A STRUCTURAL EQUATION MODEL ANALYZING THE RELATIONSHIP OF STUDENTS' ATTITUDES TOWARD STATISTICS, PRIOR REASONING ABILITIES AND COURSE PERFORMANCE

DIRK T. TEMPELAAR
Maastricht University, The Netherlands
D.Tempelaar@ke.unimaas.nl

SYBRAND SCHIM VAN DER LOEFF
Maastricht University, The Netherlands
S.Loeff@ke.unimaas.nl

WIM H. GIJSELAERS
Maastricht University, The Netherlands
W.Gijselaers@erd.unimaas.nl

ABSTRACT

Recent research in statistical reasoning has focused on the developmental process in students when learning statistical reasoning skills. This study investigates statistical reasoning from the perspective of individual differences. As manifestation of heterogeneity, students' prior attitudes toward statistics, measured by the extended Survey of Attitudes Toward Statistics (SATS), are used (Schau, Stevens, Dauphinee & DeVecchio, 1995). Students' statistical reasoning abilities are identified by the Statistical Reasoning Assessment (SRA) instrument (Garfield 1996, 1998a, 2003). The aim of the study is to investigate the relationship between attitudes and reasoning abilities by estimating a full structural equation model. Instructional implications of the model for the teaching of statistical reasoning are discussed.

Keywords: *Statistics education research; Statistical reasoning; Achievement motivations; SATS; SRA; Structural equation modelling*

1. INTRODUCTION

Recent research into statistical reasoning about variation, distribution, and sampling distributions has created important insights into the developmental process of statistical reasoning skills. Most research has focused on the identification of subsequent, hierarchically-ordered stages of reasoning development by means of qualitative research methods such as thinking-aloud sessions and in-depth interviews. Two recent special issues of this journal (*SERJ*, Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005) and an edited volume (Ben-Zvi & Garfield, 2004a) contain a wealth of such empirical studies into the cognitive process of developing reasoning abilities and of instructional tools that might foster these developments. The present research investigates statistical reasoning from a somewhat different perspective. It examines individual differences among students learning statistics and statistical reasoning. These individual differences

demonstrate much variability: Students enter learning processes with different background characteristics and different perceptions of the learning context. As a manifestation of students' heterogeneity, this study uses students' prior attitudes toward statistics. The main aim of this study is to investigate the relationship between students' attitudes toward statistics and their prior statistical reasoning abilities when entering an introductory statistics course.

Contemporary research in statistics education distinguishes an array of different but related cognitive processes in learning statistics: statistical literacy, statistical reasoning, and statistical thinking. See for example the special section of the *Journal of Statistics Education* (Short, 2002), the two special issues of *SERJ* (Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005), Ben-Zvi and Garfield (2004a), and Pfannkuch and Wild (2004). The demarcation of these three cognitive processes not being complete, it is well accepted that statistical literacy represents the most basic skills (Ben-Zvi & Garfield, 2004c). Gal (2004) distinguishes two interrelated components in statistical literacy: the ability to "interpret and critically evaluate statistical information, data-related arguments, and stochastic phenomena," and the ability to "discuss or communicate" these (see also Rumsey, 2002). Statistical reasoning is the ability to "explain why a particular result is expected or has occurred, or explain why it is appropriate to select a particular model or representation" (delMas, 2004a; see also Garfield & Chance, 2000; Garfield, 2002). Statistical thinking involves an "understanding of why and how statistical investigations are conducted and the 'big ideas' that underlie statistical investigations" (Ben-Zvi & Garfield, 2004; see also Pfannkuch & Wild, 2004; Chance, 2002). Literacy, reasoning, and thinking are to some extent achieved even before formal schooling in statistics takes place. Those naïve conceptions learned outside school can be correct or incorrect in nature.

In the 1970s, cognitive research into statistical and probabilistic reasoning revealed several categories of fallacies in human reasoning, with examples such as the 'Law of small numbers,' the 'Representativeness misconception' (Kahneman, Slovic, & Tversky, 1982), the 'Outcome orientation' (Konold, 1989), and the 'Equiprobability bias' (Lecoutre, 1992). Most of that research is documented in the seminal work of Kahneman et al. (1982), as cited in Garfield and Ahlgren (1988). In the decades thereafter, following the reform movement in statistics education, research shifted its focus from probabilistic reasoning to reasoning with data (Pfannkuch & Wild, 2004), as evidenced in the topics of the recent series of SRTL research forums and the compilation of their major contributions in Ben-Zvi and Garfield (2004a).

Another important development in recent decades is the design of assessment instruments for statistical literacy, reasoning, and thinking (delMas, 2002; Garfield & Ben-Zvi, 2004a). Paraphrasing Chance (2002), 'if not assessed, it cannot be valuable,' and assessment instruments were needed to match the focus on literacy, reasoning, and thinking. Several instruments also grew out of the need for assessment tasks that could be used in the context of research projects. Quantitative assessment instruments are still scarce, and are all derived from the first and most prominent instrument in the field: Statistical Reasoning Assessment (SRA). The SRA was developed by Konold and Garfield (Konold, 1989; Garfield, 1996, 1998a, 2003) as part of a project evaluating the effectiveness of a new statistics curriculum in U.S. high schools. The instrument is based on the well-described classes of misconceptions and their antipodes, the learned or unlearned correct conceptions, that emerged from the cognitive science research into reasoning fallacies (Garfield, 2003; Garfield & Ahlgren, 1988). In current terminology – the SRA was developed long before recent discussions on the demarcation of literacy, reasoning, and thinking – fallacies addressed in the SRA are of all three types. Being

designed in the earlier stages of the reform movement in statistics education (Ben-Zvi & Garfield, 2004c), the SRA focuses both on statistical and probabilistic reasoning. Newer assessment instruments, related to the SRA but focusing more strongly on reasoning with data, are currently being developed in the framework of the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project (delMas, 2004b; see also https://app.gen.umn.edu/artist/). As newer instruments were not yet available, the SRA was the most appropriate tool at the time of this study to assess students' reasoning abilities in the large-scale applications typical of educational practice.

Empirical studies on statistical reasoning focus predominantly on the cognitive developmental process students go through when learning reasoning abilities, and on the instructional tools that may foster these developments. The large majority of these studies are empirical in nature in that they use descriptions, often achieved by thinking-aloud sessions or interviews of the cognitive states of students, to reconstruct a developmental trajectory (Ben-Zvi & Garfield, 2004a). Garfield and Ben-Zvi (2004b, p. 399) ascertain "It may seem strange, given the quantitative nature of statistics, that most of the studies … include analyses of qualitative data, particularly videotaped observations or interviews." Yet such studies allow identification of different states of students' reasoning abilities and subsequent stages in the developmental process. Our study chooses a different perspective based on individual differences in student-related factors by investigating the role of non-cognitive individual differences in the cognitive development of students. This type of study has, at least in the context of statistics and mathematics education, a long tradition (Gal & Garfield, 1997; McLeod, 1992). In conceptualizing the non-cognitive domains of education, McLeod (1992) distinguishes among emotions, attitudes and beliefs. In most studies of learning processes in statistical education, the focus is on beliefs and attitudes, rather than emotions; see for example Gal and Ginsburg (1994) and Gal and Garfield (1997). Probably the best known, and certainly most validated, model on the role of attitudes in learning statistics is the model developed by Schau and co-authors (Schau, Stevens, Dauphinee & DeVecchio 1995). The Schau-model is based on the expectancy-value model for achievement motivations designed by Eccles and Wigfield (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000, 2002). In that model, students' expectancies for success and the value they contribute to succeeding are important determinants of their motivation to perform achievement tasks. Expectancy for success crystallizes in two different concepts: belief in one's own ability to perform a task, and a perception of the task demand. Subjective task value is generally modeled in a single concept, comprising several aspects: attainment value (importance of doing well on a task), intrinsic value (interest in and enjoyment gained from doing the task), utility value (usefulness), and costs (spent efforts) (Eccles, 2005). The contribution of Schau and co-authors to the development of the expectancy-value model of achievement motivations is two-fold. First, they designed the SATS measurement instrument to adapt the generic expectancy-value model to the statistical domain (Schau et al., 1995; Dauphinee, Schau & Stevens, 1997). Second, they extended the generic model by introducing new concepts obtained by disentangling the broad task-value concept of the expectancy-value model. In the first 28-item version of SATS, the task-value concept is broken up into an affective concept, focusing most on the enjoyment aspect of intrinsic values, and a valuation concept, focusing on the remaining components of attainment and utility values. The model of the first version thus contains two expectancy factors that deal with students' beliefs about their own ability and perceived task difficulty, Cognitive Competence and Difficulty, and two subjective task-value concepts that encompass students' feelings toward and attitudes about the value of the subject, Affect and Value (Schau, 2003). Empirical research, both within the statistics

domain (Dauphinee et al., 1997; Sorge & Schau, 2002; Hilton, Schau, & Olsen, 2004) and in other academic domains (Tempelaar, Gijselaers, Schim van der Loeff, & Nijhuis, 2007) supports the distinction of these affective and valuation aspects. In a second, 36-item version of SATS (C. Schau, personal communication, November 30, 2003), two more concepts are introduced: Interest and Effort. The Interest concept shapes the interest aspect of the intrinsic value component in the expectancy-value model, whereas the Effort concept shapes the perceived costs component in the subjective task-value (Eccles, 2005). To the knowledge of the authors, no empirical studies based on the extended SATS instrument have yet been published. Empirical studies of the 28-item version of SATS, referred to above, focus on the structure of attitudes alone, or on the structure of attitudes in relation to statistics course performances. The context of these studies is thereby slightly different from most studies in the expectancy-value framework that focus primarily on the relation between attitudes and learning task choices (such as course selection) rather than learning task outcomes.

