Puzzles in Statistical Reasoning

Dirk T. Tempelaar
Wim H. Gijselaers
Sybrand Schim van der Loeff
Maastricht University

*Journal of Statistics Education* Volume 14, Number 1 (2006),
ww2.amstat.org/publications/jse/v14n1/tempelaar.html

Copyright © 2006 by Dirk Tempelaar, Wim Gijselaers, and Sybrand Schim van der Loeff, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

**Key Words:** Assessment; Attitudes toward statistics; Learning approaches; Statistical reasoning assessment.

**Abstract**

The Statistical Reasoning Assessment or SRA is one of the first objective instruments developed to assess students’ statistical reasoning. Published in 1998 (*Garfield, 1998a*), it became widely available after the *Garfield (2003)* publication. Empirical studies applying the SRA by Garfield and co-authors brought forward two intriguing puzzles: the ‘gender puzzle’, and the puzzle of ‘non-existing relations with course performances’. Moreover, those studies find a, much less puzzling, country-effect. The present study aims to address those three empirical findings. Findings in this study suggest that both puzzles may be at least partly 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. In distinction with earlier studies, we administered the SRA at the start of our course. Therefore measured reasoning abilities, correct as well as incorrect, are to be interpreted unequivocally as preconceptions independent of any instruction in our course. Implications of the empirical findings for statistics education are discussed.

**1. Introduction**

The present study aims to explore the role of statistical reasoning in learning statistics. To this purpose the definition of statistical reasoning by *Garfield and Chance (2000)* will be employed. Statistical reasoning, according to their definition, is 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.

In order to measure statistical reasoning the Statistical Reasoning Assessment or SRA, developed by
Garfield (1998a, 2003) has been used. Garfield and co-authors have performed several empirical analyses on the SRA (Garfield 1998b, 2003; Garfield and Chance 2000; Liu 1998). One of the striking outcomes of this research is the puzzle of "non-existing relations with course performances": correlations between aggregated reasoning skills demonstrate low or zero correlations with course performances. A second puzzle emanating from this empirical work is the "gender puzzle": female and male students demonstrate striking differences in their reasoning abilities. In addition to these two puzzles, a third, though less surprising, effect is found: a country or nationality effect. This paper addresses these puzzles with the purpose to further the understanding of statistical reasoning and the assessment of it through the instrument SRA.

Statistical reasoning, and the related concepts of statistical thinking and statistical literacy, are at the center 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 American Educational Research Association (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. In the closing summary of the JSE Winter 2002 series, delMas (2002b) emphasizes the role of assessment. Although a lot of progress has been made in the delineation of the concepts reasoning, thinking and literacy, and the elaboration of instructional implications of research findings in each of the areas, we are still rather empty-handed with regard to instruments that assess students’ abilities. 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 SRTL forums on Statistical Reasoning, Thinking and Literacy, of which the first two editions are reported in Ben-Zvi and Garfield (2004)]. Objective instruments that can be applied on a broad scale in classes as large as the one reported on in this study are, to our knowledge, limited to the SRA instrument.

The relationship between statistical reasoning (and related concepts) and the learning of statistics is a complex one. 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 of sampling distributions. This is what Gal and Garfield (1997) call the outcome consideration. Expressed by Garfield (2002, p. 9): “it [statistical reasoning] appears to be universally accepted as a goal for students in statistics classes”. But in addition to being an important output of statistics education, statistical reasoning is also a crucial input in the process of learning statistics: the process consideration. Students enter our classes with prior reasoning skills; to the extent that these prior skills correspond to true knowledge being part of the course achievements aimed at, these prior skills will ease the learning process. However, an important category of prior knowledge is formed by misconceptions, or intuitive but faulty reasoning mechanisms. Both types of preconceptions are, according to modern learning theories (Bransford, Brown and Rodney 2000) crucial determinants in learning; if preconceptions are not properly addressed, newly learned knowledge might appear much more volatile than existing preconceptions brought into class. Research on learning in general (see e.g. Bransford, et al. 2000), and on statistical reasoning in particular (Garfield and Ahlgren 1988; Shaughnessy 1992), make clear that the intuitive misconceptions are of a stubborn nature. It has been demonstrated 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. They seem to rely on incorrect intuitions already present when entering the course. Therefore teaching correct conceptions – no matter how successfully – is no guarantee for students not applying misconceptions anymore. Examples of stubborn fallacies in student’s statistical reasoning are 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).

In the above mentioned studies, empirical analyses into statistical reasoning on the basis of the SRA-instrument has been performed by Garfield and co-authors. In all of these analyses the SRA was...
administered at the end of a course, parallel to the final exam. Their main aim was to investigate the mastery of reasoning skills and its relationship to course performances. In such a design, measured skills are a mixture of those newly achieved in the course, and those already present at the start of the course. In contrast to these studies, we administered the SRA in the very beginning of the first introductory course. Its outcomes are, thus, to be regarded as students’ preconception levels achieved outside class or, in some cases, in high school programs, independent of our own curriculum. This difference in timing of administering SRA makes it possible to focus on the role of prior conceptions and misconceptions in learning statistics in our course.

Since reasoning abilities are measured in the very beginning of the course, instruction-related variables are excluded as a possible ‘contaminant’. Differences in statistical reasoning can thus be wholly attributed to, what Garfield and Ben-Zvi (2004) call the ‘diversity of students’. That diversity can refer to several student-related aspects. Ben-Zvi and Garfield (2004) contains a range of studies on statistical literacy, reasoning, and thinking in which student diversity expresses itself primarily in differences in prior education and mastery. But student diversity has more manifestations than these cognitive aspects, and in this study, we will include two non-cognitive factors expected to have an impact on learning statistics. The first of these is constituted by the affective factor students’ attitudes towards statistics. Attitudes are found to be important factors in learning statistics for several reasons: see for example Nasser (2004), Gal and Ginsburg (1994), Gal and Garfield (1997), Gal and Garfield (1997) distinguish e.g. between access considerations (the willingness of students to elect statistics courses), process considerations (their influence on the learning and teaching of statistics, the focus of this study), and outcome considerations (their role in influencing students’ statistical behavior after leaving university). Analogous to its role in learning statistics, we hypothesise that positive attitudes contribute to a better state of prior reasoning abilities and misconceptions. A second aspect of students’ diversity incorporated in this study is the typical way students tend to study: their learning strategies, or more generally, their learning approaches. In statistics education, this theme has received less attention than the role of attitudes, in contrast to empirical research in learning in general; see for example Biggs (2003), Bransford et al. (2000). Typically, learning theories based on student approaches to learning distinguish between deep and surface learning (Biggs 2003; Vermunt and Vermetten 2004; Duff, Boyle, Dunleavy, and Ferguson 2004). Students taking a deep learning approach are more or less our ideal students: triggered by an interest in the topic under study, these students focus on underlying meaning, on main ideas, principles and applications. In contrast, students taking a surface approach to learning are characterised by a focus on memorisation, root learning, but no real attempt of understanding. In agreement with findings of general learning theory based on learning approaches, we hypothesise that a deep approach has a positive impact on, and a surface approach has a negative impact on, the level of statistical reasoning abilities students possess at the start of the course.

