Paper presented at the ARTIST Roundtable Conference on Assessment in Statistics
held at Lawrence University, August 1-4, 2004

Statistical Reasoning Assessment: an Analysis of the SRA Instrument
Dirk Tempelaar, UM
Version 3.2, October 22, 2004

Abstract
The Statistical Reasoning Assessment or SRA is one of the first instruments developed to assess students' statistical reasoning. Published in 1998 (Garfield, 1998a), it became widely available after the Garfield (2003) publication. Empirical studies applying SRA by Garfield and co-authors brought forward two intriguing puzzles: the 'gender puzzle', and the puzzle of 'non-existing relations with course performances'. This present study aims to address those two puzzles, and in doing so to contribute to the validation of SRA, by applying the instrument to a relative large group of students participating in an introductory statistics class. Different from earlier empirical studies, we administered the SRA at the start of our course, what enables us to study the role of preconceptions, both being correct and incorrect in nature, in learning statistics. Findings in this study suggest that both puzzles may be understood in terms of differences in effort students invest in studying: students with strong effort-based learning approaches tend to have lower correct reasoning scores, and higher misconception scores, than students with different learning approaches. Implications of these findings for statistics education are discussed.

1. Introduction

Statistical reasoning, and the related concepts of statistical thinking and statistical literacy, are at the centre of interest of the educational statistics community. For example, the Winter 2002 edition of the Journal of Statistics Education provides a series of articles based on an AERA 2002 symposium: delMas (2002a), Garfield (2002), Chance (2002), Rumsey (2002) and delMas (2002b). The articles explore definitions, distinctions and similarities of statistical reasoning, thinking, and literacy, and discuss how these topics should be addressed in terms of learning outcomes for educational statistics courses. The relationship between statistical reasoning (and related concepts) and the learning of statistics is a complex one. To cite Garfield, the author of the contribution on statistical reasoning: '*Although the term "statistical reasoning" is often used in different ways, it appears to be universally accepted as a goal for students in statistics classes. It has been shown that statistical reasoning used in everyday life as well as in classes is often incorrect, due to the different intuitions and rules that people use when evaluating statistical information.*' (Garfield, 2002, p. 9). So first of all, statistical reasoning is an achievement aimed for in most introductory statistics courses, comparable to traditional achievements as e.g. the understanding of the concept sampling distributions. But in addition to being an important output of statistics education, statistical reasoning is also a crucial input in the process of learning statistics. Students enter our classes with prior reasoning skills; to the extent these prior skills correspond to true knowledge being part of the aimed course achievements, this will ease the learning process. However, an important category of prior knowledge if formed by misconceptions, or intuitive but faulty reasoning mechanisms. Both types of preconceptions are, according to modern learning theories (Brandsford et al, 2000) crucial determinants in learning; if preconceptions are not properly addressed, newly learned correct knowledge might appear much more volatile than existing preconceptions brought into class.

For both aspects of the role of statistical reasoning in learning statistics, the measurement of statistics reasoning, either at the end or at the start of the course, is an important issue. In the closing summary of the JSE Winter 2002 series, delMas (2002b) emphasizes the role of assessment. It seems to be a critical issue: although a lot of progress has been achieved in the delineation of the related concepts reasoning, thinking and literacy,

and the elaboration of instructional implications of research findings in each of areas, we are still rather empty-handed with regard to instruments to assess students' abilities. Yes, in small-scale experimental settings, a range of techniques based on interviewing students, or think-aloud problem solving has been documented (see e.g. the contributions to the yearly SRTL forums on Statistical Reasoning, Thinking and Literacy, of which the second edition was reported in the opening issue of Statistics Education Research Journal (SERJ, 2002)). But instruments that could be applied on a broad scale in classes as large as the one this study reports on, thus necessarily of closed or objective format, are scarce, not to say unique. Garfield (2003, 1998a) describes one such instrument, the Statistical Reasoning Assessment or SRA, and this contribution provides an extensive analysis of this instrument. The SRA is a multiple-choice test consisting of 20 items developed by Konold and Garfield as part of a project to evaluate the effectiveness of a new statistics curriculum in US high schools (Konold, 1989; Garfield, 1996, 1998a, 2003). Each item in the SRA describes a statistics or probability problem. Responses to items include a statement of reasoning, explaining the rationale for the particular choice. Some of these responses are instances of correct reasoning, but the majority demonstrate characteristic patterns of intuitive, incorrect reasoning. Garfield and co-authors have performed several empirical analyses on the SRA (Garfield: 1998b, 2003; Garfield & Chance, 2000; Liu, 1998). In all of these analyses, the SRA was administered at the end of course, parallel to the final exam, with the main aim to investigate the mastery of reasoning skills and the relationship of this mastery and course performances. One of the striking outcomes of these research contributions is what we called above the puzzle of 'non-existing relations with course performances': correlations between aggregated reasoning skills (both total of correct reasoning, as total of incorrect reasoning), demonstrate low or zero correlations with course performances.

The present study aims to explore the role of statistical reasoning in learning statistics. Different from other studies, we administered the SRA in the very beginning of the first introductory course, providing its outcomes the nature of students' preconception levels achieved outside class or, in some cases, in high school programs. This difference in timing in the administration of SRA precludes the possibility to investigate how well the course succeeds in getting across statistical reasoning; however, it does allow investigating the role of prior conceptions and misconceptions in learning statistics. Our analysis of the SRA benefited from an rich dataset built by two shifts of students in an introductory Quantitative Methods class. The richness partially stems from the number and diversity of those students: between 800 and 1000 each year, from the Netherlands, from Germany, and from other, mostly European, countries. But beyond the large number: data collected was very diverse, ranging from answers to several surveys, some specific for the domain of statistics, like the SRA or the SATS, the Survey of Attitudes towards Statistics, some more general, e.g. data on prior education and prior knowledge. These data were collected in the context of student projects, in which students study their learning habits in comparison to their colleagues' ones using personal and aggregated data. This data set allowed us to make a large scale study of the SRA instrument, analysing its characteristics as such and its relationship to other relevant student characteristics. In reporting on this analysis, we will restrain as much as possible from a theoretical discussion of the instrument, or the theoretical background of statistical reasoning, but rather restrict to the more empirical aspects of the analysis. For the theoretical context, we refer to the JSE 2002(3) series of articles, and to Garfield (1996, 1998a, 1998b, 2003), Garfield and Chance (2000), and Lovett (2001).

Beyond contributing to the knowledge of statistical reasoning and the assessment of it through the instrument SRA, a second aim of this paper is to contribute to the new tradition of classroom-based or action research: Garfield and Chance (2000), Jolliffe (1998). We realize that, mainly due to the scale of the project, we made a restricted use of the broad range of

instruments available for this type of research, but in doing so we were able to maintain the most natural setting possible and combine this with a very active role for the students.

Section 2 of this study provides a short background of research into statistical reasoning and its assessment, while section 3 sketches the settings of our research project and the characteristics of the students participating in it. Sections 4 and 5 analyse the SRA data: in section 4 on a descriptive level, while in section 5 a factor model for the SRA data is developed. Since earlier empirical studies by Garfield and co-authors found strong gender and country effects in SRA data, both of these variables are incorporated in our SRA analysis in these sections. We find similar country and gender effects in our data, thereby reinforcing the existence of such puzzles. Thanks to richness of the dataset, we can pursue our analysis further by bringing in additional factors that might explain SRA levels and exhibit at the same time differences between countries and sexes: potential cofounders. In section 6, prior education is introduced, the most plausible factor to explain inter-student differences in SRA levels, given the huge differences in secondary schooling systems in Europe. Prior education appears to have only very modest impact on SRA levels, and in this way can explain country differences, but certainly not gender differences. As a further step in the search for factors responsible for the gender effect, in section 7 the affective personality factor attitudes towards statistics is introduced. Once again, although statistics attitudes of females and males appear to be quite different, attitudes themselves demonstrate no relationship to reasoning levels, excluding these affective factors as background factor with the potential to explain the gender effect in SRA. Stimulated by the last puzzle in Garfield's work, the puzzle of non-existing relations with course outcomes, and knowing that our instruments to assess course outcomes are very diverse in nature, we introduce course outcome in section 8. Of all factors investigated in our SRA study, this one demonstrates to be most promising, in several aspects. When course performance is incorporated on disaggregated levels and not an aggregated level, different performance measures are significantly (but weakly) related to reasoning levels. This relationship takes rather different forms: one performance measure is positively related to reasoning levels, but the other negatively. The major difference between the performance measures used in this study is the extent to which they are based on effort on the one side, and cognitive abilities on the other side. So in concluding that students who adopt an effortful learning approach are at risk in acquiring statistical reasoning skills, our two puzzles might have been solved, but another arises: how to accommodate our instructional processes such that effortful learners still have good chances. Educational implications and conclusions are discussed in sections 9 and 10, respectively.

2. Background

2.1. Statistical reasoning and its assessment

Garfield and Chance (2000) define statistical reasoning as the way students reason with statistical ideas and make sense of statistical information. This involves making interpretations based on sets of data, representations of data, and statistical summaries of data. Statistical reasoning is based upon an understanding of important concepts such as distribution, location and variation, association, randomness, and sampling, and aims at making inferences and interpreting statistical results. Recent research efforts are directed at isolating statistical reasoning from more general forms of reasoning, like mathematical reasoning, and at distinguishing statistical reasoning, statistical thinking and statistical literacy: see e.g. delMas (2002) and Reading (2002).

Statistical reasoning has a complex relationship to statistical education. First of all, as expressed by Garfield (2002): "*it* [statistical reasoning] *appears to be universally accepted as*

a goal for students in statistics classes". So ideally, when assessed at the end of a statistics course, students should demonstrate mastery of reasoning skills as one of the several aspired aims of the course. But then, if a course is built upon constructivist's learning principles as most courses benefiting from the education reform do, mastery of students' reasoning skills when entering the course is a relevant piece of prior knowledge students bring into class. On top of that: statistical reasoning brought into class as part of 'prior knowledge' is quite often unlearned, intuitive knowledge. Such knowledge might correspond to correct reasoning skills that are part of the agenda of our course; as such, it is 'true prior knowledge'. However, quite often this unlearned knowledge is faulty in nature, and belongs to the body of statistical misconceptions. These prior misconceptions should be addressed on and hopefully completely removed, and replaced by corresponding correct conceptions in our statistics class. Research in learning in general (see e.g. Brandsford et al, 2000), and in statistical reasoning in specific (Garfield & Ahlgren, 1988, Shaughnessy, 1992), makes however clear that unlearned, intuitive misconceptions are of stubborn nature. It suggests that even students who can correctly compute probabilities tend to fall back to faulty reasoning misconceptions when asked to make an inference or judgment about an uncertain event outside the context of doing a statistics exam, thereby relying on incorrect intuitions already present when entering the course. So teaching correct conceptions, no matter how successful, is no guarantee students will not apply misconceptions anymore. This 'last learned, first forgotten' principle justifies special attention to misconceptions brought to class.