The main contribution of this paper is to investigate the dependency of students' prior reasoning abilities on their attitudes toward statistics when entering an introductory statistics course. In the formulation of this research question, attitudes are hypothesized to be causal to statistical reasoning abilities. The hypothesized direction of causality is in agreement with process models of learning (see for example Garfield, Hogg, Schau, & Whittinghill, 2002), in which affective, student-related factors are regarded as determinants for cognitive, learning-outcome-related factors. In addition, attitudinal variables possess a trait-like nature, in contrast to reasoning abilities that possess a state-like nature. Therefore, the hypothesized causal direction follows the general modeling pattern of stable traits determining malleable states. In order to do so we start the empirical third section by developing confirmatory latent factor models for attitudes, based on the extended SATS instrument, and for statistical reasoning, based on the SRA instrument. Subsequently, these factor models are integrated into a full structural equation model that explains reasoning abilities by attitude factors. To be able to put this relationship into perspective, two further cognitive constructs are added to this model: course performance measured by quiz and final exam scores. This extension allows characterizing reasoning abilities not only by their direct relationship with attitudinal variables, but also by a comparison of that relationship with the ones between attitudes and course performances.

One of the implications of our model is that where different learning approaches provide alternative routes to achieve traditional course performances, perhaps one more efficiently than the others but all contributing to the same learning goal, this seems not to be true for statistical reasoning abilities. Some learning approaches really hinder achievement of reasoning skills. The model outcomes thus have strong implications for the development of instructional programs in statistical reasoning, which is one of the topics discussed in the concluding section.

2. METHOD

2.1. PARTICIPANTS AND PROCEDURE

In this study, the statistical reasoning of students participating in the "International Business" and "International Economics" programs of the Maastricht University was investigated. A large number of students, 842 and 776 respectively, from these two programs participated in the first year, first semester course Quantitative Methods (QM) in 2004/05 and 2005/06. This is a compulsory introduction to mathematics and statistics

for all students. Of these 1618 students, 64% were male and 36% were female. Another relevant decomposition was that 39% students had a Dutch secondary school diploma, versus 61% students with non-Dutch diplomas (most of them of German nationality).

Part of the data analyzed in this study comes from regular student quizzes and examinations. In the QM course, three assessment instruments are applied. One is a final exam, in multiple-choice format, covering both statistics and mathematics. Items in the exam focus on students' ability to apply statistical and mathematical methods; those in statistics are motivated by the Advanced Placement Statistics exams (e.g., http://apcentral.collegeboard.com). Secondly, both for statistics and mathematics, three quizzes are taken spread over the eight weeks of the course. Quizzes are optional; they give rise to bonus points for the exam score. In practice, all students participate in most of the quizzes. For this study, quiz scores are aggregated over the three quizzes. The third assessment instrument is a student project. For this project, students collect personal data by completing several self-report instruments concerning their study approach and preferred strategies. Later on, they perform an explorative analysis of these data. Students are informed that the self-reported data are also used for three additional purposes: to provide study advice to students who have adopted an inefficient study approach, for course-improvement purposes, and for research. The project is compulsory, and assessed with pass/fail. Because students can acquire feedback on their project in several stages of its development, the final assessment of it is not very informative, and is not included in this study.

The SATS and the SRA were the first self-report instruments to be administered during the first days of the course. Responses to both surveys therefore reflect students' prior attitudes and beliefs toward statistics and their prior reasoning abilities. Scores cannot be influenced by impressions of the educational process, nor by knowledge achieved in the course itself.

Both instruments are quantitative in nature, and generate observations that can be regarded as proxies for the underlying, but unobservable, theoretical constructs. Therefore, the investigation of the relationship between attitudes and reasoning abilities requires the estimation of two confirmatory latent factor models for attitudes on the one side, and for statistical reasoning on the other, as well as the integration of both these factor models into a full structural equation model. To this model, we add two indicators of course performance: latent variables measuring the strongly cognitive-based scores in the final exam, and the more effort-based scores in quizzes. The primary reason for doing so is that it allows for characterization of the particular position statistical reasoning takes within the spectrum of different performance indicators.

2.2. MEASURES

Statistical reasoning abilities The Statistical Reasoning Assessment (SRA) is a test consisting of 20 multiple-choice or multiple-answer items developed by Konold and Garfield as part of a project evaluating the effectiveness of a new statistics curriculum in U.S. high schools (Konold, 1989; Garfield, 1996, 1998a, 2003). Each item in the SRA describes a statistics or probability problem and offers four to eight choices of responses. Most responses include a statement of reasoning, explaining the rationale for a particular choice. For every item, one response corresponds to a category of correct reasoning; all or most of the other responses correspond to categories of misconceptions. For a full description of the individual items and the eight correct reasoning scales and eight misconceptions scales, see Garfield (1998a, 2003); Table 1 summarizes the scales of the description of the individual items and the eight correct reasoning scales and eight

Table 1. SRA Correct reasoning scales and misconceptions scales; based on Garfield (2003).

Correct Reasoning Scales:
Prob: Correctly interprets probabilities. Assesses the understanding and use of ideas of randomness and chance to make judgments about uncertain events.
Aver: Understands how to select an appropriate average. Assesses the understanding of what measures of center tell about a data set, and which are best to use under different conditions.
Comp: Correctly computes probability, both understanding probabilities as ratios, and using combinatorial reasoning. Assesses the knowledge that in uncertain events not all outcomes are equally likely, and how to determine the likelihood of different events using an appropriate method.
Indep: Understands independence.
Sampl: Understands sampling variability.
Correl: Distinguishes between correlation and causation. Assesses the knowledge that a strong correlation between two variables does not mean that one causes the other.
2Way: Correctly interprets two-way tables. Assesses the knowledge of how to judge and interpret a relationship between two variables, knowing how to examine and interpret a two-way table.
LrgS: Understands the importance of large samples. Assesses the knowledge of how samples are related to a population and what may be inferred from a sample; knowing that a larger, well-chosen sample will more accurately represent a population; being cautious when making inferences made on small samples.
Misconception scales:
AverMc: Misconceptions involving averages. This category includes the following pitfalls: believing averages are the most common number; failing to take outliers into consideration when computing the mean; comparing groups based on their averages only; and confusing mean with median.
OutcO: Outcome orientation. Students use an intuitive model of probability that leads them to make yes or no decisions about single events rather than looking at the series of events; see Konold (1989).
High%: Good samples have to represent a high percentage of the population. Size of the sample and how it is chosen are not important, but it must represent a large part of the population to be a good sample.
Small: Law of small numbers. Small samples best resemble the populations from which they are sampled, so are to be preferred over larger samples.
Repre: Representativeness misconception. In this misconception the likelihood of a sample is estimated based on how closely it resembles the population. Documented in Kahneman, Slovic, & Tversky (1982).
Cause: Correlation implies causation.
EquiPr: Equiprobability bias. Events of unequal chance tend to be viewed as equally likely; see Lecoutre (1992).
Groups: Groups can be compared only if they have the same size.

description of the individual items and the eight correct reasoning scales and eight misconceptions scales, see Garfield (1998a, 2003); Table 1 summarizes the scales of the instrument. In the design process of the instrument, the authors included several stages directed at achieving good validity and reliability. With regard to criterion-related validity, Garfield (2003) reports extremely low correlations with different course outcomes, suggesting statistical reasoning and misconceptions are unrelated to course performance. In addition, Garfield (2003) reports satisfactory test-retest reliabilities, but

low internal consistency reliability coefficients, implying that scales and misconception scales respectively appear not to measure one single ability or trait.

In terms of the classification into the more recently developed categories of statistical literacy, reasoning, and thinking, the allocation of individual reasoning abilities and misconceptions to these three classes is not obvious. Aver, TWay, AverMc, High%, and Groups refer to basic data-related skills, and seem to fit best in the literacy category. At the other extreme, Comp, Sampl, Correl, Small, Cause, and EquiPr involve probability and statistical theory related concepts, and might better suit the thinking category. The remaining scales, referring to notions of probability and uncertainty, would then fit the reasoning category. We return to this issue when discussing descriptive statistics of SRA data obtained from this study and a limited number of other studies that provide empirical data on the instrument: Garfield (1998b, 2003), Garfield and Chance (2000), Liu (1998) and Sundre (2003).