In Subsection 3.1 it is established that both puzzles and the nationality effect are indeed present in our data. The first puzzle, that of the non-existing relation between reasoning abilities and course performances, is studied in Subsection 3.2. In Subsection 3.3 we examine if part of gender and nationality effects can be retraced to differences in attitudes toward statistics on the basis of the hypothesis, made above that positive attitudes towards statistics contribute to a better state of prior reasoning abilities and misconceptions. In Subsection 3.4 we examine if part of gender effect can be retraced to differences in learning approaches on the basis of the hypothesized different impacts on the level of statistical reasoning abilities that students possess at the start of the course. Integrating cognitive and non-cognitive factors explaining reasoning abilities, regression equations are presented in Subsection 3.5 for both reasoning abilities and misconceptions, where the role of gender appears to be restricted. It is not so much gender itself, but a complex of gendered characteristics describing a preferred learning approach of students that explains a limited but consistent part of statistical reasoning abilities. That integrative model addresses two of the puzzles and effects discussed in this contribution: the gender puzzle, and the nationality effect. Section 4 closes with discussion and educational implications.
2. Method

2.1 Setting and Subjects of this Study

The present study was conducted in the setting of a course Quantitative Methods (QM) being part of both Economics and Business first-year programs. It is an introductory course covering regular level-100 subjects from mathematics, statistics and computer skills. The mode of instruction of the course is one where students meet in small groups of approximately twelve students with a tutor to discuss their solutions to any – usually homework – problems supplemented by lectures.

Data were collected on three shifts of students: approximately 900 first-year students participating in the 99/00 QM course, and approximately 850 students participating both in 03/04, and in 04/05. In addition to those first-year students, another 10% percent of the students are ‘repeat’ students who did not manage to pass that specific course in previous years. All courses are taught in English. The faculty attracts a relatively large proportion of foreign students. In 99/00, the share of foreign students was 46%, a figure that has risen to 65% in 04/05. 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 to nationality is important since major differences exist between secondary school systems in and outside Europe.

Most data used in this study are collected by students to be analyzed in their student projects. The topic of these projects has been ‘a statistical analysis of my study behavior,’ in which the course participants compare their study habits with that of fellow students. In order to provide data for such a comparison, all students have completed several questionnaires in the first weeks of the course. The results, both individual data and aggregated group data, have been 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. Other questionnaires that were administered are the Survey on Attitudes Towards Statistics (SATS,) and the Inventory of Learning Styles (ILS). The several questionnaires were administered in the tutorial sessions (99/00) or through web based forms (03/04 and 04/05). Due to the prospect of achieving bonus points for the student project, participation in the questionnaires was attractive and responses have been quite high. It is not possible to express the response rates as single figures, because different questionnaires were administered in different sessions (days), with different students being present. Most of the analyses reported here are based on the responses of about 2000 students (720, 580, and 700 in shifts 99/00, 03/04, and 04/05, respectively). The majority of the other students officially enrolled in the course would typically participate in the exam, but not in any educational activities.

2.2 Instruments: Statistical Reasoning Assessment, or SRA

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). In contrast to most other assessment instruments, it consists of closed format items, and it is therefore one of the few available instruments for large-scale assessment of statistical reasoning abilities of students at a pre-university level (see e.g. Gal and Garfield 1997, for a survey of assessment tools). 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. 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:**

Correctly interprets probabilities. Assesses the understanding and use of ideas of randomness,

CC1: chance to make judgments about uncertain events.

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.

Correctly computes probability, both understanding probabilities as ratios, and using combinatorial reasoning.

CC3: 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

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

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.

Understands the importance of large samples. Assesses the knowledge of how samples are related to a population

CC8: 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:**

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.

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).

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.

Law of small numbers. Small samples best resemble the populations from which they are sampled,

MC4: so are to be preferred over larger samples.

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).
Studies reporting empirical data on the application of SRA are limited, and partly overlap in experiments they describe: Garfield (1998b, 2003), Garfield and 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 and Chance 2000). The resulting correlations were low, suggesting that statistical reasoning and misconceptions were rather unrelated to students’ performance in that first statistics course.

Garfield (1998b), Garfield and 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, all of these studies analyze 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. We will follow the tradition of earlier studies in analyzing aggregated scores.

### 2.3 Instruments: Survey of Attitudes Toward Statistics, or SATS

Research in the affective domain of statistics education has lead to the development of several self-scoring instruments in the eighties, all using statements for which respondents mark their agreement or disagreement on 5-point or 7-point Likert-type; see Hilton, Schau, and Olsen (2004) for an overview. As each of these instruments had some drawbacks Schau, Stevens, Dauphinee, and DelVecchio (1995) developed the Survey of Attitudes Toward Statistics (SATS) in the nineties.

The SATS consists of 28 seven-point Likert-type items measuring four aspects of post-secondary students’ statistics attitudes. The SATS contains four scales, see Schau, et al. (1995), Dauphinee, Schau and Stevens (1997), and Gal and Garfield (1997):

- **Affect**: measuring positive and negative feeling concerning statistics;
- **Cognitive Competence**: measuring attitudes about intellectual knowledge and skills when applied to statistics;
- **Value**: measuring attitudes about the usefulness, relevance, and worth of statistics in personal and professional life;
- **Difficulty**: measuring attitudes about the difficulty of statistics as a subject.

In a recent extension of the instrument, two more scales were added, each covered by four items: Interest, and Effort (better called planned effort, since the instrument is used as an ex ante measurement) (Schau 2004, personal communication). This extended version was available for the last of the three shifts of students incorporated in this study only. 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.