Assessment instruments for statistical reasoning are based on theoretical studies in statistical reasoning that stem from what Lovett (2001) calls the first phase of research in statistical reasoning: the theoretical focus of the 1970s. This theoretical research primarily directed at the explanation of fallacies in student's statistical reasoning, led to the discovery of e.g. the 'Law of small numbers' and the 'Representativeness misconception', both described in Kahneman, Slovic, and Tversky (1982), the 'Outcome orientation' described in Konold (1989), and the 'Equiprobability bias' described in Lecoutre (1992). These are well-documented examples of aspects of statistical reasoning (all of the type: faulty reasoning), see Garfield and Ahlgren (1988) and Shaughnessy (1992) for extensive surveys. Based on classifications of types of faulty statistical reasoning (and their correct counterparts), assessments instruments have been designed, the most well-known being the Statistical Reasoning Assessment (SRA).

2.2. Statistical Reasoning Assessment: the SRA instrument

The Statistical Reasoning Assessment, shortly SRA, is a multiple-choice test consisting of 20 items developed by Konold and Garfield as part of a project to evaluate the effectiveness of a new statistics curriculum in US high schools (Konold, 1989; Garfield, 1996, 1998a, 2003). It contrast to most other assessment instruments, it consists of closed format items and it is therefore one of the few objective instruments for assessing the statistical reasoning abilities of students at pre-university level (see e.g. Gal and Garfield (1997) for a survey of assessment tools). Each item in the SRA describes a statistics or probability problem. Most responses include a statement of reasoning, explaining the rationale for the particular choice. For a full description of the individual items and the eight correct reasoning scales and eight misconceptions scales, we refer to Garfield (1998a, 2003); Table 1 summarizes the several scales of the instrument.

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

Correct Reasoning Scales:
CC1: *Correctly interprets probabilities*. Assesses the understanding and use of ideas of randomness, chance to make judgments about uncertain events.

CC2: *Understands how to select an appropriate average*. Assesses the understanding what measures of center tell about a data set, and which are best to use under different conditions.

CC3: *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.

CC4: *Understands independence*.

CC5: *Understands sampling variability*.

CC6: *Distinguishes between correlation and causation*. Assesses the knowledge that a strong correlation between two variables does not mean that one causes the other.

CC7: *Correctly interprets two-way tables*. Assesses the knowledge how to judge and interpret a relationship between two variables, knowing how to examine and interpret a two way table.

CC8: *Understands the importance of large samples*. Assesses the knowledge 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:

MC1: *Misconceptions involving averages*. This category includes the following pitfalls: averages are the most common number; failing to take outliers into consideration when computing the mean; comparing groups on their averages only; and confusing mean with median.

MC2: *Outcome orientation*. Students use an intuitive model of probability that lead them to make yes or no decisions about single events rather than looking at the series of events; see Konold (1989).

MC3: *Good samples have to represent a high percentage of the population*. Size of the sample and how it is chosen is not important, but it must represent a large part of the population to be a good sample.

MC4: *Law of small numbers*. Small samples best resemble the populations from which they are sampled, so are to be preferred over larger samples.

MC5: *Representativeness misconception*. In this misconception the likelihood of a sample is estimated on the basis how closely it resembles the population. Documented in Kahneman, Slovic, & Tversky (1982).

MC6: *Correlation implies causation*.

MC7: *Equiprobability bias*. Events of unequal chance tend to be viewed as equally likely; see Lecoutre (1992).

MC8: *Groups can only be compared if they have the same size*.

2.3. Empirical studies of the SRA instrument

Studies reporting empirical data on the administration of SRA are limited, and partly overlap in experiments they describe: Garfield (1998b, 2003), Garfield & Chance (2000), Liu (1998) and Sundre (2003).

In an attempt to determine the criterion-validity of the SRA, Garfield administered the instrument to students at the end of an introductory statistics course and correlated their total correct and total incorrect scores with different course outcomes: final score, project score, quiz total (Garfield, 1998b; Garfield & Chance, 2000). The resulting correlations were ‘extremely low’, suggesting that statistical reasoning and misconceptions were rather unrelated to students' performance in that first statistics course.

Garfield (1998b), Garfield & Chance (2000) and Liu (1998) report that the intercorrelations between items are quite low, implying a low reliability from an internal consistency point of view. In spite of these low intercorrelations, and the fact that items do not appear to measure a single trait, all of these studies analyse the total correct reasoning score and the total misconceptions score, so aggregated scores. The test-retest reliability for these two total scores turns out to be 0.7, and 0.75, respectively. Liu (1988) performs a cross-cultural comparison of USA and Taiwanese students to identify possible gender differences, at the level of separate scales as well as at the level of total correct and total misconception scores. She finds a significant country effect and a significant gender effect in Taiwanese students but not in USA students. The gender effect indicates that male students score higher on correct reasoning and lower on misconception compared to female students; see Garfield (1998b), Garfield & Chance (2000) and Liu (1998).

The Sundre study (Sundre, 2003) is somewhat different in nature: it takes the SRA as a starting point, but derives a new instrument, called the Quantitative Reasoning Quotient (QRQ), essentially by splitting single SRA items into several QRQ items. Presenting different rationales as separate items, in stead of offering them within one item in the format of a check list, it is hoped to get a better impression of correct reasoning going hand in hand with specific misconceptions in students. As a consequence, scores on the QRQ are not easily comparable to scores on the SRA instrument.

The aim of this study is to investigate the characteristics of the SRA instrument itself and to contribute to its validation. Given that aim, no attempt is made to modify the instrument, and the reported statistics are similar to those described in the Garfield and co-author studies.

3. Setting and Subjects of this Study

The course Quantitative Methods (QM) is part of both the first-year Economics and Business programs in the Faculty of Economics and Business Administration of the University of Maastricht (UM). The course covers subjects from mathematics, statistics and computer skills. The material is regarded as being difficult and unattractive by most students. Traditionally, the results have been less than expected both in terms of the rate of passing the course and in terms of retention of the material. Over the years the mode of instruction of the course has evolved from predominantly class-room teaching to a setting where students meet in small groups of approximately twelve students with a tutor to discuss their solutions to any – usually homework – problems supplemented by mass lectures. The latter feature had to be retained because of budgetary constraints, in spite of the fact that problem based learning has always been a hallmark of the teaching at the University of Maastricht.

Data were collected on two shifts of students: approximately 1000 students participating in the QM course in the academic year 99/00, and approximately 800 students participating in 03/04. About 10% percent of the students are ‘repeat’ students that did not manage to pass that specific course in previous years. Another relevant decomposition of our freshmen is according to nationality. Since all studies in the faculty are taught in the English language, the faculty attracts a relatively large proportion of foreign students. In ’99, the share of foreign students was 46%, a figure that has risen to 57% in ’03. Of all foreign students, roughly two third has German nationality, the remainder being mostly from other European countries. Only the last couple of years, a growing but still rather small inflow of Asian students is visible.

Distinguishing students according nationality is important since major differences exist between secondary school systems in Europe. All entering Dutch students participated in a final, national exam in at least seven subjects, including either basic mathematics (calculus

oriented), or advanced mathematics (algebra and geometry oriented), or both. In contrast, German students have four subjects in their final exam, two at an advanced level, two at a basic level. Given that they choose mathematics, either at the basic level ('Grundkurs') or at the advanced level ('Leistungskurs'), the level of their mathematical knowledge is somewhat comparable to that of Dutch students that choose the corresponding level of mathematics. However, a sizeable proportion of German students, mainly of the '99 shift, did not select mathematics at any level for their final exam, and their level of mathematical schooling is really incomparable to that of Dutch students. Besides that, the share of statistics and probability theory in mathematical courses will differ from state to state in Germany. In Dutch secondary education, an important structural break took place that provided us with a kind of pseudo experimental condition. Basically, the '99 inflow of Dutch students consists of two rather different groups: students with a profound interest in math and science, but without any schooling in statistics or probability (since they took the advanced mathematics program in high school), and students with much less (or absent) interest in math and science who did however receive proper schooling in statistics and probability. In the redesign of the curriculum, this strange situation abolished; statistical topics were added to the advanced mathematics program, so in the '03 inflow, all Dutch students have identical prior knowledge in statistics.

Prior education data constitute an important part of data used in this study. The most important part of the data set consists of information derived from the student projects. The topic of these projects is 'a statistical analysis of my study behaviour', in which students compare their study habits with that of companion students. In order to provide students with data allowing them to make such a comparison, all students complete several questionnaires in the first weeks of the course. The results, both individual data and aggregated group data, are made available in the later weeks of the course. The SRA survey was one of the self-report instruments that students had to fill out in the first weeks of course. Another questionnaire that was administered is the Survey on Attitudes Towards Statistics (SATS).

The several questionnaires were administered in the tutorial sessions ('99) or through web based forms ('03). The response is quite high: the chance to achieve bonus points for their student project made it attractive for students to participate. It is not possible to express the response rates as single figures: different questionnaires were administered in different sessions (days) with different student being present or absent. Most of the analyses reported here are based on the responses of about 1300 students (720 in shift '99, 580 in shift '03). The majority of the other students officially enrolled in the course would typically participate in the exam, but not in any educational activities.

Three assessment instruments for student's performance form another source of information. The performance measures are the grades for the tests, the bonus points for home work assignments and those for quizzes. These were used as indicators for the outcome of the learning process for both mathematics and statistics separately. The results of the student projects were not used as performance indicators because the project was graded by a pass/fail judgement, and students were allowed to improve their projects using the tutor's feedback on a preliminary version.