Attitudes and beliefs toward statistics Attitudes are measured with the Survey of Attitudes Toward Statistics (SATS) developed by Schau and co-authors (Schau et al., 1995; Dauphinee et al., 1997). There are two existing versions of the SATS, both consisting of seven-point Likert-type items measuring aspects of post-secondary students' statistics attitudes. The 28-item version of SATS contains four scales, as indicated below. Each scale is accompanied by two examples of items, one positively and one negatively worded:

- Affect (six items) - measuring positive and negative feeling concerning statistics, the enjoyment aspect of intrinsic value: *I like statistics*; *I am scared by statistics*.
- Cognitive Competence (six items) - measuring attitudes about intellectual knowledge and skills when applied to statistics, the self-concept of one's ability component in the expectancy-value model: *I can learn statistics*; *I have no idea of what's going on in statistics*.
- Value (nine items) - measuring attitudes about the usefulness, relevance, and worth of statistics in personal and professional life, the utility and attainment components of task value: *I use statistics in my everyday life*; *I will have no application for statistics in my profession*.
- Difficulty (seven items) - measuring attitudes about the difficulty of statistics as a subject, the perception of the task demand: *Statistics formulas are easy to understand*; *Statistics is highly technical*.

Schau et al. (1995), Dauphinee et al. (1997), and Harris and Schau (1999) elaborate on the development process of the instrument. The instrument is freely available from the internet (Schau, Dauphinee, Del Vecchio, & Stevens, 1999). Validation research in two very large samples of undergraduate students has shown that a four-factor structure provides a good description of responses to the SATS-instrument (Dauphinee et al., Hilton et al., 2004).

Recently, Schau has developed a 36-item version of the SATS, containing two additional scales, each covered by four, positively worded, items (Schau, personal communication, November 30, 2003). These scales, with one item example, are

- Interest (four items) - students' level of individual interest in statistics, the interest aspect of intrinsic value: *I am interested in learning statistics*.
- Effort (four items) - amount of work the student expends to learn statistics, the perceived cost component of task value: *I plan to work hard in my statistics course*.

2.3. DATA ANALYSIS

Parceling The very first step in the data analysis is to reverse the negatively worded items in the SATS instrument, such that for all items a higher score corresponds to a more positive attitude. This step is worthwhile to mention because it requires attentiveness in the interpretation of the construct Difficulty. High scores for Difficulty express a more positive attitude, implying that a better name for the Difficulty scale would have been 'perceived lack of difficulty.' The second step of analysis is the parceling of the SATS data, following earlier empirical work by Schau and co-authors (Schau et al., 1995; Dauphinee et al., 1997; Hilton et al., 2004). The technique of item parceling, where items from the same subscale are aggregated into several parcels or miniscales, has been adopted in empirical studies for several reasons: to obtain more continuous and normally distributed observed data, to reduce the number of model parameters to achieve a more attractive variable to sample size ratio, and to get more stable parameter estimates (Bandalos, 2002; Hau & Marsh, 2004; Marsh, Hau, Balla, & Grayson, 1998).

In parceling items, Hau and Marsh (2004) advise not to reduce the number of indicators for each latent construct beyond a minimum of three. Next, they recommend to counterbalance skewness in the presence of strong non-normality by creating parcels out of item pairs with opposite skew. In order to determine the relevance of this recommendation of counterbalancing skewness for our data set, the degree of non-normality of the data was calculated as a preliminary step to parceling. In the data of the first four SATS factors, no indications of non-normality were found in any of the self-reported questionnaires beyond Hau and Marsh's (2004) category of 'moderately non-normal,' implying skew = 1.0 and kurtosis = 1.5. Items corresponding to the constructs Interest and especially Effort were however much more strongly skewed.

In the empirical analyses of their 28-item SATS data, Schau et al. (1995), Dauphinee et al. (1997), and Hilton et al. (2004) adopt an item parceling scheme based on balancing with respect to the positively and negatively worded items, size of parcel means, standard deviations, and skew (see Schau et al.). Their parceling solution contains two parcels for Affect, Cognitive Competence, and Difficulty each; only Value contains three. Given the rule of thumb of at least three parcels per factor and the advice to counterbalance skew as much as possible, it was decided to apply a parceling scheme different from Schau and co-authors, based only on skewness, and resulting in exactly three parcels per factor.

Statistical analyses This study integrates several techniques of structural equation modeling (SEM). A SEM model is distinct from a path or regression model in that it hypothesizes that crucial variables, such as attitudes in this study, are not directly observable and are better modeled as latent variables than as observable ones. In doing so, a SEM model makes it possible to distinguish two different types of errors: errors in equations, as does the path model, and errors in the observation of variables. Making this distinction is especially worthwhile when errors in important constructs have rather different sizes. Studying reliabilities of several achievement motivations, and their variation over subjects, suggests that this argument applies to this study. In this study, SEM models were estimated with LISREL (version 8.54) using maximum likelihood estimation. For further discussion of SEM see for example Byrne (1998), Kline (2005), and Schumacker and Lomax (2004).

The standard approach to estimate a SEM distinguishes two steps (Schumacker & Lomax, 2004). In the first phase of the two-step model building approach, measurement models for all latent variables in the model are estimated. Measurement models are in general factor models that allow factors, also called traits, and the uniqueness, that is the

errors in indicators, to be correlated. In our study, we need to estimate three of such 'correlated trait' (CT), 'correlated uniqueness' (CU), and 'confirmatory factor analysis' (CFA) models: for the SATS data, for the SRA data, and for course performance data. In the second model building step, the structural part of the SEM is estimated. This structural part specifies the relationships between the independent and dependent latent variables. In contrast to the estimation of the measurement models, the estimation of structural relationships is to some extent explorative in nature. The structural part of the full structural equation model is not a priori restricted, except for several hypotheses with regard to the direction of the relationship. For the estimation of these structural parts, two different model modification procedures are applied. The first is called model trimming (Kline, 2005) or backward search (Schumacker & Lomax, 2004). Starting from a full matrix of structural path coefficients, one by one, parameters are restricted to zero if they prove non-significant, until all remaining structural parameters are significant. The second approach is called model building (Kline, 2005) or forward search (Schumacker & Lomax, 2004). It starts from a zero matrix of structural paths coefficients, and frees parameters one by one, in the order indicated by the value of the modification indices, up to point where no more significant improvement in fit is achieved. Because in both approaches subsequent models are nested, the chi-square difference statistic can be used to assess model fit. In all five subjects, both forward and backward searches converge to the same final model. Model modification is a form of explorative analysis, and brings along the risk of capitalization on chance.

With large sample sizes as in our study, the χ^2 test statistic is known to always reject in any formal test of significance (Byrne, 1998; Marsh & Yeung, 1996). For that reason, and following Marsh and Yeung (1996), and Hilton et al. (2004), emphasis is placed on the Root Mean Square Error of Approximation (RMSEA), the Goodness-of-Fit Index (GFI), the Non-Normed Fit Index (NNFI; termed Tucker-Lewis Index or TLI in Marsh & Yeung, 1996), the Comparative Fit Index (CFI) and the Relative Fit Index (RFI, termed Relative Noncentrality Index or RNI in Marsh & Yeung, 1996), and the normed version of the χ^2 test statistic: χ^2/df. For the last index, no clear-cut guidelines exist; values in the range of 2.0 to 5.0 are acceptable, with lower values indicating better fit. For RMSEA, values ≤ 0.05 indicate good fit, values ≤ 0.08 indicate reasonable fit. The indices GFI, NNFI, CFI, and RFI, all normally lie in the range 0.0 – 1.0, with higher values indicating better fit. As a benchmark for good fit, the value 0.90 is often used (Kline, 2005).

The covariance matrixes required for estimation are available from the authors upon request.

3. RESULTS

3.1. DESCRIPTIVE STATISTICS OF ATTITUDES AND BELIEVES TOWARD STATISTICS

Descriptive statistics of the SATS scales are exhibited in Table 2 and Figure 1. All attitudes are measured using a Likert 1-7 scale. Because all scale means, except for Difficulty, are larger than the neutral value of four, students in our sample express positive attitudes toward statistics for Affect, Cognitive Competence, Value, Interest, and Effort. Means and standard deviations are in line with values reported in Schau (2003) found as pre-test scores in a large class of undergraduate U.S. students; Affect, Cognitive Competence, and Value are slightly more positive in our sample, Difficulty is equal. In comparing our European data with data from U.S. studies, it is important to realize that participants in our study are all in economics and business programs. These programs

require students to take math classes in high school through at least intermediate level. Cronbach α reliability coefficients of these four scales are satisfactory, and again in line with intervals of values reported in Schau (2003) from several empirical studies by Schau and co-researchers. No empirical studies exist at this moment that incorporate the new scales of the 36-item SATS version: Interest and Effort. In our study, both these attitudes are clearly positive on average, with (planned) Effort taking a very strong position with a mean of 6.37 on a 1-7 scale. Figure 1 indicates that due to the high scores on Effort, skewness is an issue for this scale, and not for the other scales.