### 2.4 Instruments: the Inventory of Learning Styles, or ILS

Students participating in our study made a profile of their own learning preferences using the instrument: Inventory of Learning Styles (ILS). The ILS aims at measuring the following components of student learning: cognitive processing strategies, metacognitive regulation strategies, conceptions of learning, and learning orientations (Vermunt and Vermetten, 2004; and numerous references in that source). The ILS consists of 120 statements covering all learning components. Students are asked to indicate, on a five-point
scale, the degree to which they use the described learning activities in their studies, or to what degree the described views and motives correspond to their own views and motives. **Table 2** describes the several scales within each of the learning components.

| Processing strategies          | Regulation strategies              | Learning orientations       | Mental models of learning |
|-------------------------------|-----------------------------------|-----------------------------|---------------------------|
| Relating and structuring      | Self-regulation of learning       | Personally interested       | Construction of knowledge |
|                               | processes                         |                             |                           |
| Critical processing           | Self-regulation of learning       | Certificate directed        | Intake of knowledge       |
|                               | content                           |                             |                           |
| Memorising and rehearsing     | External regulation of learning   | Self test directed          | Use of knowledge          |
|                               | processes                         |                             |                           |
| Analysing                     | External regulation of learning   | Vocation directed           | Stimulating education     |
|                               | results                           |                             |                           |
| Concrete processing           | Lack of regulation                | Ambivalent                  | Co-operation              |

The first two Processing strategies, Relating & structuring and Critical processing, together constitute deep processing strategies, while the next two, Memorizing & rehearsing, and Analyzing, represent stepwise or surface processing strategies. Applications of the ILS by Vermunt and co-authors reveal four typical styles or profiles for university students in the first years of their studies (Vermunt and Vermetten 2004). The first style demonstrates high scores on the Relating & structuring, and Critical processing strategies, both Self-regulation scales, Construction of knowledge as conception of learning, and Personal interest as learning orientation. This style is interpreted as a deep or meaning-directed learning pattern. The second style represents a surface or reproduction-directed learning pattern, with high scores on the ILS scales Memorizing & rehearsing, Analyzing, both External regulation scales, Intake of knowledge as conception of learning, and Certificate and Self-test-directed learning orientations. The third and fourth style, representing undirected learning and application-directed learning, typically occur less frequent than the first two.

### 2.5 Instruments: Course performance

In agreement with assessment literature (Gal and Garfield 1997; Jolliffe 1997), in our Quantitative Methods course learning performances are measured with a portfolio containing several instruments, each of them focusing on different aspects of the mastery of mathematical and statistical knowledge. Besides the before mentioned student project, the assessment instruments are:

- Final exams of the multiple choice format. To create a kind of external anchor, 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; both parts are graded separately.
Quizzes of multiple choice and short answer format (in the 03/04 and 04/05 academic years and experimental in the 99/00 academic year). The quizzes allow 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 homework assignments of open type (only in the 99/00 academic year). The discussion of these assignments and the (partial) student solutions constitute the main agenda of the weekly, small-group, tutorial sessions. To get the discussions started, students were credited with some bonus for 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 assistants are explicitly instructed to assess the efforts put in by the students in trying to solve the homework problems, instead of assessing the correctness of the solution handed in. The success of the experiment with quizzes in the 99/00 shift led to the abandonment of the assessment of homework in later shifts.

3. Results

3.1 Analysis of SRA data

Descriptive statistics of the present SRA data, similar to those reported in Garfield (1998b, 2003), Garfield and Chance (2000) and Liu (1998), are reported in Table 3. The exhibit presents the means of the several scales of all female and male students, and all Dutch and international students, expressed as a proportion, that is on a [0-1] scale. In addition to scores on eight reasoning skills (CC1 ... CC8), and eight misconceptions (MC1 ... MC8), the aggregated reasoning score (CCtot) and aggregated misconceptions (MCtot) are reported. The aggregated scores are obtained in the same way as in the studies by Garfield and co-authors by respectively taking the sum over all correct reasoning and misconception items, and re-expressing them 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. Added to the proportional scores are two measures that signal the existence of gender-effects and nationality-effects: the $p$-value of the independent samples $t$-test, and the Cohen’s $d$ measure of effect size, calculated as the difference in means divided by the pooled standard deviation (Cohen, 1988).

Table 3. Means of SRA Correct Reasoning scales and MisConceptions scales for Male and Female, Dutch and international students, and corresponding gender and nationality effects

| Correct Reasoning                  | Females (N=779) | Males (N=1209) | $p$-value | Effect size | Dutch (N=1080) | International (N=899) | $p$-value | Effect size |
|------------------------------------|----------------|----------------|-----------|-------------|----------------|------------------------|-----------|-------------|
| CCtot: total Correct Reasoning     | 0.54           | 0.57           | 0.000     | 0.24        | 0.59           | 0.54                   | 0.000     | 0.44        |
| CC1: Correctly interprets...       | 0.70           | 0.72           | 0.066     | 0.08        | 0.71           | 0.71                   | 0.929     | 0.00        |
| CC2: Understands...                | 0.68           | 0.74           | 0.000     | 0.24        | 0.77           | 0.68                   | 0.000     | 0.36        |
| Misconceptions | Females (N=779) | Males (N=1209) | p-value | Effect size | Dutch (N=1080) | International (N=899) | p-value | Effect size |
|----------------|-----------------|----------------|--------|-------------|----------------|------------------------|--------|-------------|
| MCtot: total Misconceptions | 0.33 | 0.30 | 0.000 | 0.27 | 0.29 | 0.33 | 0.000 | 0.34 |
| MC1: Misconceptions involving averages | 0.46 | 0.41 | 0.000 | 0.21 | 0.38 | 0.48 | 0.000 | 0.39 |
| MC2: Outcome orientation | 0.24 | 0.21 | 0.001 | 0.15 | 0.23 | 0.22 | 0.038 | 0.09 |
| MC3: Good samples have to represent high % of population | 0.17 | 0.14 | 0.004 | 0.13 | 0.15 | 0.15 | 0.566 | 0.03 |
| MC4: Law of small numbers | 0.33 | 0.25 | 0.000 | 0.29 | 0.24 | 0.31 | 0.000 | 0.24 |
| MC5: Representativeness misconception | 0.11 | 0.17 | 0.000 | 0.22 | 0.15 | 0.14 | 0.449 | 0.03 |
| MC6: Correlation implies causation | 0.24 | 0.26 | 0.201 | 0.06 | 0.19 | 0.30 | 0.000 | 0.33 |
| MC7: Equiprobability | 0.60 | 0.55 | 0.001 | 0.16 | 0.56 | 0.59 | 0.056 | 0.09 |
Outcomes in this and earlier studies are remarkably similar: Garfield (2003) e.g. reports as aggregate reasoning scores (CCtot) 0.56 and 0.60 for the U.S. and Taiwanese students, compared to 0.58 as the overall mean of CCtot in our study.

Similarity of our outcomes and those found in earlier studies is not limited to aggregated scores: also scale scores demonstrate very similar patterns. Of the correct reasoning scales, CC7 and CC8 are amongst those with highest mastery level, and CC3 and CC5 with lowest. Of the misconception scales, MC7 and MC8 are high in all studies (in our sample, MC8 somewhat less), and MC3, MC5 and MC6 are low.

In the Liu-study, reported in Garfield (1998b, 2003), Garfield and Chance (2000), and Liu (1998), the analysis of gender and country/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 the USA), and have significantly lower total misconceptions scores. These results can be generalized to our study with a remarkable regularity. We find significant gender effects in both aggregated scores in the same direction. Moreover, we find that CC2, CC3, CC5, and CC7 are significantly higher and MC1, MC3, MC4, MC7, and MC8 are significantly lower for males than for females among our students (where MC5 plays the role of the exception which proves the rule). All effects are quite strong in a statistical sense, having p-values below 0.005. The gender effect is rather substantial: males score more than 5% higher in total correct reasoning, and more than 9% lower in total misconceptions, than females, with Cohen’s d effect size ranging between small and medium. Performing an ANOVA indicates that no interaction effects are present in our data; p-values of the interaction effect for CCtot and MCtot are e.g. 0.247 and 0.875, respectively. For that reason, no further ANOVA results are incorporated in this and subsequent subsections.