So except for the course performance measures, all data was measured in the first week of the course, that is, essentially before the student learning has started. This implies that these data describe entry characteristics of students, unrelated to the learning taking place in our QM course. This annotation is above all essential for properly understanding SRA levels: the instrument measures correct reasoning skills and misconceptions obtained prior to going to university. This distinguishes this study from other empirical studies of SRA data, where the

SRA was administered at the end of the course, assessing so both skills achieved in and outside the statistics class (Garfield: 1998b, 2003, Garfield and Chance: 2000, and Liu: 1998).

4. Descriptives of SRA data

It is interesting to compare the descriptive statistics of the present SRA data with those reported in Garfield (1998b, 2003), Garfield and Chance (2000) and Liu (1998) for samples of USA and Taiwanese students. Table 2 presents the means of the several scales of all students and of the several subsets of the present sample. All means are expressed as a proportion, on a [0-1] scale. In addition the corresponding statistics, pertaining to USA and Taiwanese college students adapted from Garfield (1998b, 2003), are presented. Those scores are re-expressed on a [0-1] scale, to allow for comparison with the outcomes of our study. The SRA was administered to both these groups of students at the end of an introductory course in business statistics. The UM students filled the questionnaire out at the start of the course, making it a prior knowledge assessment.

In addition to scores on eight reasoning skills (CC1 … CC8), eight misconceptions (MC1 … MC8), aggregated reasoning score (CCtot) and aggregated misconceptions (MCtot), two adapted scores are reported. When taking the SRA with two statistics lecturers, one striking deviation with the scoring rubric was found. In the fifteenth item of the SRA, data from two experimental groups are to be compared, one group having somewhat higher scores than the other. Six rationales are given, including 'one group did better because its average appears to be a little higher than the average of the other group' and 'there is no difference between the two groups because the difference between their averages is small compared to the amount of variation in the scores', are given, and students are asked to indicate the one they agree most with. Both the two lecturers, and 85% of all students, choose the first mentioned rationale, according to the scoring rubric a misconception involving averages. None of the lecturers, and only 8% of the students, choose the correct second rationale. We judged that, through the wording of the supposed misconception, one can not be certain that all students (and lecturers) choosing that rationale are really caught in the misconception, and decided to calculate, besides the scores including this item, scores excluding this item: CC5A and MC1A, as adapted scores for CC5 and MC1.

The aggregated scores total correct reasoning (CCtot) and total misconceptions (MCtot) are obtained in the same way as in the studies by Garfield and co-authors: taking the sum over all correct reasoning and misconception scales, and re-expressing as a proportion. Since 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.

In addition to the calculation of SRA scores of the whole sample, several decompositions were applied to allow for comparison of different groups: by gender, and by nationality, following Garfield and co-authors, and for obvious reasons, by academic year. That last decomposition resulted in some minor differences between the two shifts. However, we found much stronger differences for nationality, and since the share of foreign students increases over time, the shift differences could be contributed to underlying changes in nationality composition. For that reason, we focus in Table 2 on gender and nationality effects; differences that were found to be statistically significant at the .01 level are indicated in bold..

Table 2. *SRA Correct reasoning scales and misconceptions scales as proportions.*

Correct Reasoning Scales:	UM (n=1303)	UM Female (n=518)	UM Male (n=778)	UM Dutch (n=632)	UM Foreign (n=666)	USA (n=245)	Taiwan (n=267)
CC1	0.69	0.68	0.69	0.69	0.69	0.68	0.68
CC2	0.71	**0.67**	**0.74**	**0.76**	**0.67**	0.61	0.60
CC3	0.41	**0.38**	**0.43**	0.42	0.39	0.45	0.46
CC4	0.60	0.61	0.59	**0.62**	**0.58**	0.63	0.74
CC5	0.25	**0.20**	**0.29**	**0.29**	**0.22**	0.22	0.23
CC5A	0.41	**0.28**	**0.50**	**0.48**	**0.34**		
CC6	0.70	0.73	0.69	**0.78**	**0.63**	0.52	0.65
CC7	0.77	**0.73**	**0.80**	**0.83**	**0.72**	0.65	0.79
CC8	0.72	0.72	0.72	0.73	0.71	0.68	0.76
CCtot	0.58	**0.56**	**0.59**	**0.61**	**0.55**	0.56	0.60
Misconception scales:	UM (n=1303)	UM Female (n=518)	UM Male (n=778)	UM Dutch (n=632)	UM Foreign (n=666)	USA (n=245)	Taiwan (n=267)
MC1	0.37	**0.40**	**0.34**	**0.32**	**0.41**	0.30	0.22
MC1A	0.16	**0.20**	**0.13**	**0.12**	**0.19**		
MC2	0.23	0.25	0.22	0.24	0.22	0.23	0.22
MC3	0.15	**0.18**	**0.14**	0.15	0.15	0.09	0.09
MC4	0.28	**0.33**	**0.25**	**0.25**	**0.32**	0.29	0.34
MC5	0.15	**0.13**	**0.17**	0.15	0.15	0.17	0.11
MC6	0.23	0.21	0.25	**0.18**	**0.28**	0.10	0.10
MC7	0.57	**0.62**	**0.54**	0.56	0.58	0.56	0.56
MC8	0.26	**0.31**	**0.24**	0.25	0.28	0.60	0.39
MCtot	0.29	**0.31**	**0.27**	**0.27**	**0.30**	0.27	0.24

In addition to testing on differences in means for gender and nation separately, Analysis of Variance was applied to investigate simultaneously gender and nation effects, and their interaction. These outcomes are integrated in the following paragraphs.

Broadly speaking, the patterns in the UM data are similar to those found for the USA and Taiwan students in the sense that, of the correct reasoning scales, the means CC7 and CC8 are highest and those of CC3 and CC5 are lowest for all three samples. Of the misconception scales, MC7 and MC8 are highest for the USA and Taiwan students and, although for UM students MC7 ranks highest as well, MC8 ends up somewhere in the middle region. MC3, MC5 and MC6 are lowest for all three samples. Two other general patters emerge in the UM data. Similar to Garfield (2003), we find a nationality effect in half of all scales, and both aggregate scores. That effect has always the same direction: Dutch students have higher correct reasoning and lower misconception scores than foreign students. Once again similar to Garfield (2003), we find a gender effect in the UM data: in 11 out of 16 of the individual scale, and in both total scales. The gender effect has, except for MC5, a consistent direction: males score higher on correct reasoning and lower on misconceptions than females. Most effects are quite strong in statistical sense, having p-values below .0005; for that reason, p-values are not reported in the discussion of the individual scales.

CC1 (*Correctly interpreting probabilities*) is a paragon of constancy; no effects are found in the ANOVA and the *t*-tests.

For CC2 (*Understanding how to select an appropriate average*) significant differences both between genders and nationalities were found for UM students, but no interaction effect. UM students score higher than USA and Taiwan students; under the assumption that the standard deviations in the last two groups, not reported in Garfield (2003), are similar to those found in the UM students, these differences are significant.

The difference in timing, taking the test as pre-test as is the case in our study, or as a post-test as is the case in the studies reported by Garfield (2003), will have its greatest impact for conceptions that are explicitly on the agenda of any introductory statistics course. CC3 and CC4 seem to be typical examples of such topics, and not surprisingly, UM students score slightly lower than students in the Garfield-report. Between UM subgroups there is a significant gender effect for CC3 (*Good samples have to represent a high percentage of the population*), and a significant nationality effect in CC4 (*Understanding independence*), but no interaction effects.

Very significant gender, nationality and interaction effects are found for CC5 (*Understanding sampling variability*). Surprisingly, all these effects enlarge to really huge proportions when item 15 is excluded (e.g., *F*-statistic for gender equals 32.5).

The scores in UM students for the CC6 (*Distinguishing between correlation and causation*), and CC7 (*Correctly interpreting two-way tables*) scales are higher than those of USA/Taiwan students. These scales, together with CC2, represent concepts that may be characterised as general reasoning skills more than as statistical reasoning skills; Higher 'European' scores may reflect the general level of secondary education in Europe.

Conceptions for which UM-students achieve higher scores than students in the Garfield-report, CC2, CC6, and CC7, may be characterised as general reasoning skills more than as statistical reasoning skills, and will not play that prominent of a role in a statistics course; higher 'European' scores in general, and higher Dutch scores in specific, may reflect the general level of secondary education.

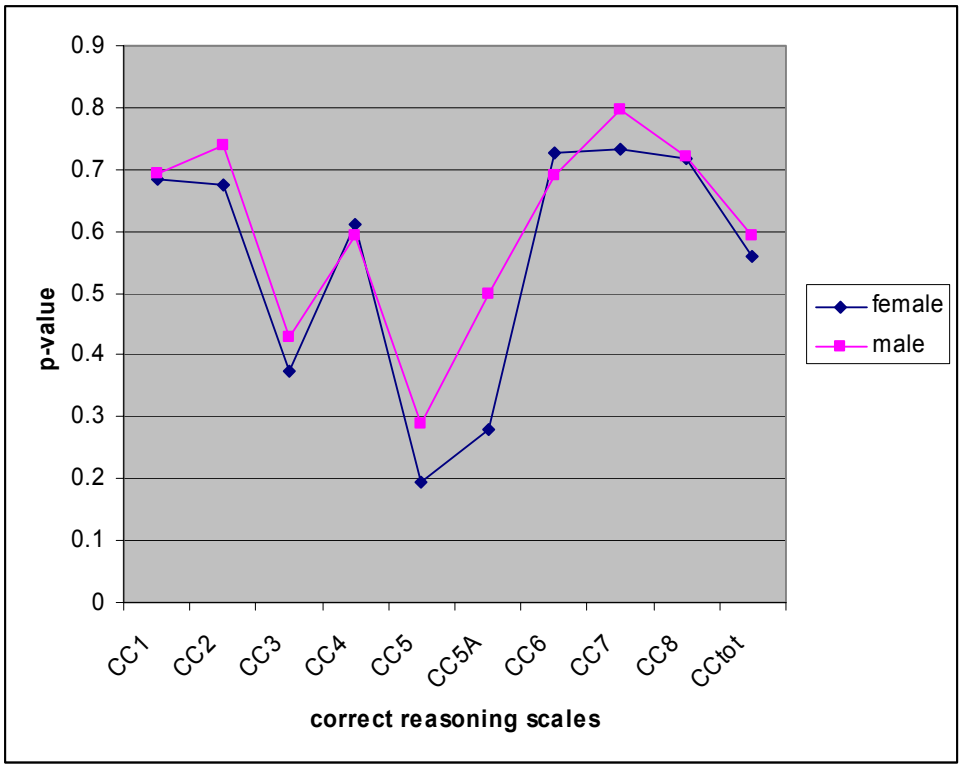

***Figure 1**: Reasoning abilities of male and female students*

Similar conclusions apply to the several misconception scales. High UM scores relative to the Garfield-report are found for MC1 (*Misconceptions involving averages*), MC3 (*Good samples have to represent a high percentage of the population*) and MC6 (*Correlation implies causation*), all referring to topics that will be covered in any introductory course, so that the timing of the test administration once again plays a crucial role. In contrast, MC8 (*Groups can only be compared if they have the same size*) shows remarkably low misconception scores, especially relative to the score of US-students. Significant gender effects were found in six, and significant nationality effects in three MC's; the two MC's having both, MC1, and MC4 (*Law of small numbers*) also demonstrating interaction effects.