Table 2. Scale means, standard deviations, and Cronbach α's for attitudes toward statistics in our study (n=1458) and as reference, values reported in Schau (2003)

	Mean (Standard deviation)		Cronbach α	
	this study	Schau (2003)	this study	Schau (2003)
Affect	4.52 (1.10)	4.03 (1.14)	0.82	0.80 – 0.89
Cognitive Competence	5.08 (0.89)	4.91 (1.09)	0.78	0.77 – 0.88
Value	5.05 (0.83)	4.86 (1.01)	0.78	0.74 – 0.90
Difficulty	3.59 (0.77)	3.62 (0.78)	0.68	0.64 – 0.81
Interest	5.07 (0.99)		0.80	
Effort	6.37 (0.72)		0.76	

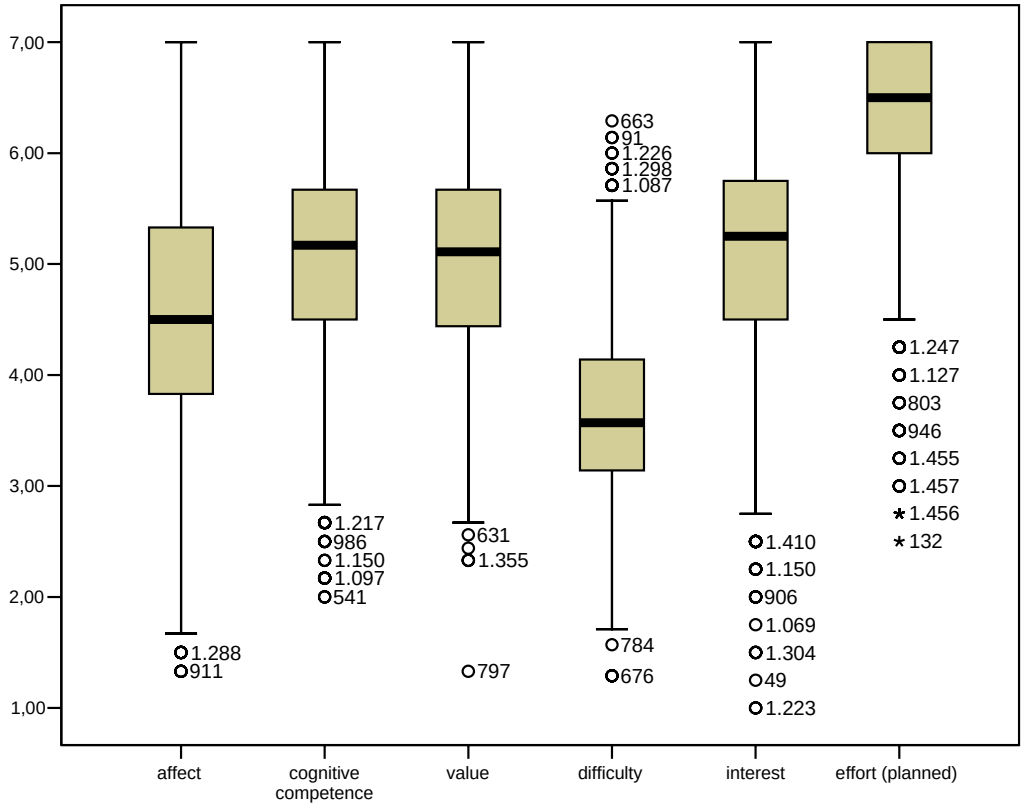

Figure 1. Descriptives of SATS scales (n=1458)

3.2. DESCRIPTIVE STATISTICS OF STATISTICAL REASONING ABILITIES

Descriptive statistics of the SRA data, similar to those reported in Garfield (1998b, 2003), Garfield and Chance (2000) and Liu (1998), are exhibited in Table 3. Because the maximum score of the several scales varies with the total number of answer options corresponding to the scale, the table presents the means of the several scales expressed as a proportion, that is, on a [0-1] scale. In addition to scores on eight reasoning skills, and eight misconceptions, the aggregated correct reasoning score (Correct) and aggregated misconceptions (Misconcep) are reported. The aggregated scores are obtained in the same way as in the studies by Garfield and co-authors by taking the sum over all correct reasoning and misconception items, and re-expressing them as a proportion. Because the number of items per scale ranges from 1 to 5, different scales have a different weight in the total score, so aggregated scores are to be regarded as weighted averages. Data reported by Garfield and co-authors are restricted to means.

Table 3. Scale means and standard deviations for statistical reasoning abilities in our study (n=1499) and as reference, post-course values US college students reported in Garfield (2003)

	Mean (Standard deviation)			Mean (Standard deviation)	
	this study	Garfield (2003)		this study	Garfield (2003)
Prob	0.75 (0.29)	0.68	AverMc	0.46 (0.27)	0.30
Aver	0.71 (0.27)	0.61	OutcO	0.22 (0.17)	0.23
Comp	0.40 (0.25)	0.46	High%	0.15 (0.23)	0.09
Indep	0.64 (0.29)	0.63	Small	0.28 (0.27)	0.29
Sampl	0.28 (0.30)	0.22	Repre	0.12 (0.22)	0.17
Correl	0.66 (0.47)	0.51	Cause	0.28 (0.37)	0.10
Twow	0.74 (0.40)	0.65	EquiPr	0.57 (0.33)	0.56
LrgS	0.71 (0.33)	0.68	Groups	0.29 (0.46)	0.60
Correct	0.58 (0.13)	0.55	Misconcep	0.29 (0.10)	0.27

Outcomes of our study and those reported in Garfield (2003) are remarkably similar, although the composition of groups of participating students is rather different. Garfield's study refers to U.S. college students surveyed at the end of an introductory course statistics, our study to European university students at the start of such an introductory course. Of the correct reasoning scales, Prob and Twow are amongst those with highest mastery level, and Comp and Sampl with lowest. Of the misconception scales, EquiPr and Groups are high in all studies (in our sample, Groups somewhat less), and High%, Repre and Cause are low.

Conceptions for which we find higher scores than reported in the Garfield studies are Aver, Correl, and Twow. The misconception for which our data indicate a remarkably low relative score is Groups. Of these four scales, three are characterized earlier as being part of the category of statistical literacy. This agrees with the difference in timing of the instrument, as a pre-test in our study, and a post-test in other studies. Not (recently) educated in introductory statistics, it is not surprising that students in our study score relatively high on statistical literacy components, but low on a statistical thinking related component as MC6 (correlation implies causation), typically an important concept to be taught in an introductory course.

As a last observation on average levels of reasoning skills and misconceptions, the high rate of correct answers is noticeable. Of the eight correct reasoning skills, five have

means of above 65% correct. Of the eight misconception scales, only two have means larger than 30%.

3.3. MEASUREMENT MODEL OF ATTITUDES AND BELIEFS TOWARD STATISTICS

As a first step in the modeling of the SATS data, an explorative factor analysis was performed (principal components, varimax rotation). The eigenvalue criterion identifies six factors. The scree-criterion demonstrates a large jump at four factors, and a smaller jump at six factors. The newly created scales Interest and Effort clearly qualify as independent factors. The same is true for the scale Value. However, items in the scales Affect, Cognitive Competence, and Difficulty are strongly correlated. This finding coincides with other empirical studies on SATS: Schau et al. (1995), Dauphinee et al. (1997), Hilton et al. (2004), and Cashin and Elmore (2005). On the basis of these high correlations, Cashin and Elmore (2005) decide to reduce the three scales Affect, Cognitive Competence, and Difficulty into one latent factor, whereas in the other three studies they are modeled as separate, but correlated, latent factors. We followed the last approach estimating a six-factor confirmatory factor model on parcelled attitudes data allowing a correlated traits (CT) structure but without cross-loadings in the factor loading matrix and no correlated uniqueness (CU) factor. Table 4 contains fit indices of this CT factor model, Figure 2 the structure of the factor model, including estimated trait correlations.

Table 4. Fit indices of six-factor correlated traits confirmatory factor models of attitudes toward statistics

	χ^2	*df*	RMSEA	GFI	NNFI	CFI	RFI
CT 6CFA model	701.80	123	.057	.95	.97	.97	.96

Fit indices indicate that the hypothesized correlated traits factor model fits the data quite well. Having confirmed the six-factor model, the correlation structure of latent factors depicted in Table 5 deserves prime interest. Table 5 demonstrates that twelve out of fifteen trait correlations are significant. Only three trait correlations appear to be non-significant and are restricted to zero in the estimation of the final version of the factor model, with the other correlations freed.