Conceptions for which we find higher scores than reported in the Garfield-studies, CC2, CC6, and CC7, may be characterized as general reasoning skills more than as statistical reasoning skills; higher ‘European’ scores in general, and higher Dutch scores in particular, might simply reflect the general level of secondary education. Similar conclusions apply to the several misconception scales. We find high scores relative to the Garfield-reports for MC1, MC3, and MC6, all referring to topics that will be covered in any introductory course, so that the timing of the test administration might play a crucial role in explaining this difference (prior versus post assessment). In contrast, MC8 shows remarkably low misconception scores in our sample.

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. For both aggregated scales, Dutch students have an 11% higher total reasoning score, and a 9% lower misconception score, than non-Dutch students; effect sizes are in the range of medium. The nationality effect is about as stable as the gender effect, but much better
explainable: 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, which is rather uncommon in secondary school programs in other European countries, apparently provides Dutch students with a head start. Does this nationality effect possibly contribute to (part of) the gender effect? The answer is no; the female/male composition of Dutch and foreign student groups is very similar.

The second pattern refers to the high variability in prior mathematics education. Both Dutch students and students from most other European countries have taken mathematics in secondary school either as a major, or at advanced level, or alternatively as minor, or at basic level. Although the dummy ‘math major’ is a rather imprecise indicator of prior mathematics education, given the huge differences in mathematics programs in different European secondary school systems, it does contribute to the explanation of reasoning skills to a similar degree as nationality. Students with a math major have a 10.5% higher total correct reasoning score, and a 9.5% lower total misconception score, than students with a math minor. Apart from nationality, the math major dummy is a potential confounder explaining the gender puzzle since prior math education is somewhat biased, with 36% of the males versus 30% of the females having pursued a math major at high school level. However, the gender effect can only partially be contributed to differences in prior math education. After splitting the sample into two sub samples, corresponding to different levels of prior math education, most scales still demonstrate significant gender effects.

As a last observation on average levels of reasoning skills and misconceptions, the high rate of correct answers is noticeable. Of the eight correct reasoning skills, five have means of above 65% correct. Of the eight misconception scales, only two have means larger than 35%. Given the circumstance that only a minority of our inflow did attend any formal education in statistics in secondary school, and a majority did not, one might doubt whether the level of the instrument is appropriate for (European) high schools and what impact the restricted discriminative power might have on the reliability of the instrument.

Correlations between the several SRA scale scores are low, and in many cases not significant. For correct reasoning skills, they range between -0.17 and +0.14, and for misconceptions, from -0.29 to +0.14. This finding is in line with other studies, see Garfield (1998b), Garfield and Chance (2000), Liu (1998), and Garfield (2003). As a consequence, the Cronbach alpha reliabilities of the aggregated scales, taking the eight correct reasoning scales and the eight misconception scales as components, are low: 0.29 and 0.11, respectively, and the focus on aggregated scales has therefore certain drawbacks. We will not pursue the issue of the reliability of aggregated scales here further, but will instead refer to Tempelaar (2004a, b) for alternative representations of the reasoning skills scales that avoid the reliability problems of aggregate scales.

3.2 SRA and student performance: the ‘non-existing relation puzzle’

3.2.1 Descriptive data about SRA and student performance

In this subsection we will focus on one of the three shifts of students: the 99/00 shift. Data of other shifts demonstrate similar patterns, but are less rich, since they lack one course performance instrument: the assessment of homework. The assessment portfolio that measures students’ course performance in the 99/00 shift contains three instruments: final tests, graded home work assignments and quizzes, each for mathematics and statistics, and each for three different periods. Descriptive analysis of the performance indicators shows, first of all, that 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 in 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
assessment 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 both the quiz scores and bonuses achieved for homework assignments. This gender effect is present in mathematics and statistics, both for Dutch and international students, and always in the same direction: female students outperform male students. The effect is large, especially for the homework component. Third: there exists an even much stronger nationality effect in both performance indicators, where international students outperform Dutch students, both for mathematics and statistics, in all periods, both for females and males. Differences are again large.

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 and for international students. In the second and third statistics exam, this pattern tends to reverse, females scoring higher than males; differences are however not significant. The nationality effect in exam scores demonstrates a somewhat similar development. In the first exam, Dutch students do significantly better than international students, both in math, showing a very large difference, and in statistics. In the second exam, Dutch and foreign students approach each other in math, whilst international students significantly outperform their Dutch counterparts in statistics. Finally, in the third exam, international students outperform Dutch ones both for math and for statistics significantly.

Most of these apparent differences have natural explanations. First of all the match between secondary education and university study is much better for Dutch students than for international students. The counter veiling force, though, is that international 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 for homework already from the very first period onwards, and starts to pay off in the more cognitive based indicators in the second period. The picture for the gender issue is similar: female students are willing to spend more efforts on their study than male students. This pays off starting from the very first period onwards, 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.

3.2.2 SRA as predictor for performance indicators

What is the relationship between course performances and SRA scores, and how strong is this relationship? Regarding students’ reasoning abilities as a relevant part of their prior knowledge base when entering the course, one would expect that correct conceptions would positively contribute to performance indicators, whereas misconceptions do the reverse, given that prior knowledge is in general one of the better predictors of course performance. Table 4 contains the correlations between aggregated SRA scales and performance indicators.

Table 4: Correlations of SRA scales and course performance indicators and their two-sided $p$-values: Homework bonus, scores in quizzes and final exam (N=680)

| Performance indicator:                          | CCtot | $p$-value | MCtot | $p$-value |
|------------------------------------------------|-------|-----------|-------|-----------|
| Homework bonus: Statistics period 1            | -0.02 | 0.653     | 0.08  | 0.043     |
| Homework bonus: Statistics period 2            | -0.09 | 0.017     | 0.08  | 0.032     |
Performance indicators are ranked such that they start in Table 4 with the most ‘effort-based’ indicators, the bonus for the weekly homework assignments, through the weekly quizzes, and finish with the least effort-based but strongly cognitive oriented final exams. This design is advantageous, because striking differences between the three assessment categories evolve. 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 CCtot and performance indicators are positive and, although not very large, still substantial of size (up to 0.28). At the same time, all significant correlations with misconceptions are negative, but somewhat smaller in size. Weekly quizzes demonstrate a different pattern in that their relationship to SRA scales is absent. Going one step further into more effort-based indicators, the least intuitive result stems from the correlations between weekly homework bonus 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 explain why relationships between SRA scores and course performance can be weaker than the relationship between SRA scores and specific components of course performance. If the final course 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. Alternatively, if progress tests like quizzes or mid term exams contribute strongly to grades, again a condition is created in which dependencies with SRA scales remain hidden. It is only through the two extremes, traditional final exams focusing on the cognitive aspect on the one side, and scores for homework assignments on the other, that the impact of reasoning abilities and misconceptions becomes visible. In our analysis, we assume, as a working hypothesis, effort to be the mediating variable.