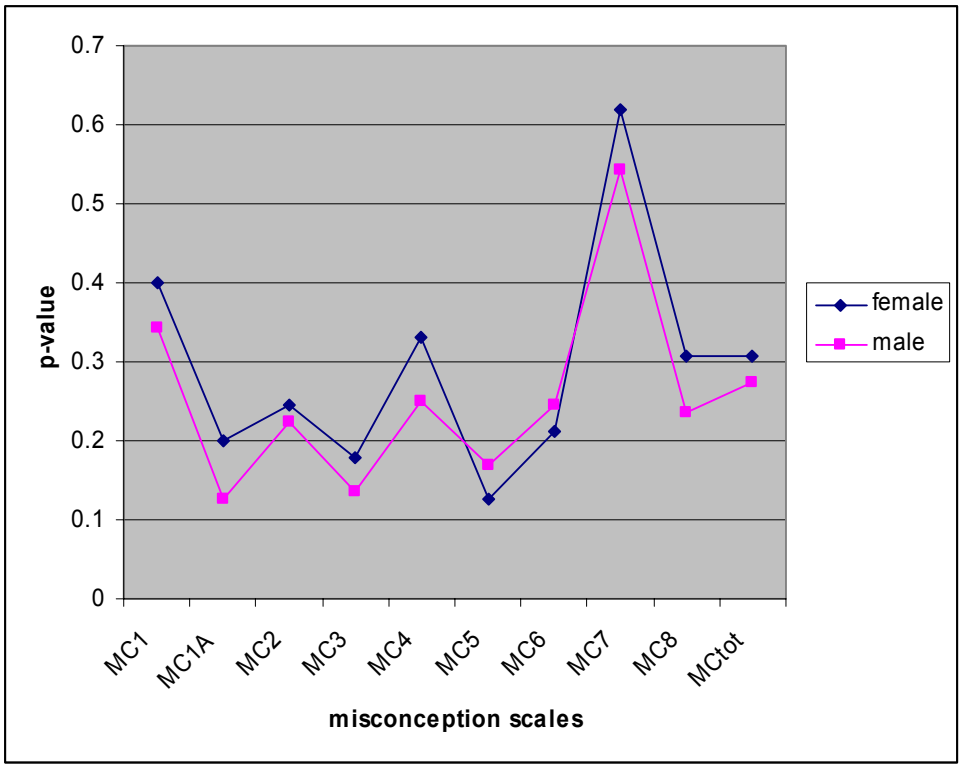

***Figure 2**: Misconceptions of male and female students*

In the Liu-study, reported in Garfield (1998b, 2003), Garfield and Chance (2000), and Liu (1998), the analysis of gender and culture/nationality effects was restricted to the aggregated total correct and total misconceptions scores, instead of the individual scales. Based on an ANOVA of aggregated scores with country and gender as factors, Garfield (2003, p. 30) concludes: "*It is interesting to see that despite the seemingly similar scale scores for the students in the two countries, there are actually striking differences when comparing the male and female groups. … it will be interesting to see if replications of this study in other countries will yield similar results*." 'Similar' should here be understood to mean that males have significantly higher total correct reasoning scores (except for USA), and have significantly lower total misconceptions scores. These results indeed do generalize to our study, with a remarkable regularity. We find significant gender effects in both aggregated scores of the same direction. Moreover, noting that the reliability of the aggregated scores is questionable because of low inter-correlations - see next section for a discussion of that issue - we found that CC2, CC3, CC5, and CC7 are significantly higher and MC1, MC3, MC4, MC7, and MC8 significantly lower for males than for females among the UM students (where MC5 plays the role of the exception which proves the rule).

Much of the remainder of this paper will focus on the question: what can explain these remarkable gender effects? We will start with one paragraph of what we expected to be an good explanation, but appears not to be. Due to its problem-based character of all programs of the University of Maastricht, we have a lot of experience with tests in true/false/? format. In analysing the outcomes of these kind of test, we typically find the following gender effect: female students answer less items than male students. Both the number of correctly answered items, as the number of incorrectly answered items, are less than for males (but the difference,

total correct minus incorrect, is in general higher). The best explanation of that pattern is females to be more risk averse than males. Or, in a somewhat different formulation: female students tend to underestimate their knowledge level, male students to overestimate. This pattern is so general, that we expected it to be present in our SRA data. However, if risk aversion would have been the trigger, than females should have both lower correct score, and lower incorrect score and thus in total check less answers than males. For the SRA however, the number of checked answers for females and males are rather equal (19.4 versus 19.5), but their distribution over correct answers (11.7 versus 12.5) and incorrect answers (7.8 versus 7.1) is unequal.

The nationality effect is about as stable as the gender effect, but les sharp and much less puzzling with regard to its explanation: Dutch secondary education seems to offer Dutch students a better preparation than most other European school systems, which shows up, amongst other things, in better general and statistical reasoning abilities. The focus on mathematics in Dutch secondary education, including an introduction into statistics and probability, apparently provides Dutch students a lead. Could this nationality effect contribute to (part of) the gender effect? The answer is no: first of all, the gender effect is stronger than the nationality effect, but more important: the female/male composition of Dutch and foreign student groups is similar.

Besides the gender and nationality effect, a third general effect is striking: the high p-values of most scales. Of the eight correct reasoning skills, five have p-values above .65. Of the eight misconception scales, only one has a p-values clearly larger than .35. Given the circumstance that only a minority of our inflow did attend formal education in statistics in secondary school, and a majority did not, one can wonder if the level of the instrument is appropriate for (European) high school. Next section will continue with that question, in the framework of the reliability of the instrument.

5. A second order analysis of SRA data

Table 4 presents the correlations of all scales. Bold-face figures represent estimates that are significantly different from zero at the 1% level. This table demonstrates the questionable reliability of the total or aggregate correct reasoning and misconceptions measures, referred to in Section 4. This finding is in line with what was found for USA and Taiwan students, see Garfield (2003). The Cronbach α reliabilities of the aggregated scales, taking the eight correct reasoning scales and the eight misconception scales as components, are 0.24 and 0.06, respectively. Replacing CC5 and MC1 by CC5A and MC1A has only a marginal effect: 0.25 and 0.07, respectively. Deleting individual items with extreme p-values, as suggested in Liu (1998), turns out to have little impact on reliabilities.

Table 4. *Correlations between all SRA Correct reasoning scales and misconceptions, based on 1303 UM student records; values in bold are significant at the .01 level.*

	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	MC1	MC2	MC3	MC4	MC5	MC6	MC7
CC1															
CC2	**.10**														
CC3	**.11**	.03													
CC4	.01	**.09**	**-.21**												
CC5	.03	.04	**.10**	-.04											
CC6	-.01	.06	-.01	.04	.05										
CC7	.04	**.13**	.06	**.08**	.01	.03									
CC8	.02	**.13**	**.08**	.05	-.01	.03	.07								
MC1	.04	**-.30**	.04	-.02	**-.19**	-.05	-.05	-.06							
MC2	**-.47**	-.05	**-.21**	-.04	.03	.05	-.02	**-.27**	.02						
MC3	**-.07**	-.06	-.03	.05	-.01	**.10**	-.02	**.08**	.04	**.08**					
MC4	-.00	**-.08**	**-.13**	.07	**-.68**	-.06	-.03	**-.14**	.07	-.04	.01				
MC5	.04	-.03	**.25**	**-.74**	.06	-.02	-.01	-.02	.03	**-.08**	-.03	-.06			
MC6	.07	-.05	.01	-.02	-.05	**-.47**	-.00	-.02	**.10**	-.05	**-.09**	.02	.03		
MC7	-.04	.06	**-.84**	**.21**	**-.13**	.02	-.06	-.04	-.03	-.02	.02	**.15**	**-.24**	.00	
MC8	-.02	**-.08**	-.04	.04	-.03	.01	-.05	-.02	.05	.04	**.12**	.02	-.05	.04	.03

Low correlations, few significant ones, and several wrong signs, even amongst the significant correlations, all together complement the conclusions of Garfield (1998b), Garfield & Chance (2000) and Liu (1998) with regard to the aggregated totals as a measure of reasoning abilities or misconceptions: '…the reliability of the instrument has yielded less than impressive results'.

Further study of the correlation matrix does suggest an alternative approach to summarize the outcomes of the SRA-instrument. Correlations within the group of correct reasoning scales, and within the group of misconceptions are, without exception, low. In the bottom-left rectangle, however, containing the correlations between correct reasoning skills and misconceptions, the first six columns each contain exactly one highly significant and strongly negative correlation. This is not surprising: from the definition of e.g. CC1 and MC2 it becomes clear, 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 correlation within a correct reasoning and misconception pair is due to the fact that these scales are based on the same multiple answer items. This leads to a negative correlation by construction if each multiple answer item has a single correct answer. In the SRA, several items are of the multiple answer format, which implies that choosing the correct answer does not exclude choosing one (or even more) incorrect answers. However, given the extreme p-values described in the last paragraph, this non-exclusiveness is more theoretical than empirical in nature, thus making specific pairs of correct reasoning and misconception negatively dependent by 'empirical construct'.

An exploratory factor analysis indeed provides evidence that the appropriate way of aggregating the SRA scales is according to the pattern visible in the correlation matrix. Factor analysing the data produces a seven factor solution, in which the first five factors, explaining together 54% of total variation, are composed of pairs consisting of one correct conception, and one misconception, with factor loading of about equal size and opposite signs (CC3 & MC7; CC5 & MC4; CC1 & MC2; CC6 & MC6; and CC4 & MC5).