Table 5. Estimated latent factor correlations of attitudes toward statistics

	Affect	Cognitive Competence	Value	Difficulty	Interest	Effort
Affect	1.00					
Cognitive Competence	0.80	1.00				
Value	0.40	0.43	1.00			
Difficulty	0.61	0.62	-	1.00		
Interest	0.42	0.35	0.63	-	1.00	
Effort	-	0.17	0.34	-0.28	0.44	1.00

Note. All reported correlations are significant at $p < 0.000001$.

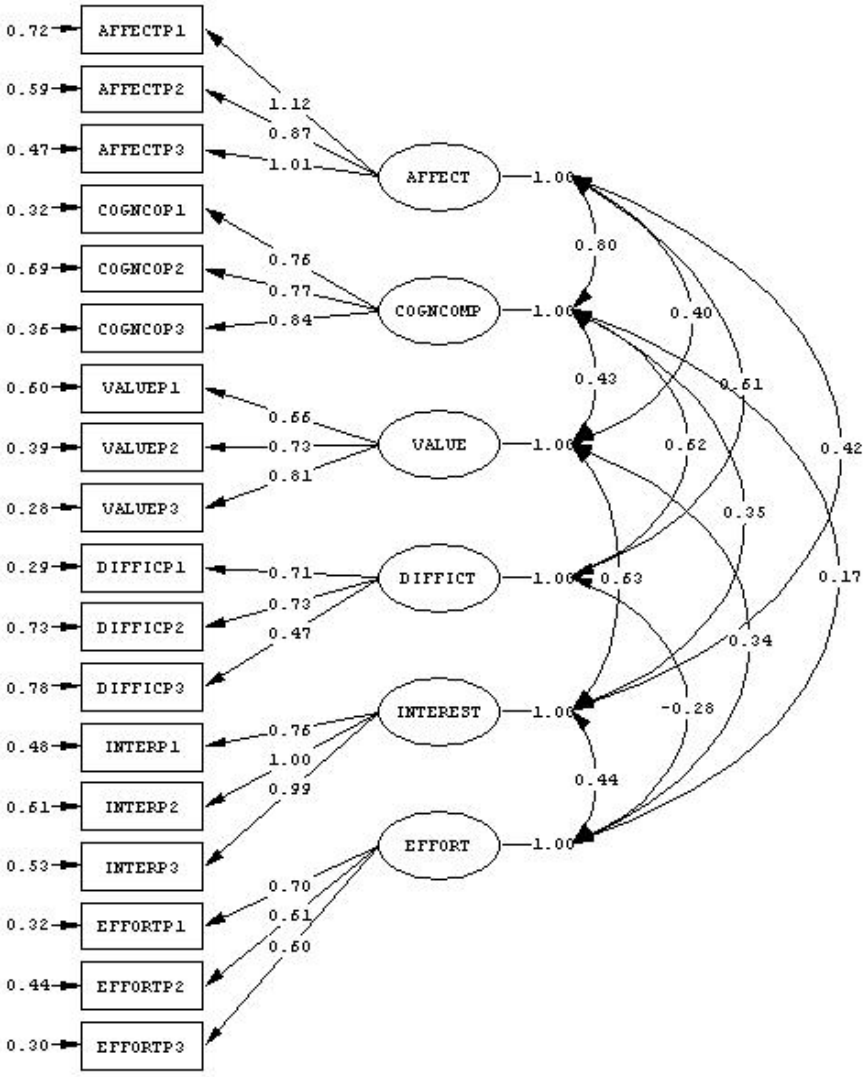

Figure 2. Correlated traits factor model as measurement model for attitudes toward statistics. Values are standardized parameter estimates. All values shown are statistically significant, $p < 0.05$. AFFECT, Affect; COGNC, Cognitive Competence; VALUE, Value; DIFFIC, Difficulty; INTEREST, interest; EFFORT, (planned) effort.

When interpreting the trait correlation structure, the first issue that comes up is the effect of disentangling the broad task value concept into Affect, related to liking the subject, and Value, related to the importance attached to the subject. The correlation between latent factors Affect and Value (r = 0.40) is, relative to other correlations, modest. This indicates that Affect and Value are clearly empirically distinguishable constructs. The correlation between Value and Difficulty is insignificant, indicating that the attached value is independent to the lack of perceived difficulty. A third observation refers to the by far largest correlation, namely between Affect and Cognitive Competence. This is in itself a remarkable fact: Affect is achieved by decomposing the task value component into affective and utility-related factors, but from this analysis it appears that Affect is much more strongly related to the expectancy component Cognitive

Competence, than to Value. This once again confirms the usefulness of the affect extension of the expectancy-value model. The strong correlation we found is comparable to the results found in Dauphinee et al. (1997) and Hilton et al. (2004).

The relationship between the two factors Interest and Effort and the other four factors is primarily through Value. Interest is unrelated to Difficulty, and Effort is unrelated to Affect and negatively related to Difficulty. That last negative relationship seems to be an consequence of rational study behavior; students who regard statistics as difficult plan to invest more study effort than students regarding the subject as less difficult. However, it is at odds with the expectancy-value model, where that relation has the opposite sign. The different outcome is best explained by the context in which the model is used; whereas the expectancy-value model is primarily based on the selection of learning tasks (such as choosing one course in favor of another), the context of this study is the intensity of performance, given the required learning tasks. In the expectancy-value model, Effort is assumed to be an intermediate outcome variable. For this interpretation to be true, the correlations between Effort and its predictors are expected to be strongly positive. This is not the case, except possibly for Interest. Two potential explanations for the weaker than expected relationship between Effort and its predictors are available. First, Effort is an ex-ante measure, and planned effort might quite well diverge strongly from ex-post measured, realized effort. Second, planned Effort scores seem to be a composition of two rather different underlying mechanisms that can make the relationships of this variable to other attitudinal constructs ambiguous. On the one side, students with high achievement motivation are assumed to spend large efforts in their learning, so planned effort acts as a proxy for achievement motivation. On the other side, planned effort might act as a proxy for students' learning approaches; students with a tendency to a memorizing type of learning tend to invest more effort in their learning than students with a learning approach focused on understanding. In general, the latter deep learning approach is regarded as better, and at least more efficient, than the first mentioned surface learning approach. For that reason, it might be expected that students with a tendency towards deep learning will have more positive attitudes, making deep learning positively related to the several attitudinal variables, and surface learning negatively related. If this is true, the relationship between Effort and attitudinal variables is the result of two counterbalancing forces: higher planned effort levels when being motivated, but lower planned effort levels when relying on efficient, deep learning approaches. In the subsection discussing the outcomes of the full structural equation model, we further elaborate on this issue.

3.4. MEASUREMENT MODEL OF STATISTICAL REASONING ABILITIES

Previous empirical studies of the SRA instrument have used aggregated correct conceptions, and aggregated misconceptions, as scales, with the eight correct reasoning ability scores and the eight misconception scores as items. This would suggest a measurement model with the two aggregated reasoning abilities as latent constructs, and the correct reasoning ability and aggregated variables as indicators. However, Garfield (1998b), Garfield and Chance (2000) and Liu (1998) point out that this modeling approach has important drawbacks. In their studies, as in ours, the correlations between reasoning ability scores are low, mostly insignificant, and quite often of opposite signs. This is problematic in terms of scale construction, because it gives rise to low values of instrument reliability. In the present data set analyzed in this study, the Cronbach-α reliability of the correct reasoning scales is 0.34, whereas for the misconception scales, the reliability α is 0.10. These values are too low to warrant meaningfulness of aggregated constructs. Elsewhere, we have investigated the reliability of aggregated

scales for a much larger sample, and have come to similar conclusions (Tempelaar, 2004). Deleting individual items with extreme p-values, as suggested in Liu (1998), appears to have little impact on reliabilities in our data.

Inspection of the correlation matrix depicted in Table 6 does however expose a pattern in correlations that suggests an alternative approach for modelling the outcomes of the SRA-instrument. Correlations within the group of correct reasoning scales, and within the group of misconceptions are, without exception, low. However, in the rectangular part of the correlation matrix containing the correlations between correct reasoning skills and misconceptions, seven out of eight columns contain exactly one highly significant and strongly negative correlation. This is not surprising; from the definition of for example Prob and OutcO, it is apparent that outcome orientation, that is the use of an intuitive and incorrect probability model, is at odds with correctly interpreting probabilities. And in some cases, the strong negative correlations between several correct conceptions and misconceptions find their origin in the fact that the concepts are based on different options of the same multiple choice items, which would lead to negative correlations by construct (although several multiple choice items allow for multiple answers).