### 3.3 SRA and Attitudes toward Statistics

| Measure                  | Correlation | Significance | Other Correlation | Significance |
|--------------------------|-------------|--------------|-------------------|--------------|
| Homework bonus: Statistics period 3 | -0.13 | 0.001 | 0.10 | 0.006 |
| Homework bonus: Mathematics period 1 | -0.12 | 0.001 | 0.14 | 0.000 |
| Homework bonus: Mathematics period 2 | -0.14 | 0.000 | 0.10 | 0.009 |
| Homework bonus: Mathematics period 3 | -0.06 | 0.093 | 0.03 | 0.388 |
| Quiz score: Statistics period 1 | 0.01 | 0.792 | 0.00 | 0.922 |
| Quiz score: Statistics period 2 | -0.01 | 0.808 | 0.02 | 0.599 |
| Final exam: Statistics period 1 | 0.24 | 0.000 | -0.17 | 0.000 |
| Final exam: Statistics period 2 | 0.06 | 0.131 | -0.07 | 0.055 |
| Final exam: Statistics period 3 | 0.07 | 0.072 | -0.05 | 0.196 |
| Final exam: Mathematics period 1 | 0.28 | 0.000 | -0.18 | 0.000 |
| Final exam: Mathematics period 2 | 0.18 | 0.000 | -0.17 | 0.000 |
| Final exam: Mathematics period 3 | 0.13 | 0.001 | -0.17 | 0.000 |
Except for Difficulty, students express positive attitudes towards statistics. This is true for all relevant subgroups of students; see Table 5. In contrast, mean scores for Difficulty are below the neutral level, expressing that students perceive the subject as difficult (the naming of the Difficulty-scale is somewhat counterintuitive: 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).

Table 5: Average scores for SATS scales Affect, Cognitive Competence, Value and Difficulty, and the added scales Interest and Effort, and corresponding gender- and nationality effects, expressed by p-values and effect sizes

| SATS Scales: | Females (N=822) | Males (N=1290) | p-value | Effect size | Dutch (N=987) | International (N=1060) | p-value | Effect size |
|-------------|-----------------|----------------|---------|-------------|---------------|------------------------|---------|-------------|
| Affect      | 4.35            | 4.63           | 0.000   | 0.28        | 4.70          | 4.37                   | 0.000   | 0.33        |
| Cognitive Competence | 4.79        | 5.06           | 0.000   | 0.33        | 4.93          | 4.99                   | 0.123   | 0.07        |
| Value       | 5.01            | 5.01           | 0.848   | 0.01        | 4.97          | 5.07                   | 0.005   | 0.12        |
| Difficulty  | 3.51            | 3.66           | 0.000   | 0.20        | 3.76          | 3.46                   | 0.000   | 0.42        |

| Added SATS Scales 2004: | Females (N=276) | Males (N=439) | Dutch (N=287) | International (N=428) |
|------------------------|-----------------|---------------|---------------|-----------------------|
| Interest               | 5.27            | 5.05          | 4.94          | 5.27                  |
| Effort                 | 6.55            | 6.24          | 6.08          | 6.55                  |

Table 5 indicates that both gender and nationality effects are present. Male students have significantly higher scores in Affect, Cognitive Competence and Difficulty, but significantly lower scores in Interest and Effort, than female students (all p-values being less than 0.005, and effect sizes ranging from small to medium); for Value, no significant difference exists. In comparing Dutch and international students, Dutch students express significantly higher Affect and Difficulty than international students, but lower Value, Interest and Effort; Cognitive Competence is invariant across nationalities (again at 0.005 level, with effect sizes ranging from medium to large). Attitude scores of our students are comparable to those reported in other studies; Schau (2003) e.g. reports pre-test scores for Affect, Cognitive Competence, Value, and Difficulty of 4.03, 4.91, 4.86, and 3.62, respectively.

Do attitudes as measured by SATS have any impact on students’ state of reasoning abilities? If so, we expect this impact to be positive for the reasoning abilities, and negative for the misconceptions. The SATS instrument is based on the expectancy-value model of behavior, developed by Eccles and her colleagues (see, for example, Wigfield and Eccles 2000, 2002; Eccles and Wigfield 2002). According to this theory of achievement motivation, students’ expectancies for success and the value they contribute to succeeding are important determinants of their motivation to perform achievement tasks. Expectation of success includes two components: belief about one’s own ability in performing a task (the SATS scale Cognitive Competence), and a perception of the task demand (Difficulty). From empirical research, these two aspects of success expectation are known to be positively related to the student’s (prior) knowledge.
state (Wigfield and Eccles 2000, 2002; Eccles and Wigfield 2002). Therefore the expectation of positive correlations with reasoning, and negative with misconceptions, is most explicit for these two affects. These expectations turn out to be true, with the exception of the recently introduced variables Interest and Effort, as can be seen in the correlation matrix of Table 6.

Table 6: Correlations between SRA and SATS scales and their two-sided $p$-values (N=2031 for first four scales, N=687 for last two scales)

| Affect          | CCtot | $p$-value | MCtot | $p$-value |
|-----------------|-------|-----------|-------|-----------|
| Cognitive Competence | 0.12  | 0.000     | -0.07 | 0.005     |
| Value           | 0.12  | 0.000     | -0.06 | 0.012     |
| Difficulty      | 0.11  | 0.000     | -0.10 | 0.000     |
| Interest        | 0.02  | 0.533     | 0.05  | 0.174     |
| Effort          | -0.07 | 0.058     | 0.17  | 0.000     |

Although most correlations are very strongly significant, their size is moderate to small. In a joint analysis, SATS variables explain 2.2% of the variation in correct reasoning, and 4.5% in variation of misconception scores. However, the size of the gender effect is smaller, and since SATS variables are gender biased, the possibility of a gender effect induced by differences in attitudes is open.