6. Prior education and SRA

As indicated in section 3, data on prior education is different for students with a Dutch secondary school diploma, and students with an alternative kind of diploma (the German 'Abitur' being the most frequent 'other diploma'). Dutch students of our '99 shift took in high school a composition of at least seven different subjects (out of Dutch, French, German, English, Latin, or Greek language, Physics, Chemistry, Biology, Mathematics I, Mathematics II, Economics, Management, Geography, to mention the most important ones). Combinations were required to obey the following constraints: Dutch and English languages are obligatory, as is one of Mathematics I and Mathematics II. Math I is mainly calculus oriented and seen as the best preparation for social sciences, languages, arts and humanities. Math I contains an introduction to statistics and probability theory that is quite comparable to the Advanced Placement Program Statistics in the USA. Math II is mainly algebra and geometry oriented, and seen as the best preparation for sciences; it does not contain an introduction to statistics and probability. Math I and Math II are complementary, so a sizeable group of students, mainly in the social sciences preparing track and striving for an ambitious program, opted for both Math I & II. As indicated, that structure changed, and the '03 shift is roughly composed of only two groups of Dutch students: those who took basic mathematics (now called Math A, a program quite similar to the old Math I), and those who took advanced mathematics (now called Math B, similar in content to Math I & II). Since Math A and math B are not complementary anymore, the option to combine them is no longer available.

For foreign students, not that much is known about their prior education. Students were asked to classify their high school education into 'math major', 'math minor', and 'no math'. Since the meaning of this classification will be different for students of different countries, and Dutch students provide ample data, we will focus on Dutch students.

In this section, two questions will be posed:

- Does prior education, in specific differences in the prior math education students received, contributed to the explanation of differences in reasoning abilities and misconceptions?
- If so, might differences in prior education between female and male students help to explain the gender effect in SRA scores?

We will start by answering one aspect of the second question: are there differences in prior math education between females and males. The answer is yes, be it rather small. In the '03 shift, 10.3% of females and 12.8% of males have advanced mathematics as their mathematical preparation. These numbers are much lower than in the '99 shift: 16.3% of females and 25.5% of males took Math II (either in isolation, or combined with Math I). That the interest in advanced mathematics decreased that drastically is no surprise, given the high level of the new Math B program. Apparently, males are (still) somewhat more ambitious with regard to the science preparing topics in the high school program. This is partly offset by the fact that females tend to learn more in taking these programs: grades achieved by females in all topics are higher than those achieved by males, be it that the difference is small and insignificant.

6.1. Impact of prior math education on SRA

As can be concluded from Figures 3 & 4, the impact of prior math education on both correct conceptions and misconceptions is very small. The more advanced programs (Math II, Math B) tend to have higher correct conceptions scores and lower misconceptions scores, but only very few differences are significant, mainly in CC1, CC8, MC1A, and MC2.

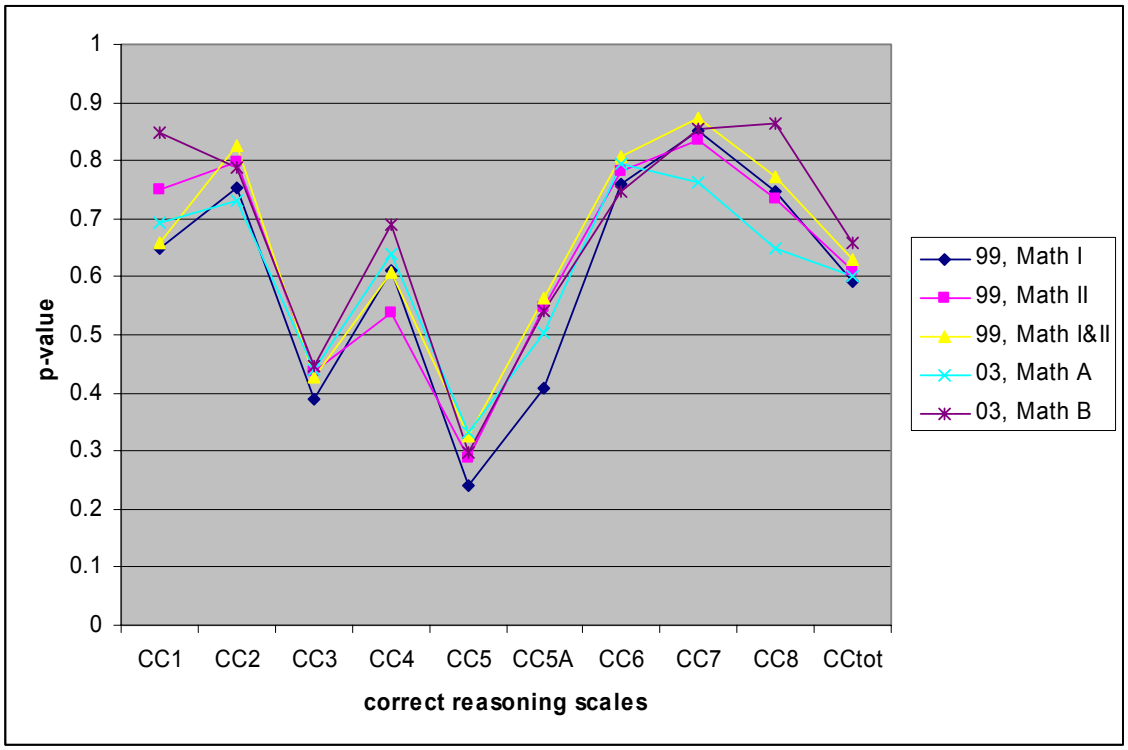

***Figure 3**: Reasoning abilities of students with different levels of math prior education*

Once again, in judging differences between the groups, one has to realize that two effects mix up: having received an introduction into statistics and probability (all groups do, except group '99, Math II'), being in a science preparing track or not ('99, Math II' and '03, Math B' do), and lastly, being in a social science preparing track with interest in math ('99, Math I&II'). From both figures it is apparent that, where differences exist, interest and student orientation are more important than prior schooling.

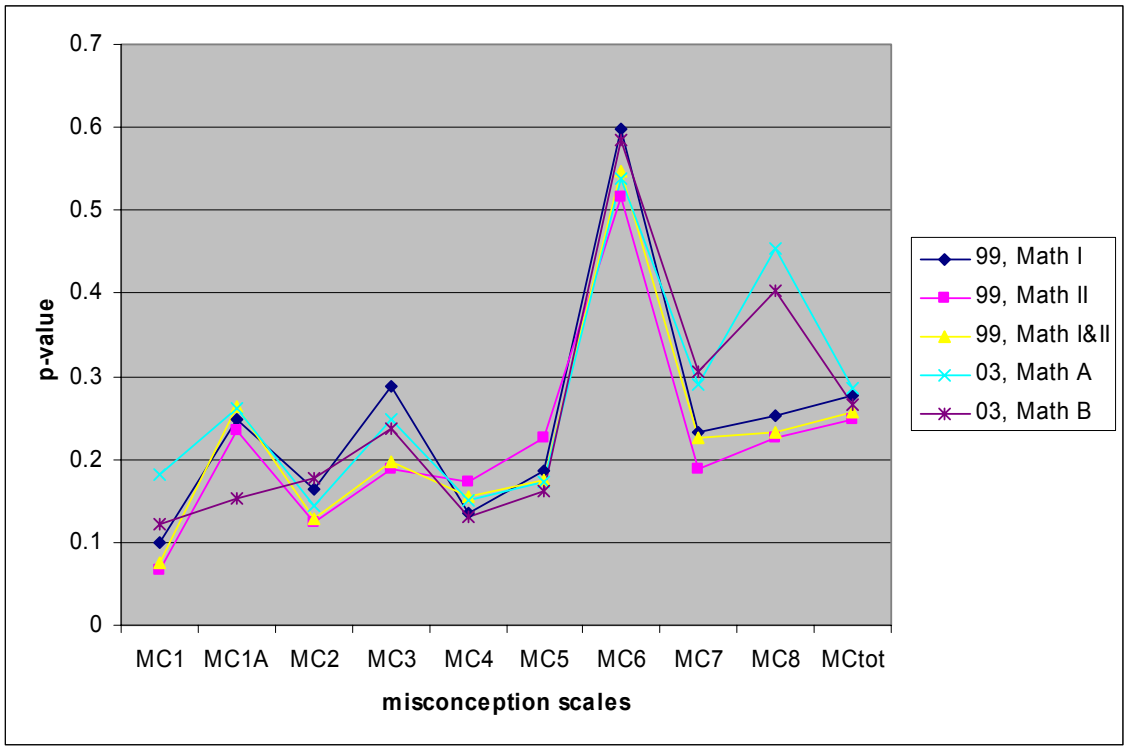

***Figure 4**: Misconceptions of students with different levels of math prior education*

6.2. Gender, prior math education and SRA

Although our inability to find strong prior education effects in reasoning abilities and misconceptions is in itself sufficient reason to exclude the possibility of gender effects being caused by differences in prior education, we continued with testing gender differences for SRA scales in each of the five subgroups separately. Outcomes are as expected: the gender effects remain but, since sample size reduces, fewer differences are significant. Particularly in CC5 (& CC5A, *understands sampling variability*), MC1 (& MC1A, *misconceptions involving averages*), and MC4 *(law of small numbers)*, strong and significant gender effects are visible in at least four out of the five subgroups. Remark that the scales with strongest gender effect are different from the scales with a prior education effect; another argument that these effects are independent.

6.3. Prior math knowledge and SRA

Prior education is not synonymous with prior knowledge: how much do students know in the subjects math and statistics when entering university. For the Dutch students, we know achieved grades in the national school exam. In this, we focus on grades for Math A and Math B. However, this restricts the sample to only Dutch students, in fact a subset of the Dutch students: those who took the subject in their final exam. In addition to these data, we administered a very small entry test on calculus and algebra (both five items, multiple choice). The level of the entry test appeared to be too low, making the discriminative power of the test rather low, but at least it provides information on all groups of students. Significant correlations (level .01) between the four indicators of prior knowledge and SRA scores are reported in Table 5; all data refer to the ’99 shift.