Table 6. Correlations between SRA correct reasoning and misconceptions scales being significant at p = 0.01; values in bold exceed 0.30 in absolute value

	Prob	Aver	Comp	Indep	Sampl	Correl	Twow	LrgS
Prob	1.00							
Aver		1.00						
Comp	0.09		1.00					
Indep		0.08	-0.16	1.00				
Sampl		0.10	0.08	-0.07	1.00			
Correl	0.09	0.17				1.00		
Twow	0.13	0.13				0.09	1.00	
LrgS		0.10	0.09		0.07	0.09	0.09	1.00
AverMc		**-0.43**			-0.26			
OutcO	**-0.42**		-0.22	-0.13				**-0.32**
High%						0.08		0.11
Small		-0.10	-0.09	0.07	**-0.69**			-0.16
Repre			-0.21	**-0.69**	0.08			
Cause		-0.08			-0.07	**-0.46**		
EquiPr			**-0.80**	0.20	-0.12	0.09		
Groups								

	AverMc	OutcO	High%	Small	Repre	Cause	EquiPr	Groups
AverMc	1.00							
OutcO		1.00						
High%			1.00					
Small				1.00				
Repre					1.00			
Cause	0.14		-0.07		0.10	1.00		
EquiPr				0.12	-0.10		1.00	
Groups	0.07		0.09					1.00

Taking this pattern of correlations into account, we suggest a different method of aggregating scales scores instead of calculating total correct and misconception scores. On the basis of the strong negative correlations between seven pairs of one correct reasoning scale and one misconception scale, a pair-wise aggregation process seems to be more appropriate than aggregation over all correct, and all incorrect answers. To

investigate this option, an exploratory factor analysis was performed. This factor analysis resulted in a seven-factor solution, with five factors composed of pairs of one correct conception and one misconception, having factor loadings of opposite signs: Comp and EquiPr, Sampl and Small, Indep and Repre, Prob and OutcO, and Correl and Cause. The remaining two factors are composed of Aver, Twow, LrgS, and AverMc; and High% and Groups, respectively. All factor loadings have the expected signs: positive for correct conceptions, negative for misconceptions.

Subsequently, a measurement model was estimated taking the outcome of the explorative factor analysis as its basis. No cross-loadings were allowed but, similar to the estimation of the attitudes measurement model, trait correlations were allowed. In addition, uniqueness correlations were allowed for those reasoning abilities and misconceptions that shared an item. Of the 21 trait correlations, only four appear to be significant. This does not come as a surprise, given the many insignificant correlations in Table 6. All 10 uniqueness correlations appear to be significant. The final measurement model for reasoning abilities is depicted in Figure 3; the fit indices of the final model are reported in Table 7. The fit of the CTCU 7 CFA model is good.

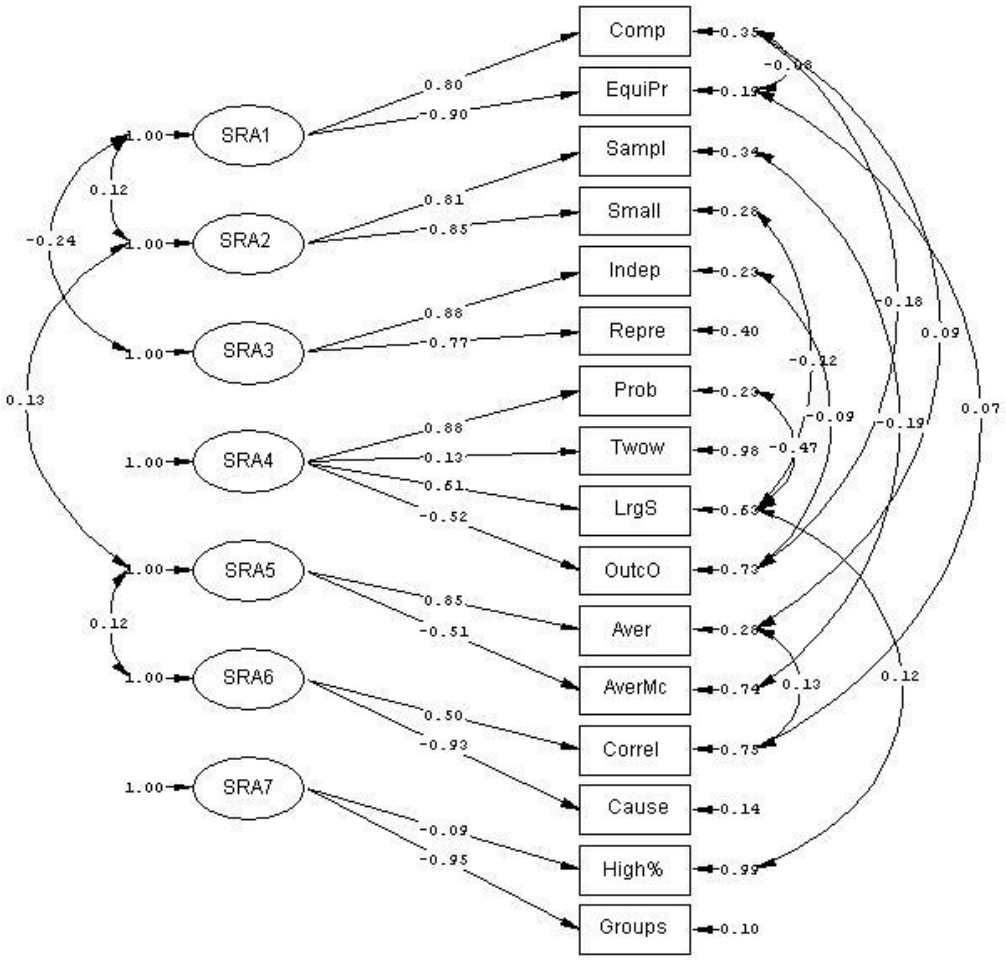

Figure 3. Correlated traits, correlated uniqueness factor model as measurement model for statistical reasoning abilities. Values are standardized parameter estimates. All values shown are statistically significant, $p < 0.05$. CC, SRA1..7, latent reasoning factors.

Table 7. Fit indices of seven-factor correlated traits confirmatory factor models of statistical reasoning abilities

	χ^2	*df*	RMSEA	GFI	NNFI	CFI	RFI
CTCU 7CFA model	355.00	98	0.042	0.97	0.93	0.94	0. 90

Judging from the good fit of this measurement model, an important conclusion with regard to the SRA instrument becomes apparent. When using SRA as an instrument to assess statistical reasoning, it is less attractive to aggregate all correct scales and all misconception scales into constructs like total correct reasoning and total misconceptions, given the limited reliability of such constructs. As an alternative, composing latent reasoning constructs on which both correct and misconception scales load seems to offer higher reliability.

3.5. FULL STRUCTURAL EQUATION MODEL OF ATTITUDE AND BELIEFS, STATISTICAL REASONING ABILITIES, AND COURSE PERFORMANCE

The final step in the analysis regards the integration of both measurement models. This includes the not explicitly elaborated model for course performances, specifying the two latent course performances EXAM and QUIZ. Both course performance constructs are measured by two indicators: a score for mathematics and a score for statistics. The relationships that link the latent factors in the three measurement parts constitute the structural part of the model. The estimation of the structural parameters is similar to the estimation of trait correlations in the measurement models; no a priori restrictions apply as to what parameters are restricted to zero and which are set free. Two modification directions were applied: model building and model trimming. Both methods converge to the model depicted in Figure 4. Figure 4 does not make explicit the estimated correlations between latent factors; the same correlation structure as visible in Figures 2 and 3 was however used in the estimation of the full model. Table 8 reports fit indices of that model and indicates good fit. Table 9 describes the standardized parameter estimates or β-coefficients of the structural part of the model.