By far the strongest correlation is the one between total MisConceptions and planned Effort in learning. This correlation is positive, a fact that contradicts the general hypothesis that positive attitudes will contribute to higher reasoning abilities and lower misconceptions levels, but it corroborates our working hypothesis formulated in the last section: learning approaches, characterized by investing large efforts, might result in inferior learning outcomes. However, there is another mechanism that has the potential to explain a positive relationship between the misconception level and planned effort: students realizing their deficient prior knowledge, might plan to compensate by spending above average efforts on their study. For this mechanism to apply, one would require a negative relationship between planned effort and prior knowledge. In our sample, we have three measurements that can be used to indicate prior knowledge: the SRA total reasoning abilities score, the students’ score in math in the national exam (only for Dutch students), and most relevant, the students’ self-scored Cognitive Competence in the SATS instrument. Table 6 indicates that the correlation between SRA total reasoning score CCtot and Effort is absent. The same is true for the correlation between grade for the national exam, and planned effort. The third correlation, between Effort and Cognitive Competence, is significant, but its sign is opposite to what a compensation mechanism would predict. Higher self-concept is associated with higher planned efforts, thereby making the existence of a compensating mechanism very improbable, and in stead favoring the hypothesis of inadequate learning approaches.

3.4 SRA and Student learning approaches

Analyzing the relationship between SRA and ILS produces correlations that are in line with other research
into the relationship between learning approaches and course performance. Several significant correlations exist, but the size of them is restricted, typically being no larger than 0.1 (see e.g. Duff, et al. 2004). Deep processing typically contributes to better course performance, surface processing to inferior course performance. This pattern is also visible in our data on students’ reasoning abilities: the deep processing component ‘Critical processing’ correlates positively to SRA correct reasoning and negatively to SRA misconceptions. The reverse is true for the surface component Analyzing. Table 7 contains all correlations and their significance levels.

| ILS scale                           | CCtot | p-value | MCtot | p-value |
|-------------------------------------|-------|---------|-------|---------|
| Relating and structuring            | 0.03  | 0.242   | 0.02  | 0.366   |
| Critical processing                 | 0.10  | 0.000   | -0.10 | 0.000   |
| Memorizing and rehearsing           | -0.02 | 0.466   | 0.01  | 0.687   |
| Analyzing                           | -0.06 | 0.011   | 0.04  | 0.057   |
| Concrete processing                 | -0.01 | 0.808   | -0.01 | 0.799   |
| Self-regulation of learning processes | 0.00  | 0.894   | 0.00  | 0.832   |
| Self-regulation of learning content | 0.00  | 0.981   | -0.07 | 0.005   |
| External regulation of learning processes | -0.03 | 0.162   | 0.07  | 0.003   |
| External regulation of learning results | 0.00  | 0.977   | 0.04  | 0.058   |
| Lack of regulation                  | -0.01 | 0.806   | -0.04 | 0.107   |
| Personally interested               | -0.05 | 0.057   | 0.01  | 0.697   |
| Certificate directed                | -0.02 | 0.483   | 0.05  | 0.037   |
| Self test directed                  | -0.04 | 0.064   | 0.11  | 0.000   |
| Vocation directed                   | -0.04 | 0.125   | 0.12  | 0.000   |
| Ambivalent                          | -0.06 | 0.016   | -0.02 | 0.355   |
| Construction of knowledge           | -0.08 | 0.000   | 0.09  | 0.000   |
| Intake of knowledge                 | -0.09 | 0.000   | 0.15  | 0.000   |
| Use of knowledge                    | -0.05 | 0.034   | 0.11  | 0.000   |
| Stimulating education               | -0.02 | 0.336   | 0.07  | 0.004   |
| Co-operation                        | -0.08 | 0.001   | 0.06  | 0.007   |
The largest numbers of significant correlations are found amongst the last five scales, the mental models of learning. All these scores are positively related to the level of misconceptions, MC\textsubscript{tot}, and negatively associated with the level of correct conceptions, CC\textsubscript{tot} (except for Stimulating education). This finding deviates somewhat from the deep versus surface learning hypothesis; according to that hypothesis, one would expect that Construction of knowledge contributes to reasoning abilities, whereas Intake of knowledge would hinder it. From Table 7, one is inclined to adopt a different kind of mechanism; students with very outspoken mental models of learning (scoring high on one or two of the individual scales) tend to do worse in terms of reasoning abilities than students without outspoken mental models of learning who combine all or most of the individual models without being strongly dependent on any of them. A similar conclusion can be drawn from learning orientations, although the effect is weaker and restricted to Misconceptions. The learning orientations Vocation directed and Self test directed contribute positively to the MC\textsubscript{tot} score, as does Certificate directed, but with lower significance. This is to be interpreted that a unidirectional learning orientation puts a student at a disadvantage in terms of misconceptions. Self-regulation of learning content and external regulation of the learning process have a, be it very modest, impact on Misconceptions of expected direction: students who do a better job in regulating their study themselves, achieve lower misconception scores. A similar impact on correct reasoning is absent. In general, it is noticeable that the strongest relations are those between learning approaches and misconceptions rather than between learning approaches and correct conceptions. In a joint analysis, learning approaches explain 5.1% of variation in MC\textsubscript{tot}, against 4.1% of variation in CC\textsubscript{tot}.

Can learning approach contribute to the explanation of the gender puzzle, and the corroboration of our effort hypothesis? The answer to both questions is affirmative. Correlations in Table 8 demonstrate that Effort is positively correlated with all four components of deep and surface processing. The strongest correlation is to be found between Analyzing and the SATS scale, where analyzing is the surface processing component correlated with Misconceptions. The weakest correlation score can be observed for Critical processing, the deep processing component positively correlated with Correct reasoning. Finally, Effort is strongly positively correlated with all five learning orientations, each in their turn correlated with the Misconception score. Other attitudes are also correlated to learning approaches but, except for the two deep processing scales, these correlations are much weaker than those of the Effort scale.

Table 8: Correlations between selected ILS scales and SATS scale effort, and their two-sided \(p\)-values (\(N=675\))

| ILS scale                | SATS Effort | \(p\)-value |
|--------------------------|-------------|-------------|
| Relating and structuring | 0.22        | 0.000       |
| Critical processing      | 0.12        | 0.002       |
| Memorizing and rehearsing| 0.19        | 0.000       |
| Analyzing                | 0.29        | 0.000       |
| Construction of knowledge| 0.33        | 0.000       |
| Intake of knowledge      | 0.22        | 0.000       |
| Use of knowledge         | 0.31        | 0.000       |
| Simulating education     | 0.17        | 0.000       |
With regard to the gender effect, Table 9 contains the outcomes of tests on differences of means for the relevant ILS scales. The pattern is identical to that of Table 8: significant negative gender effects exist in scales that correlate strongly with the SATS Effort variable (Analyzing and all mental models of learning). In contrast, the deep learning component that correlates most weakly with Effort, Critical processing, demonstrates the only significant positive gender effect.