Table 5: *Correlations between math prior knowledge and SRA scales at .01 level (N=510, 510, 254, 122).*

	Correlations			
	Calculus entry exam	Algebra entry exam	Math A grade	Math B grade
CC2		.12		
CC5		.11		-.25
CC5A				-.24
CCtot		.15	.18	
MC1A		-.16	-.15	
MC3			-.17	
MC8			-.16	-.26
MCtot		-.14	-.20	

Several SRA scales, and most clearly the two aggregated scores, do depend on the prior math knowledge of students. Somewhat surprising is the rather strong negative relation between the grade in advanced mathematics and CC5(A). Apparently, the concept of sampling variability is at odds with science paradigm most of these students live in.

7. Attitudes and Beliefs towards Statistics and SRA

7.1. Attitudes and beliefs and the SATS instrument

In the context of mathematics education, the study of affective factors in learning processes has a long tradition and has given rise to terms like 'mathematics anxiety' that seem to be reserved for the mathematics domain only. In conceptualising the affective domain of mathematics education, McLeod (1992) distinguishes between emotions, attitudes and beliefs. 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. In many studies of learning processes, the focus is on beliefs and attitudes, rather than emotions, which are transient and hard to measure directly, but serve as a source for the development of attitudes and are thus measured indirectly; see e.g. Gal and Garfield (1997). Borrowing from this research tradition, there has grown a large body of literature on the role of attitudes and beliefs towards statistics has developed, in which one question keeps reappearing: Are attitudes and beliefs towards learning statistics distinct from the more general ones, such as towards learning mathematics, or towards exams in general? Gal and Ginsburg (1994) and Gal and Garfield (1997) are examples of this line of research. According to Gal and Garfield (1997), several reasons exist to consider the role of attitudes and beliefs about statistics in statistics education: the willingness of students to elect statistics courses (access considerations), their influence on the learning and teaching of statistics (process consideration), and their role in influencing students' statistical behaviour after leaving university (outcome considerations).

The area of research on developing instruments to assess attitudes and beliefs towards statistics is well developed. In the eighties, several instruments were developed, all using statements for which respondents mark their agreement or disagreement on 5-point or 7-point Likert-type. This contains the Statistics Attitude Survey, see Roberts and Bilderback (1980),

and, Roberts and Saxe (1982), the Statistical Anxiety Rating Scale, see Cruise, Cash and Bolton (1985), the Statistical Anxiety Inventory, see Zeidner (1991), and the Attitudes Towards Statistics, see Wise (1985). As each of these instruments had some drawbacks Schau, Stevens, Dauphinee, and DelVecchio (1995) developed the Survey of Attitudes Towards Statistics (SATS) in the nineties. The SATS consists of 28 seven-point Likert-type items measuring four aspects of post-secondary students' statistics attitudes. It has two forms, with minor differences in wording: a 'pre' form for students who have not yet taken a statistics course, and a 'post' form for administration during or after a course. The SATS contains four scales, see Schau et al. (1995), Dauphinee, Schau and Stevens (1997, and Gal and Garfield (1997), each accompanied by two examples of items, one positively and one negatively worded:

- Affect: measuring positive and negative feeling concerning statistics: *I like statistics; I am scared by statistics*; 6 items.
- Cognitive Competence: measuring attitudes about intellectual knowledge and skills when applied to statistics: *I can learn statistics; I have no idea of what's going on in statistics*; 6 items.
- Value: measuring attitudes about the usefulness, relevance, and worth of statistics in personal and professional life: *I use statistics in my everyday life; I will have no application for statistics in my profession*; 9 items.
- Difficulty: measuring attitudes about the difficulty of statistics as subject: *Statistics formulas are easy to understand; Statistics is highly technical*; 7 items.

In our research, we opted for the SATS instrument on grounds of the theoretical reasons that led to its development and the fact its statistical properties are well documented.

Several studies of gender differences in attitudes towards statistics are reviewed in Dauphinee, Schau, and Stevens (1997). Although the general conclusion of all these studies tends to be the same, there are some restrictions in comparing their outcomes since different researchers used different instruments, each having different scales. Roberts and Sache (1982), using the one-dimensional Statistics Attitude Survey developed by Roberts and Bilderback (1980), concluded that male students exhibit on average more positive attitudes than female students both at the beginning and at the end of an introductory statistics course. This is, in short, also the general conclusion of other research, but then with much more nuance brought forward by the use of multi-dimensional attitude scales. In an application of Wise's (1985) Attitudes Towards Statistics, Waters, Martelli, Zakrajsek, and Popovich. (1988) found that male students have more positive Course attitudes than female students, whereas no gender differences exits with respect to the other attitude scale.

7.2. SATS data and their relation to SRA

In our study, SATS was administered in the very first week of the course and can thus be viewed as an entry characteristic of the student. Since students would encounter several other inventories all expressed in terms of 5-point Likert type scale, it was decided to administer the SATS according to that scale instead of the 7-point Likert-type scale. Anchors were however the same: strongly disagree, neither disagree nor agree, strongly agree.

Table 6: *Average scores for scales Affect, Cognitive Competence, Value and Difficulty.*

SATS Scales:	(n=1392)	Female (n=427)	Male (n=646)	Dutch (n=556)	Foreign (n=518)
AFFECT	3.33	3.23	3.41	3.43	3.25
COGNC	3.58	3.47	3.67	3.59	3.60
VALUE	3.64	3.63	3.67	3.63	3.69
DIFFIC	2.73	2.68	2.77	2.81	2.65

All scale averages for Affect (AFFECT), Cognitive Competence (COGNC), and Value (VALUE) are far above (and statistically significant different from) the neutral level of three: students of different background have positive attitudes and beliefs in these aspects. In contrast, all mean scores for Difficulty (DIFFIC) are below the neutral level, expressing that students perceive the subject as difficult (the naming of the Difficulty-scale is somewhat counter intuitive: all scales are defined such that higher values correspond to more positive attitudes and feelings; a name like 'lack of perceived difficulty' would better catch this meaning). The table demonstrates some imbalance: students evaluate themselves as rather cognitive competent in learning statistics (overall mean score of 3.52 on the 5-point scale), but at the same time regard statistics as a somewhat difficult topic. Table 6 suggests that both gender and nationality effects may be present. Performing independent samples *t*-tests confirms this impression: male students have significantly higher scores on Affect, Cognitive Competence and Difficulty (all *p*-values less than .001); for Value, no difference exists. In comparing Dutch and foreign students, two scales demonstrate significant differences. foreign students score significantly lower in Affect and Difficulty than Dutch students, while they score higher (but not significantly) on Value and the Cognitive Competence score is invariant across nationalities.

With regard to the correlation structure of the four attitudes scales, our findings support the results reported in Gal and Garfield (1997): Affect and Cognitive competence are strongly related; Value and Difficulty are moderately related to Affect and Cognitive competence but not interrelated. See Table 7 for the correlations between the several scales, all significant at the .01 level.

Table 7: *Correlations between the four SATS attitude scales (N=1392)*

	Correlations			
	AFFECT	COGNC	VALUE	DIFFIC
AFFECT				
COGNC	.69			
VALUE	.33	.34		
DIFFIC	.48	.48	.10	

Do attitudes and beliefs have any impact on reasoning and misconceptions? If so, we expect this impact to be positive for the reasoning abilities, and negative for the misconceptions. At least two of the SATS scales, Cognitive competence and Difficulty, contain aspects of self-reported self-efficacy, and it is well known that self-efficacy is positively related to knowledge. Selecting only those correlations that are significant at the .01 level, our expectations come out: five positive correlations for CC2, CC3, and CC7, five

negative for MC3, MC7, and MC8: see Table 8, where rows without significant entries are deleted.

Table 8: *Correlations at .01 level between SATS and SRA scales (N=1076)*

	Correlations			
	AFFECT	COGNC	VALUE	DIFFIC
CC2				.10
CC3		.11		.13
CC7	.10	.11		
CCtot	.10	.15		.14
MC3	-.08	-.09		
MC7				-.08
MC8	-.10	-.09		
MCtot	-.10	-.10		-.09

Although of correct sign and statistically significant, the size of all correlations is very moderate, implying that attitudes play no substantial role in the explanation of the gender effect in SRA (in a structural equation model of attitudes, reasoning abilities, and course outcome, nor reported here, all paths between attitudes and reasoning abilities appear to non-significant).

8. SRA and course performance

8.1. Performance indicators

In our Quantitative Methods courses learning outcomes are measured with several instruments, each of them focussing on different aspects of the mastery of mathematical and statistical knowledge, see Gal and Garfield (1997) and Jolliffe (1997). The most important assessment instruments are:

- Final exams of the multiple choice format. To create a kind of external anchors, these exams are partly inspired by released Advanced Placement Statistics Exam. Like in the AP exam, our final exams will have a strong emphasis on conceptual issues, and students are allowed to use an extensive formula sheet, making the exam nearly of the 'open book' type. The exam covers both statistics and mathematics; those parts will be separately graded.
- Quizzes of multiple choice and short answer format (in the '03 shift, and experimental in the '99 shift). The quizzes allowed students to achieve a bonus score. The level of the items is more basic than in the final exam, the main purpose being to stimulate student to spread their learning efforts evenly in time. It is hypothesized that the quiz score is stronger effort-based than the exam score.
- Weekly hand-in assignments of open type (only in the '99 shift). The discussion of these assignments and the (partial) student solutions constitute the main agenda of the weekly, small-group, tutorial session. To start these discussions in full drive, students were credited with some bonus by doing preparatory work on these assignments outside the tutorial group. Even more than the bonus for quizzes, these scores are assumed to be very strongly effort-based: teaching assistant are explicitly instructed to assess the efforts put in by the students in trying to solve the homework problems, in stead of assessing if the correct solution is contained in the handed in work.

The program is divided over several half-semester periods (three in ’99, two in ’03). Each such period will have its own assessment, implying the availability of performance indicators for different subjects (math and stats), different assessment formats (homework, quiz, exam) and different periods (1, 2, and 3, or 1 and 2). For several reasons (the prime being the replacement of homework assignments by the use of an electronic learning tool in ‘03), we will focus on the ’99 shift.