Table 8. Fit indices of full structural model of attitudes toward statistics, statistical reasoning abilities, and course performance

	χ^2	*df*	RMSEA	GFI	NNFI	CFI	RFI
SEM	1599.26	620	0.035	0.94	0.96	0.97	0.94

Table 9. Standardized estimates of the structural part of the full structural model of attitudes toward statistics, statistical reasoning abilities, and course performance

	Affect	Cog Comp	Value	Difficulty	Interest	Effort	SRA4
SRA1						-0.10	
SRA2				0.09		-0.09	
SRA3							
SRA4			0.08			-0.10	
SRA5				0.18			
SRA6				0.10	-0.08		
SRA7			0.07			-0.12	
EXAM		0.39		-0.17			0.56
QUIZ		0.34		-0.17		0.09	

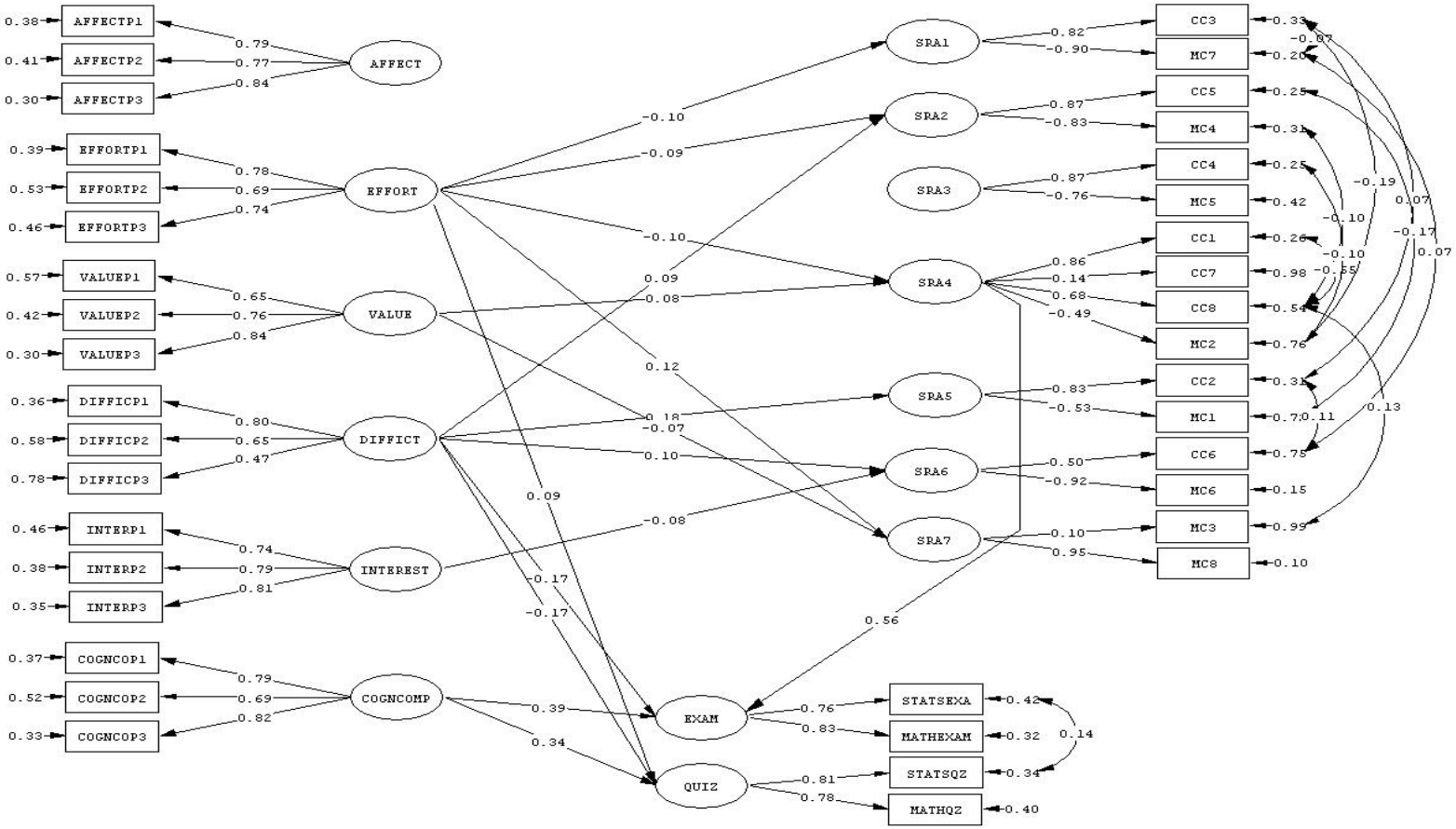

Figure 4. Structural equation model of attitudes toward statistics, statistical reasoning abilities, and course performance. Values are standardized parameter estimates. All values shown are statistically significant, $p < 0.05$. AFFECT, Affect; COGNC, Cognitive Competence; VALUE, Value; DIFFIC, Difficulty; INTEREST, interest; EFFORT, (planned) effort, SRA1..7, latent reasoning factors; EXAM, QUIZ, latent course performance factors

The final structural model allows several interpretations. Students' self-ability belief, Cognitive Competence, is a strong predictor of both latent course performance factors, with β-values of 0.39 and 0.34. This is in agreement with many studies on the expectancy-value model, and self-concept or self-efficacy research. The relationships between statistical reasoning and the two course performance factors are weak, which is in line with the low correlations between SRA constructs and course performance found in several studies. In our study, only SRA4, the latent factor composed of four correct conceptions and misconceptions related to the ability to interpret probabilities, has a significant and strong impact on the latent exam factor.

The second direct effect from attitudinal variables on course performances stems from the other expectancy construct of perceived task demand: (lack of) Difficulty. The relationship is reversed, with β-values of -0.17. This outcome is somewhat surprising; the expectancy-value model would predict a positive relationship. However, the relationship is robust; using a split-sample approach (and path analysis), it is confirmed in subsamples composed in several ways. The bivariate relationship between Difficulty and performance is however absent; the negative relation we find is present only in a simultaneous relation between Cognitive Competence, Difficulty, and course performance. It should thus be interpreted as a process of underestimation of task demand by students with an above average ability belief.

The reduced form squared multiple correlations of both course performance latent factors EXAM and QUIZ are equal to 0.10. This means that the combined effect of both direct paths from SATS variables to EXAM and QUIZ, and the indirect paths from SATS via SRA to the two course performance factors, explains 10% of the total variation in both course performances. In the decomposition of explained variation into direct and indirect effects, it becomes clear that the contribution of the indirect effect can be ignored: less than 0.5%. The dominance of direct over indirect effects is due to the fact that relations between SATS and SRA are weak, and much weaker than relations between SATS and performance. In line with the expectancy-value model, attitudes have a positive impact on reasoning abilities through the variables Value and (perceived lack of) Difficulty. In contrast to predictions based on the expectancy-value model, the Effort variable has a negative impact on four of the seven latent reasoning factors. The negative relationship is consistent: β-coefficients of Effort to the several SRAs are either significantly negative, or zero, but never positive. Although a negative relation may appear counter-intuitive, it is in line with related research on the relationship between preferred learning approaches and reasoning abilities, where it was found that a tendency to surface learning negatively influences statistical reasoning (Tempelaar, 2004; Tempelaar, Gijselaers, & Schim van der Loeff, 2006; Tempelaar, Schim van der Loeff, Gijselaers, Crombrugghe, 2007). Planned effort being a proxy of both achievement motivation and a non-efficient learning approach (see the above discussion of the measurement model of attitudes), will give rise to diverse relationships between learning outcomes and the Effort variable. Learning performances that allow for alternative learning paths – such as memorizing versus understanding – are expected to demonstrate a positive relationship with planned effort. For these learning performances, the achievement motivation component in planned effort is dominant; students who are prepared to work hard will achieve better performances. In our study, quiz scores for both mathematics and statistics are the ultimate example of such type of course performances. Quizzes are designed to be accessible for all students and the bonus points they bring about are especially helpful for students at risk of not passing the course. This makes it plausible that the motivation component in planned effort dominates the learning approach component, which explains the positive relationship between Effort and Quiz.

The opposite case is constituted by the SRA factors. Because the SRA is administered as an entry measurement unrelated to course grading, any direct effect of achievement motivation can assumed to be absent. And because statistical reasoning is not part of any secondary education of most students in this study, indirect effects – taking advantage of having been highly motivated in secondary school – will at most be very modest. As a result, the learning approach component in planned effort is expected to be dominant, which quite well explains the negative relationships found between EFFORT and four of the SRA factors. In this spectrum of course performances, the scores on the exam take an intermediate position. Being the course performance measurement, they certainly contain a strong achievement motivation component. At the same time, exams are certainly much less accessible than quizzes, which feeds the learning approach component. In the aggregation, the two effects are counterbalancing, which quite well might explain the latent factor EXAM being unrelated to Effort.

4. CONCLUSIONS

In this study the affect-extended version of the expectancy-value model (Schau et al., 1995; Dauphinee et al., 1997; Hilton et al., 2004) was adopted as an achievement motivation model. Our data corroborate this extension, in the sense that affect and value turn out to be clearly distinguishable constructs, as well as in the sense that these variables play a distinctive role in the relationships with reasoning abilities and course performance. To our knowledge this study is the first to apply the 36-item SATS version, with the new scales Interest and Effort. Both scales appear to be a valuable addition to the instrument. The latent trait correlations in Table 5 demonstrate that the two factors are well identified constructs. However, correlational analysis suggests that Effort might be composed of two rather different characteristics. Therefore, a decomposition of this scale into an achievement motivation aspect and a learning approach aspect is called for. The latter aspect has the interpretation that students with a surface learning approach will typically achieve high scores on this Effort variable, because they invest large amounts of time for learning subjects by memorization.