Table 9: Gender effect (mean difference of males to females) in selected ILS scales, and two-sided p-values in an independent samples $t$-test (N=1215, 799 for males, females)

| ILS scale                  | SATS Effort | p-value |
|----------------------------|-------------|---------|
| Relating and structuring   | -0.008      | 0.810   |
| Critical processing        | 0.100       | 0.001   |
| Memorizing and rehearsing  | 0.013       | 0.659   |
| Analyzing                 | -0.060      | 0.020   |
| Construction of knowledge  | -0.177      | 0.021   |
| Intake of knowledge        | -0.142      | 0.021   |
| Use of knowledge           | -0.142      | 0.024   |
| Simulating education       | -0.131      | 0.024   |
| Co-operation               | -0.169      | 0.025   |

Motivated by the important role Effort appears to play in the growth of correct and incorrect statistical conceptions, we investigated in this section the relation between learning approaches and SATS scores. The global picture that emerges is that Critical processing, being an important constituent of the meaning-directed learning pattern, has a positive impact on statistical reasoning, whereas Analyzing, a constituent of the reproduction-directed learning pattern, has a negative impact. In addition, an outspoken mental model of learning and an outspoken learning orientation have negative impacts on statistical reasoning, whereas a more balanced mental model of learning and learning orientation contributes to better statistical reasoning. Since all these learning approach components appear to be gendered in our sample, they help explain the gender puzzle in statistical reasoning.

3.5 Final model and conclusions

Integrating the partial models of statistical reasoning explained by attitudes as well as learning approaches – including dummies for gender and nationality – generates regression equations as described in Table 10.
Table 10. Standardized regression coefficients ($\beta$), significance levels, and $t$-values of regression models for SRA Correct Reasoning scores and MisConceptions scores of the complete models (Method Enter) and reduced models (SPSS Method Stepwise with entry significance level 5%, removal significance level 10%)

* significant at 10%; ** significant at 5%; *** significant at 1%; N=1466

| Explanatory variables | CCtot: total Correct Reasoning score | Method Enter | Method Stepwise | MCtot: total MisConceptions score | Method Enter | Method Stepwise |
|-----------------------|-------------------------------------|--------------|-----------------|----------------------------------|--------------|-----------------|
|                       | $\beta$ | $t$ | $\beta$ | $t$ | $\beta$ | $t$ | $\beta$ | $t$ | $\beta$ | $t$ |
| Constant              |         |     |         |     |         |     |         |     |         |     |
| Nationality (dummy for Dutch students) | 0.205*** | 7.092 | 0.203*** | 7.815 | -0.130*** | 4.496 | -0.120*** | 4.462 |
| Gender (dummy for female students) | -0.081*** | -0.130*** | -0.083*** | -3.227 | 0.098*** | 3.667 | 0.099*** | 3.863 |
| Relating and structuring | 0.027 | 0.712 | -0.034 | -0.892 |
| Critical processing | 0.123*** | 3.836 | 0.125*** | 4.520 | -0.092*** | 2.869 | -0.113*** | 4.352 |
| Memorizing and rehearsing | -0.024 | -0.789 | -0.024 | -0.805 |
| Analyzing | -0.060* | -1.857 | -0.066* | -2.382 | 0.008 | 0.233 |
| Concrete processing | -0.015 | -0.418 | -0.043 | -1.211 |
| Self-regulation of learning processes | 0.001 | 0.037 | 0.023 | 0.660 |
| Self-regulation of learning content | -0.013 | -0.408 | 0.011 | 0.343 |
| External regulation of learning processes | 0.015 | 0.471 | 0.026 | 0.813 |
| External regulation of learning results | -0.002 | -0.068 | 0.006 | 0.169 |
Explained variation in the two regression equations achieved by stepwise regression is 8.5%, and 7.8%, respectively for Correct reasoning and MisConceptions. Nationality and gender dummies only explain 5.8% respectively 4.3% of variation, so adding both attitudes and learning approaches as predictors has a significant, but restricted effect on explained variation. The best predictor of Correct reasoning is the nationality dummy, contributing about half of all explained variation, followed by the learning approach Critical processing. The Gender dummy is significant, but has a very restricted impact: it explains less than 1%. Conclusions for the SRA MisConceptions variable are similar: nationality dummy and Critical
processing are the main regressors, gender is significant with an impact stronger than in the correct reasoning case, but still limited. Reducing the sample to the 03/04 shift to allow the SATS variable Effort into the model has no effect in the first equation explaining Correct reasoning. It, however, has an impact on the second equation: Effort enters the equation replacing the gender dummy completely.

What can we conclude from this final model with regard to the research question of this study? First of all, there exists a solid nationality effect in both Correct conceptions and Misconceptions that overshadows all other effects. Although the nationality effect, in principal, can consist of several elements, the large differences between secondary school systems in European countries and the prominent role of statistics in the math program of Dutch high schools suggest that this effect is mainly caused by differences in prior schooling. The fact that the nationality effect is stronger in Correct reasoning than in MisConceptions, reinforces the plausibility of a schooling effect. Beyond the nationality effect, there exists a gender effect. However, SRA scales are not the only gendered phenomena relevant in statistics education; also attitudes toward statistics, as measured by SATS, and preferred learning approaches, as measured by ILS, demonstrate gendered components. For that reason, the greatest part of the gender effect in SRA (but not all of it) can be explained by differences in learning approaches and attitudes. Students with a reproduction directed learning pattern and unilateral learning orientations as well as mental models of learning are outperformed in statistical reasoning by students with a meaning-directed learning pattern along with balanced learning orientations and mental models of learning. Since female students are overrepresented in the first category (at least in our sample), a gender puzzle is created. A gender puzzle that arises most prominently in the MisConceptions part.

In doing this study, both puzzles seem to be -at least for the greatest part- resolved. What is left is the question of why statistical reasoning behaves so differently from other academic subjects, including mathematics and statistics. When confronted with a learning task, students will decide upon their preferred approach toward that task. That choice is first of all context dependent: students choose different learning approaches for different learning tasks. It is also student dependent: some students have ‘on average’ a stronger tendency to use surface approaches, others a stronger tendency to apply deep approaches. Empirical research in learning approaches generally indicates that although students with a stronger emphasis on deep approaches are somewhat more successful than students who emphasize surface approaches, approaches are best regarded as substitutes. There are different ways to reach the same goal, one maybe more efficient than the other, but in the end all resulting in mastery. The strong correlations found in this study between the several types of course performance, both for mathematics and for statistics, confirm this perspective. In this rather general pattern, statistical reasoning makes the exception. It’s negative correlations with effort, and with several of the scales of the learning styles instrument, suggest that statistical reasoning calls for a unique learning approach, excluding alternative ways to mastery. ‘Trying harder’ has not many, but at least one limitation.

4. 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, such as final exams. In general, correlations between course outcomes as assessed by these different instruments tend to be rather high. 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 types of course performance, 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, though: 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 reproduction directed and strongly effort-based might be an obstacle in developing statistical reasoning. If this interpretation is correct, it will have a strong impact on statistics education. The assessment portfolio relevant for this study demonstrates a wide range of instruments: from multiple choice final exams, via quizzes, to assessed home work. Still, for all these different instruments, both deep and surface learning approaches contribute to achieving satisfying outcomes. So, although an effort focused learning approach might be not the most efficient to pass the course, it still carries the guarantee for success, as long as effort levels are high enough. If the SRA were to be added to the portfolio of assessment instruments, this story would become different. If our sample is representative, and if the characteristics of the SRA as post-test are similar to those of a pre-test, we cannot but conclude that there are no alternative routes toward achieving reasoning abilities.