Analysing descriptive statistics of these performance indicators bring forward several conclusions. First: the several performance indicators are strongly positively correlated. The strongest correlations are amongst indicators of the same type. Correlations between final exam scores for math and stats and the three different periods range between 0.4 and 0.6; for homework assignments scores between 0.5 and 0.8, and for quizzes, even above 0.9. But correlations between scores of different types of assessments instruments are not much lower: between quiz scores and homework scores, ranging from 0.6 to 0.8, between quiz scores and final exam scores, ranging from 0.3 to 0.6, and between homework scores and final exam scores, ranging from 0.2 to 0.6.

Second: there exists a strong gender effect in the bonus scores that students achieve for their hand-in assignments. This gender effect is present in mathematics and statistics, and both for Dutch and foreign students. The gender effect is always in the same direction: female students outperform male students. The effect is large: in all independent sample *t*-tests, (for all courses), all differences are significant in the same direction, most *t*-values being larger than 4 (implying *p*-values less than .001). Second: a similar significant gender effect is found in the bonus scores achieved for quizzes for Dutch students: female students outperform male students. For foreign students, differences are in the same direction, but not significant. Third: there exists an even much stronger nationality effect, in both bonus scores achieved for hand-in assignments, as for quizzes: foreign students outperform Dutch students, both for mathematics and statistics, in all courses, both for females and males. Differences are large: the smallest *t*-value is 5.

With regard to the written exams, the picture is completely different. For all mathematics exams, and the first statistics exam, males outperform females, both for Dutch as for foreign students. The *t*-values range from 1.5 to 3, making some differences significant, others not. In the second and third statistics exam, this pattern tends to reverse, female scoring higher than males; differences are however not significant. The nationality effect in exam scores demonstrates a somewhat similar development, with foreign students in the position of female students, and Dutch students in the position of male students. In the first exam, Dutch students do significantly better than foreign students, both in math (very large difference) and in statistics. In the second exam, Dutch and foreign level off for math, whilst foreign students significantly outperform Dutch students for statistics. And in the third exam, foreign students outperform Dutch ones both for math and for statistics significantly.

It is not that difficult to provide some intuition for these apparent differences. First of all: the match between secondary education and university study is much better for Dutch students than for foreign students. The compensating force is that foreign students on average put a lot more effort in their study than Dutch students. This difference in effort pays off in the more effort-based indicators such as bonus score already from the very first course on, and starts to pay off in the more cognitive based indicators in the second course.

The picture in the gender issue is similar: female students are willing to spend more efforts in their study than male students. This pays off from the very first course on, especially in the effort-based bonus scores. However, it is not obvious why females start at a lower level

in quizzes and exams, given the circumstance that differences in prior education are between small and absent.

Gender differences with regard to study behaviour seem to be a universal fact of life. In fact, the real differences may be even stronger than those that are apparent from these data, given the fact that entries with more than one missing value are deleted. Those entries correspond to students that missed exams, or did not participate in tutorial sessions and thus did not gain any bonus score. Leaving those 'no-show' and 'irregularly-show' students out, already eliminates the group showing the least efforts in studying. With regard to the nationality effect, one should be careful to regard this as a difference in culture effect. In choosing to study fulltime in a foreign country at a relatively young age, a choice made by those students enrolling in our program, implies a self-selection effect, which will play a major role, but cannot be identified separate from a culture effect.

8.2. SRA as predictor for performance indicators

What is the relationship between course performances and SRA scores, and how strong is this relationship? One would expect that correct conceptions would contribute in positive sense to performance indicators, whereas misconceptions do the reverse. Since earlier studies with the SRA achieved disappointing results in this respect, one would not expect to find a strong impact, and the impact is expected to vary amongst the scales, since some correspond to topics within the curriculum, others not. Tables 9 and 10 contain the correlations between SRA scales and performance indicators.

Table 9: *Correlations between SRA scales and course performance: bonus points for homework and quizzes. Significant at 0.05 level (bold: 0.01 level); N=680.*

SRA Scales:	Stats1 bonus	Stats2 bonus	Stats3 bonus	Math1 bonus	Math2 bonus	Math3 bonus	Quiz1 bonus	Quiz2 bonus
CC1								
CC2			**-.12**	**-.11**	**-.11**			
CC3								
CC4								
CC5								
CC6								
CC7			**-.10**	-.08	-.10	-.08		
CC8			-.08	**-.11**	-.09			
CCtot		-.09	**-.13**	**-.12**	**-.14**			
MC1		.08	**.11**	**.12**	.09			
MC2								
MC3	.08	.10	**.11**					
MC4								
MC5								
MC6								
MC7								
MC8								
MCtot	.09	.09	**.11**	**.13**	**.10**			

Performance indicators are ranked such that they start in Table 9 with the most 'effort-based' indicators, the bonus scores for the weekly home work assignments, through the weekly quizzes, and finish with the least effort-based but strongly cognitive oriented written

exams in Table 10. This design pays out, because striking differences between the three assessment categories show up.

Table 10: *Correlations between SRA scales and course performance: scores in final exam. Significant at 0.05 level (bold: 0.01 level); N=680.*

SRA Scales:	Stats1 score	Stats2 score	Stats3 score	Math1 score	Math2 score	Math3 score
CC1	**.14**			.08		
CC2	**.17**		**-.12**	**.21**	**.12**	.09
CC3				**.11**	.08	.09
CC4	.09			**.12**		
CC5					.08	
CC6						
CC7	**.14**			**.19**	**.10**	
CC8	**.11**			.08		
CCtot	**.24**			**.28**	**.18**	**.13**
MC1	**-.12**			**-.14**	**-.14**	-.09
MC2	-.09					
MC3	-.09					
MC4					**-.11**	
MC5						
MC6						
MC7				-.08	-.08	**-.11**
MC8	-.09			-.08		
MCtot	**-.18**			**-.18**	**-.17**	**-.16**

Starting with the written exams, we find a pattern that quite well fits the expectations: all significant correlations (and in fact, also nearly all insignificant ones) between correct reasoning skills and performance indicators are positive and, although not very large, still substantial of size (up to .28). At the same time, all significant correlations with misconceptions are negative, but somewhat smaller in size.

Weekly quizzes demonstrate a different pattern: their relationship to SRA scales is absent. Going one step further into more effort-based indicators, the great surprise comes with the correlations between weekly home work bonus scores and SRA scales: all significant correlations have the 'wrong' sign, that is correct conceptions scores correlate consistently negative with bonus scores, and misconception scores correlate consistently positive with bonus scores!

This somewhat paradoxical result might quite well explain why other studies did not find any relationship between SRA scores and course performance. If final grade is composed as a weighted average of several assessment instruments, each of them having a different effort content, the aggregation process might cancel out the relationships between SRA scales and separate performance indicators. Or, as an alternative explanation, if progress tests like quizzes or mid term exams contribute strongly to grades, once again a condition is created in which decencies on SRA scales remain hidden. It is only through the two extremes, traditional final exams focussing on cognitive aspect on the one side, and scores for home assignments on the other, that the impact of reasoning abilities and misconceptions becomes visible. In our analysis, we assume effort to be the mediating variable. If this assumption is correct, remains to be investigated.

Some other conclusions of the pattern of correlations are the following.

- In general, the impact of reasoning abilities and misconceptions 'dies out in time': comparing period 1, period 2, and period 3 correlations for the same scale and performance indicator, the absolute values of the correlations typically decrease. See for example the correlation between MCtot and the Math score in the written exam: -.18 => -.17 => -.16.
- Although designed as an instrument to measure statistical reasoning abilities and misconceptions, and not general or mathematical reasoning skills, the SRA scales have higher explanatory power to math performance indicators, than to statistics performance indicators, especially with regard to the written exams
- Explanatory power for statistics exams seem to be restricted to the period 1 exam, covering descriptive statistics and probability theory. SRA does not explain period 2 and 3 exams, covering inferential statistics and the regression model, respectively. This is not surprising, since nearly all SRA items refer to topics in descriptive statistics and probability, and not to the more advanced topics covered in later periods. So the impact of SRA levels on statistics exam scores is content specific. In contrast, the impact of SRA levels on math exam scores is, by definition, not content specific and not restricted to only one period.
- Simple regression models explaining exam scores in statistics and math in period 1 by individual SRA scales, or aggregated SRA scales respectively, have reasonable explanatory power: R^2 is 8.1% (5.7% respectively) for Stats1, 10.0% (8.0% respectively) for Math1. Those numbers compare quite well the explanatory power of e.g. attitudes and beliefs, the SATS scales, towards exam scores. The Stats1 score gets explained by CC1, CC2, CC4, CC7, CC8, and MC8, the Math1 score by CC2, CC3, CC4, and CC7.

9. Discussion and educational implications

Most statistics programs adapted to the education reform movement contain a portfolio of different course assessments. Some assessment instruments are highly effort-based, as homework assignments and projects, while some are more cognitive based, as final exams. In general, correlations between course outcomes as assessed by these different instruments tend to be rather high; see for our study section 8.1. Grading students with a portfolio, instead of a single final exam, thus seems not to have a strong impact on grading decisions. Choosing for a rich portfolio is therefore better understood by the desire to stimulate students in their learning, than to drastically change the grading outcomes.

The SRA-instrument is a natural candidate for any assessment portfolio in introductory statistics. However, in comparing its outcomes with other instruments, it takes a unique position: correlations with final exam outcomes are weakly positive, correlations with effort-based instruments as homework assignments are however weak but negative. The weakness in the positive correlations found in this study might not be that problematic: it is after all a pre-test, and reasoning skills as measured by SRA are not included explicitly as course goals.

More problematic might be the negative (be it weak) relationship between study efforts (as measured by the bonus for homework assignments) and the SRA outcomes. One interpretation of this is that a learning approach that is strongly effort-based might be a hindrance in becoming skilful in statistical reasoning. If this hypothesis is correct (in a further study, in which we relate SRA scores with scores on a learning style inventory, we intend to investigate this hypothesis), it will have a strong impact on statistics education: studying effortful appears to bring one far, even in rather cognitive based final exam (given the positive correlations between exam scores and homework scores as reported in section 8.1). However,

for reasoning skills this appears not to be true: correlations between effort and reasoning skills are even negative, implying that students with an effortful learning approach are sensitive to misconceptions and lack of reasoning skills.

