dissertation*, USA, University of Minnesota.

Appendix

ATTITUDES TEN OPZICHTE VAN STATISTIEK (Wise, 1985)

Richtlijnen: Duid voor elk van volgende stellingen de categorie aan die het beste weergeeft hoe je je momenteel voelt t.o.v. de stelling. Beantwoord alstublieft alle stellingen.

	Sterk Oneens	Oneens	Neutraal	Eens	Sterk Eens
1. Ik denk dat statistiek mij van nut zal zijn in mijn beroep.	_____	_____	_____	_____	_____
2. De gedachte deel te nemen aan een cursus statistiek maakt me nerveus.	_____	_____	_____	_____	_____
3. Een goeie onderzoeker moet een opleiding in statistiek hebben.	_____	_____	_____	_____	_____
4. Statistiek lijkt mij iets heel mysterieus te zijn.	_____	_____	_____	_____	_____
5. De meeste mensen zouden er voordeel uit halen een cursus statistiek te volgen.	_____	_____	_____	_____	_____
6. Ik heb het moeilijk om het verband te zien tussen statistiek en mijn studiedomein.	_____	_____	_____	_____	_____
7. Deelnemen aan een cursus statistiek ervaar ik als een heel onaangename ervaring.	_____	_____	_____	_____	_____
8. Ik zou graag mijn statistische vorming voorzetten in een cursus voor gevorderden.	_____	_____	_____	_____	_____
9. Statistiek zal voor mij nuttig zijn in het vergelijken van relatieve verdiensten van verschillende objecten, methoden, programma's, etc.	_____	_____	_____	_____	_____
10. Statistiek is niet echt heel nuttig omdat het ons vertelt wat we al weten.	_____	_____	_____	_____	_____
11. Een statististische opleiding is relevant voor mijn prestatie binnen mijn studiedomein.	_____	_____	_____	_____	_____
12. Ik wou dat ik had kunnen vermijden deel te nemen aan mijn cursus statistiek.	_____	_____	_____	_____	_____

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
13. Statistiek is een onderdeel van mijn opleiding dat de moeite waard is.	_____	_____	_____	_____	_____
14. Statistiek is te veel op wiskunde georiënteerd om van veel nut te zijn voor mij in de toekomst.	_____	_____	_____	_____	_____
15. De gedachte aan een volgende cursus statistiek te moeten deelnemen maakt me overstuur.	_____	_____	_____	_____	_____
16. Statistische analyse zou beter overgelaten worden aan experts en zou geen deel mogen uitmaken van het werk van een professioneel die een leek is op gebiek van statistiek.	_____	_____	_____	_____	_____
17. Statistiek is onafscheidbaar van wetenschappelijk onderzoek.	_____	_____	_____	_____	_____
18. Ik voel me geïntimideerd wanneer ik moet werken met wiskundige formules.	_____	_____	_____	_____	_____
19. Ik kijk ernaar uit om statistiek daadwerkelijk te kunnen gaan gebruiken in mijn beroep.	_____	_____	_____	_____	_____
20. Statistiek studeren is tijdverspilling.	_____	_____	_____	_____	_____
21. Mijn opleiding in statistiek zal me het onderzoek dat binnen mijn studiedomein gebeurt, beter helpen te begrijpen.	_____	_____	_____	_____	_____
22. Men kan resultaten van onderzoek effectiever ‘verteren’ als men enige noties van statstiek heeft gekregen.	_____	_____	_____	_____	_____
23. Een opleiding in statistiek zorgt voor een professionele ervaring die beter afgerond is.	_____	_____	_____	_____	_____
24. Statistisch denken kan een nuttige rol spelen in het alledaagse leven.	_____	_____	_____	_____	_____
25. Met getallen werken, maakt me ongemakkelijk.	_____	_____	_____	_____	_____
26. Ik vind dat statistiek vroeg in een professionele opleiding moet verplicht worden.	_____	_____	_____	_____	_____
27. Statistiek is voor mij te ingewikkeld om effectief te kunnen gebruiken.	_____	_____	_____	_____	_____
28. Statistische training is niet echt nuttig voor de meeste professionelen.	_____	_____	_____	_____	_____

29. Statistisch denken zal ooit even belangrijk zijn Voor een goed burgerschap als kunnen lezen en schrijven. _____ _____ _____ _____ _____

DANKUWEL VOOR JE HULP!

Vragenlijst Attitudes ten opzichte van Statistiek (1)

SATS-R Pre

RICHTLIJNEN: De beweringen hieronder zijn ontwikkeld om je attitudes ten opzichte van statistiek na te gaan. De schaal heeft 7 mogelijke antwoorden, gaande van 1 (sterk oneens), via 4 (eens noch oneens) tot 7 (sterk eens). Als je geen mening hebt, kies dan antwoord 4. Lees en beantwoord alstublieft elke vraag. Duid telkens het ene antwoord aan op de 7-punten schaal dat het meest jouw attitude ten opzichte van de stelling weergeeft. Gebruik de hele 7-punten schaal om je overeenstemming of gebrek aan overeenstemming met onze stellingen weer te geven. Denk niet te diep na over elk antwoord. Duid je antwoord aan en ga snel over naar de volgende stelling.