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 to that that also specific aspects of the quality of learning, such as approaching learning tasks in a committed but reproduction directed way, do not guarantee proper reasoning skills. Chance (2002) describes several instructional tools that allow ‘thinking beyond the textbook’. The outcomes of this study emphasize the importance of using those types of activities and other tools discussed by Chance; neither traditional lecturing, nor textbook-based independent learning, can assure success. The study at the same time indicates what those tools should do beyond teaching some specific skills or knowledge: strengthen e.g. critical processing, and create a better balance in learning orientations and mental models of learning, since these are important in achieving statistical reasoning skills.

Acknowledgements

The authors would like to thank the editor and two anonymous referees for their valuable comments on an earlier version of this paper that led to an improved final version. Earlier versions of this paper appear in the Proceedings of the ARTIST 2004 Roundtable Conference on Assessment in Statistics and the American Statistical Association 2004 Proceedings of the Section on Statistical Education.

References

Ben-Zvi, D., and Garfield, J., eds. (2004), The challenge of developing statistical literacy, reasoning, and thinking, Dordrecht, the Netherlands: Kluwer Academic Publishers.

Biggs, J. (2003), Teaching for Quality Learning at University, 2nd Ed., Buckingham: Society for Research into Higher Education / Open University Press.

Bransford, J. D., Brown, A. L., and Rodney, R. C. (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, B. L. (2002), “Components of Statistical Thinking and Implications for Instruction and Assessment,” *Journal of Statistics Education* [Online], 10(3).
ww2.amstat.org/publications/jse/v10n3/chance.html

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences*, 2nd Ed., Hillsdale, NJ: Lawrence Erlbaum Associates.

Dauphinee, T. L., Schau, C., and Stevens, J. 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, R. C. (2002a), “Statistical Literacy, Reasoning, and Learning,” *Journal of Statistics Education* [Online], 10(3).
ww2.amstat.org/publications/jse/v10n3/delmas_intro.html

delMas, R. C. (2002b), “Statistical Literacy, Reasoning, and Learning: A Commentary,” *Journal of Statistics Education* [Online], 10(3).
ww2.amstat.org/publications/jse/v10n3/delmas_discussion.html

Duff, A., Boyle, E., Dunleavy, K., and Ferguson, J. (2004), “The relationship between personality, approach to learning and academic performance,” *Personality and Individual Differences*, 36, 1907-1920.

Eccles, J.S., and Wigfield, A. (2002), “Motivational Beliefs, Values, and Goals,” *Annual review of psychology*, 53, 109-132.

Gal, I. and Garfield, J. (1997), “Curricular Goals and Assessment Challenges in Statistics Education,” In: Gal, I. and Garfield, J., *The Assessment Challenge in Statistical Education*, Voorburg: IOS Press.

Gal, I. and Ginsburg, L. (1994), “The Role of Beliefs and Attitudes in Learning Statistics: Towards an Assessment Framework,” *Journal of Statistics Education* [Online], 2(2).
ww2.amstat.org/publications/jse/v2n2/gal.html

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

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

Garfield, J. (2002), “The Challenge of Developing Statistical Reasoning,” *Journal of Statistics Education* [Online], 10(3).
ww2.amstat.org/publications/jse/v10n3/garfield.html

Garfield, J. (2003), “Assessing Statistical Reasoning,” *Statistics Education Research Journal* [Online], 2 (1), 22-38.
http://www.stat.auckland.ac.nz/~iase/serv/SERJ2(1).pdf
Garfield, J., and 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 Ben-Zvi, D. (2004), “Research on statistical literacy, reasoning, and thinking: issues, challenges, and implications,” In D. Ben-Zvi & J. Garfield (Eds.), The challenge of developing statistical literacy, reasoning, and thinking, 397-409, Dordrecht, the Netherlands: Kluwer Academic Publishers.

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

Hilton, S. C., Schau, C., and Olsen, J. A. (2004), “Survey of Attitudes Toward Statistics: Factor Structure Invariance by Gender and by Administration Time,” Structural Equation Modeling, 11 (1), 92-109.

Jolliffe, F. (1997), “Issues in constructing assessment instruments for the classroom,” in The Assessment Challenge in Statistical Education, eds. J. Garfield and I. Gal, Voorburg: IOS Press.

Kahneman, D., Slovic, P. and 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.

Lecoultre, 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.

Nasser, F.M. (2004), “Structural Model of the Effects of Cognitive and Affective Factors on the Achievement of Arabic-Speaking Pre-service Teachers in Introductory Statistics,” Journal of Statistics Education [Online], 12(1).

Schau, C. (2003), Students’ attitudes: the “other” important outcome in statistics education,” Paper presented in the Joint Statistical Meetings, San Francisco, CA.

Schau, C., Stevens, J., Dauphinee, T. L., and Vecchio, A. 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:
Tempelaar, D. (2004a), “Statistical Reasoning Assessment: an Analysis of the SRA Instrument,” *Proceedings of the ARTIST Roundtable Conference on Assessment in Statistics.*

Tempelaar, D. (2004b), “Statistical Reasoning Assessment: an Analysis of the SRA Instrument,” in *2004 ASA Proceedings of the Joint Statistical Meetings*, pp. 2797-2804, Alexandria, VA: American Statistical Association.

Vermunt, J. D. and Vermetten, Y. J. (2004), “Patterns in Student Learning: Relationships Between Learning Strategies, Conceptions of Learning, and Learning Orientations,” *Educational Psychology Review, 16*(4), 359-384.

Wigfield, A., and Eccles, J.S. (2000), “Expectancy - Value Theory of Achievement Motivation,” *Contemporary Educational Psychology, 25*(1), 68-81.

Wigfield, A., and Eccles, J.S. (2002), “The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence,” In: *Development of Achievement Motivation*, Wigfield, A., and Eccles, J.S. (eds.), San Diego: Academic Press.

Dirk T. Tempelaar  
Faculty of Economics and Business Administration  
Department of Quantitative Economics  
Maastricht University  
Maastricht  
Netherlands  
*D.Tempelaar@KE.UNIMAAS.NL*

Wim H. Gijselaers  
Faculty of Economics and Business Administration  
Department of Educational Development and Research  
Maastricht University  
Maastricht  
Netherlands  
*W.Gijselaers@Educ.UNIMAAS.NL*

Sybrand Schim van der Loeff  
Faculty of Economics and Business Administration  
Department of Quantitative Economics  
Maastricht University  
Maastricht  
Netherlands  
*S.Loeff@KE.UNIMAAS.NL*