Through a factor-analytic study, we conclude that a factor model with most factors being composed of pairs of one reasoning ability and one misconception provides an appropriate measurement model. This shows that the SRA-instrument used by Garfield (1998b, 2003), Garfield and Chance (2000) and Liu (1998) is not flawed. In studies by these authors only two aggregate scales, one for statistical reasoning abilities and one for statistical misconceptions, are employed. They point out that these aggregate scales have shortcomings in view of the low values of correlations between the scales that constitute both aggregate scales, which results in low reliabilities. Our results imply that the finding of low correlations does not invalidate the instrument, but that alternative measurement models other than the one based on aggregate scales should be used.

This study adds support to previous findings of the absence of a strong relationship of misconceptions and their counterpart, the reasoning abilities, with students' course performances. This is demonstrated in studies where statistical reasoning is regarded as one of the several learning outcomes of the course and assessed simultaneously with these other course performances (Garfield, 1998b, 2003; Garfield and Chance, 2000; and Liu, 1998). In the present study along with those of Tempelaar (2004) and Tempelaar et al. (2006), it is also demonstrated in a second type of studies, where statistical reasoning is regarded as part of the prior knowledge state of the student and assessed before the start of the course. Are these studies, given their conclusions that SRA components are only weakly or even un-related to different course performance indicators, uninformative? We

would argue that the opposite is true; exactly because of these absent relationships, they are informative. In general, different components of statistical knowledge, measured as course performance scores, tend to be substantially correlated. For example, in this study the correlation between latent course performance factors EXAM and QUIZ equals r = 0.69. And investigating the relationships among three rather different types of course performances, final exam scores, quiz scores, and homework scores, we find similar substantial correlations. Because the SRA-instrument was developed to assess statistical reasoning mastery achieved in high school statistics programs, the natural hypothesis is that SRA-scores correlate with the several course performances in the same way as the other components of course performances do. But they clearly do not do so. It is these unexpected low correlations that make studies such as ours informative, rather than the case that the expected, substantial positive correlations would have been found.

The absence of substantial relationships can be well explained in the context of naïve theories that are an element of the new theory of learning, as elaborated in Bruer's (1993) 'Schools for thought.' Naïve theories or misconceptions are informal, self-acquired elements of science knowledge, inconsistent with formal science. Students can possess formal knowledge and naïve knowledge at the same time; the learning of formal knowledge does not automatically imply that naïve knowledge is unlearned. In spite of having mastered the formal knowledge, students tend to solve scientific problems with their naïve knowledge, especially when they are confronted with these problems outside a school context. And, worst of all, formal knowledge tends to be forgotten much faster than naïve knowledge. Empirical outcomes of studies using the SRA-instrument are in line with these observations. Absence of substantive relationships is compatible with the hypothesis that both statistical reasoning abilities and statistical misconceptions are part of students' naïve statistical knowledge; the first category naïve and correct, the second category naïve but incorrect. More research to investigate the role of naïve theories in learning and the development of naïve knowledge over time is necessary. This is particularly relevant because the reform movement in statistics education has called for a more prominent position of statistical reasoning, and the related domains of statistical literacy and thinking in the statistics curriculum. So it is the reformed curriculum, more than any traditional curriculum, that requires resolving the instructional challenge of unlearning statistical misconceptions before being able to replace them with proper reasoning abilities.

Empirical studies as documented in special issues of *SERJ* (Ben-Zvi & Garfield, 2004b; Garfield & Ben-Zvi, 2005) and in Ben-Zvi & Garfield (2005) conclude that in order to learn reasoning and to unlearn misconceptions, the use of specific educational tools is indispensable. This study suggests that the use of these tools is probably only part of the solution of the instructional challenge. A strong dependency on these instructional tools might be at odds with educational principles on which student-centered programs are based, in the sense that they limit students' own responsibility to organize the learning process. The outcomes of this study might bring forward some further limitations. In most learning processes students enter the learning context with a given set of background characteristics, such as a preference for deep learning versus surface learning. Most of these contexts allow all students to achieve satisfactory learning outcomes, be it along different learning paths. As a concrete example, our structural equation model suggests that both surface learning oriented students and deep learning oriented students can achieve adequate course performance scores. But our empirical analyses also suggest that statistical reasoning might be the odd man out in this context; the learning of statistical reasoning seems not easily to assimilate to the variation in students' background characteristics as preferred learning approach, as is the case with

other cognitive goals. If this conclusion is correct, it implies we need an even broader range of educational tools than already described in the sources referred to earlier; more than content, the tools should address general learning approaches.

ACKNOWLEDGEMENTS

The authors express their gratitude to the anonymous referees and the assistant editor, Beth Chance, for their helpful remarks and suggestions.

REFERENCES

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004a). *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, The Netherlands: Kluwer Academic Publishing.

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004b). Research on reasoning about variability [Special issue]. *Statistics Education Research Journal*, *3*(2).

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

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bruer, J. T. (1993). *Schools for thought: A science of learning in the classroom*. Cambridge, MA: The MIT Press.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics scale: A construct validity study. *Educational and Psychological Measurement, 65*(3), 509-524.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3).
[Online: http://www.amstat.org/publications/jse/v10n3/chance.html]

Chance, B. L. & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal, 1*(2), 38-41.
[Online: http://www.stat.auckland.ac.nz/~iase/serj/SERJ1(2).pdf]

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.

delMas, R. (2002). Statistical literacy, reasoning, and learning. *Journal of Statistics Education, 10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/delmas_intro.html and http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html]

delMas, R. (2004a).A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishing.

delMas, R. (2004b). Overview of ARTIST website and Assessment Builder. *Proceedings of the ARTIST Roundtable Conference*, Lawrence University.
[Online: http://www.rossmanchance.com/artist/Proctoc.html]

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105-121). New York: The Guilford Press.

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

Gal, I. (2004). Statistical literacy, meanings, components, responsibilities. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 47-78). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Gal, I., & Garfield, J. B. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield, *The assessment challenge in statistical education* (pp. 1-13). Voorburg, The Netherlands: IOS Press.

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

Garfield, J. B. (1996). Assessing student learning in the context of evaluating a chance course. *Communications in Statistics; Part A: Theory and Methods*, *25*, 2863-2873.

Garfield, J. B. (1998a, April). *Challenges in assessing statistical reasoning*. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA.

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

Garfield, J. B., & Ahlgren, A. (1988). Difficulties in learning basic concepts in statistics: Implications for research. *Journal for Research in Mathematics Education*, *19*, 44-63.

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

Garfield, J. B., & Ben-Zvi, D. (2004b). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 397-409). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Garfield, J. B., & Ben-Zvi, D. (Eds.) (2005). Reasoning about variation [Special section]. *Statistics Education Research Journal*, *4*(1).

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematics Thinking and Learning*, *2*(1&2), 99-125.

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

Harris, M. B., & Schau, C. (1999). Successful strategies for teaching statistics. In S.N. Davis, M. Crawford, & J. Sebrechts (Eds.), *Coming into her own: Educational success in girls and women* (pp. 193-210). San Francisco: Jossey-Bass.

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.

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.

Jolliffe, F. (1998). What is research in statistical education? In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 801-806). Voorburg, The Netherlands: International Statistical Institute.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, MA: Cambridge University Press.

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

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*, 59-98.

Liu, H. J. (1998). *A cross-cultural study of sex-differences in statistical reasoning for college students in Taiwan and the United States*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181–220.

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, a project of the National Council of Teachers of Mathematics* (pp. 575-596). New York: Macmillan.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishing.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/rumsey2.html]

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

Schau, C., Dauphinee, T. L., Del Vecchio, A., & Stevens, J. (1999). *Survey of attitudes toward statistics (SATS)*.
[Online: http://www.unm.edu/~cschau/downloadsats.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*(5), 868-875.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.

Short, T. H. (Ed.) (2002). Statistical literacy, reasoning, and thinking [Special section]. *Journal of Statistics Education*, *10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/abstracts.html]

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

Sundre, D. L. (2003, April), *Assessment of Quantitative reasoning to enhance educational quality*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
[Online: http://www.gen.umn.edu/artist/articles/AERA_2003_QRQ.pdf]

Tempelaar, D. (2004). Statistical reasoning assessment: An Analysis of the SRA instrument. *Proceedings of the ARTIST Roundtable Conference*, Lawrence University.
[Online: http://www.rossmanchance.com/artist/Proctoc.html]

Tempelaar, D. T., Gijselaers, W. H., & Schim van der Loeff, S. (2006). Puzzles in statistical reasoning. *Journal of Statistics Education*, *14*(1). [Online: http://www.amstat.org/publications/jse/v14n1/tempelaar.html]

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

Tempelaar, D., Schim van der Loeff, S., Gijselaers, W., & De Crombrugghe, D. (2007). *Preferred learning approaches and statistical reasoning.* Unpublished manuscript.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 92-122). San Diego: Academic Press.

DIRK T. TEMPELAAR
Department of Quantitative Economics,
Faculty of Economics and Business Administration
University of Maastricht
PO Box 616, 6200 MD Maastricht
the Netherlands