	Sterk oneens			Eens noch oneens			Sterk eens
1. Ik ben van plan om al mijn statistische taken te voltooien.	1	2	3	4	5	6	7
2. Ik ben van plan hard te werken voor mijn cursus statistiek**.**	1	2	3	4	5	6	7
3. Ik zal statistiek leuk vinden.	1	2	3	4	5	6	7
4. Ik zal me onzeker voelen wanneer ik statistische problemen moet oplossen.	1	2	3	4	5	6	7
5. Ik zal moeite hebben om statistiek te begrijpen, door de manier waarop ik denk.	1	2	3	4	5	6	7
6. Statistische formules zijn gemakkelijk te begrijpen.	1	2	3	4	5	6	7
7. Statistiek is waardeloos.	1	2	3	4	5	6	7
8. Statistiek is een ingewikkeld onderwerp.	1	2	3	4	5	6	7
9. Statistiek zou een verplicht onderdeel van mijn professionele opleiding moeten zijn.	1	2	3	4	5	6	7
10. Statistische vaardigheden zullen me meer inzetbaar maken op de arbeidsmarkt.	1	2	3	4	5	6	7

	Sterk oneens			Eens noch oneens			Sterk eens
11. Ik zal er geen idee van hebben waar het allemaal om gaat bij statistiek.	1	2	3	4	5	6	7
12. In ben geïnteresseerd om ertoe in staat te zijn statistische informatie naar anderen te communiceren.	1	2	3	4	5	6	7
13. Statistiek is niet nuttig voor een doorsnee deskundige.	1	2	3	4	5	6	7
14. Ik ben van plan hard te studeren voor elke statistiektest.	1	2	3	4	5	6	7
15. Ik zal gefrustreerd geraken wanneer ik examens statistiek moet afleggen.	1	2	3	4	5	6	7
16. Statistisch redeneren is niet toepasbaar in mijn leven buiten mijn beroep.	1	2	3	4	5	6	7
17. Ik gebruik statistiek in mijn dagelijks leven.	1	2	3	4	5	6	7
18. Ik zal onder stress staan tijdens de statistieklessen.	1	2	3	4	5	6	7
19. Ik zal het plezierig vinden om cursussen statistiek te volgen.	1	2	3	4	5	6	7
20. Ik ben geïnteresseerd in het gebruiken van statistiek.	1	2	3	4	5	6	7
21. Statistische conclusies worden zelden gepresenteerd in het dagelijkse leven.	1	2	3	4	5	6	7
22. Statistiek is een onderwerp dat door de meeste mensen snel geleerd wordt.	1	2	3	4	5	6	7
23. Ik ben geïnteresseerd in het begrijpen van statistische informatie.	1	2	3	4	5	6	7
24. Het leren van statistiek vereist heel wat discipline.	1	2	3	4	5	6	7

	Sterk oneens			Eens noch oneens			Sterk eens
25. In mijn beroep zullen er geen toepassingen van statistiek nodig zijn.	1	2	3	4	5	6	7
26. Ik zal veel wiskundefouten maken bij statistiek.	1	2	3	4	5	6	7
27. Ik ben van plan om elk college en practicum van statistiek te volgen.	1	2	3	4	5	6	7
28. Ik heb schrik van statistiek.	1	2	3	4	5	6	7
29. Ik ben geïnteresseerd in het leren van statistiek.	1	2	3	4	5	6	7
30. Statistiek gaat gepaard met enorme berekeningen.	1	2	3	4	5	6	7
31. Ik kan statistiek leren.	1	2	3	4	5	6	7
32. Ik zal statistische vergelijkingen begrijpen.	1	2	3	4	5	6	7
33. Statistiek is irrelevant voor mijn leven.	1	2	3	4	5	6	7
34. Statistiek is hoogst technisch.	1	2	3	4	5	6	7
35. Ik zal het moeilijk vinden om statistische concepten te begrijpen.	1	2	3	4	5	6	7
36. De meeste mensen moeten een nieuwe manier van denken aanleren om aan statistiek te doen.	1	2	3	4	5	6	7

(zie ook volgende bladzijde)

MERK OP dat de labels voor de schalen voor elk van volgende schalen verschilt van bovenstaande schalen.

37. Hoe goed bracht je het ervan af bij je wiskundecursussen in het secundair onderwijs?	Heel slecht 1	2	3	4	5	6	Heel goed 7
38. Hoe goed ben je in wiskunde?	Heel slecht 1	2	3	4	5	6	Heel goed 7
39. Hoeveel statistiek zal je gebruiken in het domein waarin je zou willen tewerk gesteld worden nadat je afstudeert?	Helemaal Geen 1	2	3	4	5	6	Heel veel 7
40. Hoe zelfzeker ben je dat je inleidend statistisch materiaal onder de knie kan krijgen?	Helemaal niet zelfzeker 1	2	3	4	5	6	Heel zelfzeker 7
41. Hoeveel computerervaring heb je?	Geen 1	2	3	4	5	6	Heel veel 7
42. Hoeveel ervaring met statistiek heb je (bijv. cursussen, onderzoeksstudies)?	Geen 1	2	3	4	5	6	Heel veel 7
43. Hoe zou je, in het algemeen, de statistische vaardigheden van jongens en meisjes vergelijken?	Meisjes veel beter 1	2	3	Meisjes en jongens ongeveer even goed 4	5	6	Jongens veel beter 7

BEDANKT VOOR JE HULP!