One of Garfield's (2002) conclusions is that the quality of teaching, and the performance of students on their exams, does not tell that much about students' reasoning skills and their level of integrated understanding. This study adds that also the quality of learning, in terms of the effort invested in it, does not guarantee proper reasoning skills, and in fact, if it guarantees anything, strong study efforts are more a hindrance than a help in achieving reasoning skills. Educational reforms in the nineties strengthened the role of independent student learning, at the expense of teaching, in most European educational systems. But neither traditional teaching, nor independent student learning seem to be the designated tools in acquiring reasoning skills. Chance (2002) describes several instructional tools that allow 'thinking beyond the textbook'. The outcomes of this study emphasise the importance of using activities and other tools discussed by Chance: they not merely supplement traditional learning, but produce learning outcomes not achieved by other means.

10. Conclusions

From the several sections, a number of conclusions are 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 as the outcome of factor analysis on which both correct and misconception scales are allowed to load seem to offer higher reliability.

We administered the SRA as an entry test for freshmen, where a majority of these freshmen did not receive any formal schooling in statistics or probability. In spite of that, most correct reasoning scales have p-values above .65, most misconception scales p-values of .35 or less. To allow for efficient discrimination in reasoning abilities for (European) freshmen, increasing the difficulty level of SRA would be attractive.

The SRA was designed with the explicit aim to assess aspects specific for the domain statistics. Although the items themselves truly focus on statistical aspects, and in that sense the design fulfils its aims, data generated by the administration of the instrument make a different picture: although relationships with learning outcomes are weak, they are stronger towards those in mathematics than in statistics. This is at least puzzling, since it suggests that statistical reasoning is not that different from mathematical reasoning (at least: in the eyes of our students, since it is self-report instruments) as is hoped for in the statistics education community.

SRA results demonstrate strong gender effects, that don't disappear by accounting for differences in prior education or prior knowledge. This is at least puzzling, and deserves further investigation.

Investigating the relationship between statistical reasoning and course performance indicators generates a remarkably dichotomous picture. The strongest effort-based performance indicators, bonus for homework assignments, is negatively related to reasoning abilities, weekly quiz outcomes, taking a more central position, is unrelated to reasoning abilities, and written exam grades, most cognitive in nature, is positively related to reasoning abilities.

If effort is indeed an important mediator, then a next puzzle arises: why and how acts 'effortful learning' as an obstacle for achieving reasoning abilities? If the assumption on the role of effort is true, this puzzle might replace the gender puzzle, since (at least in our sample) females are known to put more effort in their study than males.

The SRA-based reasoning constructs appear to be much weaker related to attitudes constructs, based on the SATS instrument, than course performance indicators. Our search to find factors influencing statistical reasoning, and so provide a partial answer to Lovett's (2001) conclusion that '*existing research on students' difficulties in learning statistical reasoning does not offer much explanation of what causes these difficulties nor provides much guidance in devising specific solutions for overcoming them*', is in this respect unsuccessful. Or, reformulated somewhat more positively: it is not through negative affects towards statistics that students opt for unlearned, intuitive reasoning above learned, statistical reasoning in solving statistical problems.

The educational implications of this study are based on the inverse relation found between study efforts and reasoning skills. Students that do well in modern, student-centred learning systems appear to do less well in reasoning. Good teaching appears not to be very helpful in acquiring reasoning skills. Attractive students' characteristics for independent learning, such as the willingness to spend much efforts in the study, qualifies even worse: it appears to stand in the way of attaining reasoning skills.

11. References

Bransford, John D. et al., eds. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Committee on Developments in the Science of Learning with additional material from the Committee on Learning Research and Educational Practice, National Research Council. Washington: National Academy Press.

Chance, Beth L. (2002). Components of Statistical Thinking and Implications for Instruction and Assessment. *Journal of Statistics Education* [Online], 10(3). (http://www.amstat.org/publications/jse/v10n3/chance.html)

Chance, Beth L. and Garfield, Joan B. (2002). New Approaches to Gathering Data on Student Learning for Research in Statistics Education. *Statistics Education Research Journal* [Online], 1(2), 38-41. (http://fehps.une.edu.au/f/s/curric/creading/serj/past_issues/SERJ1(2).pdf)

Cruise, R.J.; Cash, R.W. & Bolton, D.L. (1985). Development and validation of an instrument to measure statistical anxiety. *American Statistical Association Proceedings of the Section on Statistics Education*,92-97.

Dauphinee, Thomas L., Schau, Candace, & Stevens, Joseph J. (1997). Survey of Attitudes Toward Statistics: Factor Structure and Factorial Invariance for Women and Men. *Structural Equation Modeling: a multidisciplinary journal*, 4 (2), 129-141.

delMas, Robert C. (2002a). Statistical Literacy, Reasoning, and Learning. *Journal of Statistics Education* [Online], 10(3). (http://www.amstat.org/publications/jse/v10n3/delmas_intro.html)

delMas, Robert C. (2002b). Statistical Literacy, Reasoning, and Learning: A Commentary. *Journal of Statistics Education* [Online], 10(3). (http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html)

Fox, Marye Anne and Hackerman, Norman Eds. (2003). *Evaluating and Improving Undergraduate Teaching in Science, Technology, Engineering, and Mathematics.* Committee on Recognizing, Evaluating, Rewarding, and Developing Excellence in Teaching of Undergraduate Science, Mathematics, Engineering, and Technology, National Research Council. Washington: National Academy Press.

Gal, Iddo & Garfield, Joan B. (1997). *Curricular Goals and Assessment Challenges in Statistics Education*. In: Gal & Garfield, *The Assessment Challenge in Statistical Education*. Voorburg: IOS Press.

Gal, Iddo & Ginsburg, Lynda (1994). The Role of Beliefs and Attitudes in Learning Statistics: Towards an Assessment Framework. *Journal of Statistics Education* [Online], 2 (2). (http://www.amstat.org/publications/jse/v2n2/gal.html)

Garfield, Joan B. (1995). How students learn statistics. *International Statistical Review*, 63 (1), 25-34.

Garfield, Joan B. (1996). Assessing student learning in the context of evaluating a chance course. *Communications in statistics; Part A: Theory and methods*. 25(11), 2863-2873.

Garfield, Joan B. (1998a). *Challenges in Assessing Statistical Reasoning*. AERA Annual Meeting presentation, San Diego.

Garfield, Joan B. (1998b). The Statistical Reasoning Assessment: Development and Validation of a Research Tool. In: L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W.K. Wong (eds.), *Proceedings of the Fifth International Conference on Teaching Statistics*, 781-786. Singapore: International Statistical Institute.

Garfield, Joan (2002). The Challenge of Developing Statistical Reasoning. *Journal of Statistics Education* [Online], 10(3). (http://www.amstat.org/publications/jse/v10n3/garfield.html)

Garfield, Joan B. (2003). Assessing Statistical Reasoning. *Statistics Education Research Journal* [Online], 2(1), 22-38. (http://fehps.une.edu.au/F/s/curric/cReading/serj/current_issue/SERJ2(1).pdf)

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

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

Garfield, J.B.; Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review*, 67 (1), 1-12.

Garfield, Joan; Hogg, Bob; Schau, Candace; & Whittinghill, Dex (2002). First courses in statistical science: the status of educational reform efforts. *Journal of Statistics Education* [Online], 10(2). (http://www.amstat.org/publications/jse/v10n2/garfield.html)

Goldberg, L. R. *International personality item pool.* http://ipip.ori.org/ipip/

Gollub, Jerry P. et al. eds. (2002). *Learning and Understanding:Improving Advanced Study of Mathematics and Science in U.S. High Schools*. Committee on Programs for Advanced Study of Mathematics and Science in American High Schools, National Research Council. Washington: National Academy Press.

Jolliffe, Flavia (1997). Issues in constructing assessment instruments for the classroom. In: Gal & Garfield, *The Assessment Challenge in Statistical Education*. Voorburg: IOS Press.

Jolliffe, Flavia (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*, 801-806. Singapore: International Statistical Institute.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

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

Lecoutre, M.P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics*, 23, 557-568.

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

Lovett, Marsha (2001). A Collaborative Convergence on Studying Reasoning Processes: A Case Study in Statistics. In: Carver, Sharon M., & Klahr, David: *Cognition and*

Instruction, Twenty-Five Years of Progress; 347-384. Mahwah: Lawrence Erlbaum Associates.

McLeod, Douglas B. (1992). Research on Affect in Mathematics Education: a Reconceptualization. In: Grouws, Douglas A. (ed.) (1992). *Handbook of research on mathematics teaching and learning*, A project of the National Council of Teachers of Mathematics. New York: Macmillan. Ch. 23, pp. 575-596.

O'Connell, Ann Aileen (1999). Understanding the Nature of Errros in Probability Problem-Solving. *Educational Research and Evaluation*, 5(1), 1-21.

Reading, Chris (2002). The International Research Forum on Statistical Reasoning, Thinking and Literacy: Summaries of presentations at STRL-2. *Statistics Education Research Journal* [Online], 1(1), 30-45.
(http://fehps.une.edu.au/f/s/curric/creading/serj/past_issues/SERJ1(1).pdf)

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., & Saxe, J.E. (1982). Validity of a statistics attitude survey: A follow-up study. *Educational and Psychological Measurement*, 47, 907-912.

Rumsey, Deborah J. (2002). Statistical Literacy as a Goal for Introductory Statistics Courses. *Journal of Statistics Education* [Online], 10(3).
(http://www.amstat.org/publications/jse/v10n3/rumsey2.html)

Schau, Candace; Stevens, Joseph; Dauphinee, Thomas L.; & l Vecchio, Ann De (1995). The Development and Validation of the Survey of Attitudes Toward Statistics. *Educational and psychological measurement*, 55 (5), 868-875.

SERJ (2002): *Statistics Education Research Journal*, 1(1), 30-45. The International Research Forums on Statistical Reasoning, Thinking and Literacy: Summaries of Presentations at SRTL-2.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In: D. A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning*. New York: Macmillan, 465-494.

Sundre, D.L. (2003). *Assessment of Quantitative Reasoning to Enhance Educational Quality*. AERA annual meeting presentation, Chicago. Available through the ARTIST web site: http://www.gen.umn.edu/artist/articles/AERA_2003_QRQ.pdf.

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

Watson, Jane M.; Collis, Kevin F.; Callingham, Rosemary A.; & Moritz, Jonathan B. (1995). A Model for Assessing Higher Order Thinking in Statistics. *Educational Research and Evaluation*, 1(3), 247-275.

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

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, 61, 319-328.