Assessing Knowledge Structures in a Constructive Statistical Learning Environment

P. P. J. L. Verkoeijen
Erasmus University of Rotterdam

Tj. Imbos
Maastricht University

M. W. J. van de Wiel
Maastricht University

M. P. F. Berger
Maastricht University

H. G. Schmidt
Erasmus University of Rotterdam

Journal of Statistics Education Volume 10, Number 2 (2002),
www.amstat.org/publications/jse/v10n2/verkoeijen.html

Key Words: Active learning; Constructive learning environment; Free recall; Knowledge representation.

Abstract

In this report, the method of free recall is put forward as a tool to evaluate a prototypical statistical learning environment. A number of students from the faculty of Health Sciences, Maastricht University, the Netherlands, were required to write down whatever they could remember of a statistics course in which they had participated. By means of examining the free recall protocols of the participants, insight can be obtained into the

mental representations they had formed with respect to three statistical concepts. Quantitative as well as qualitative analyses of the free recall protocols showed that the effect of the constructive learning environment was not in line with the expectations. Despite small-group discussions on the statistical concepts, students appeared to have disappointingly low levels of conceptual understanding.

1. Introduction

In the past two decades statistics educators have grown increasingly aware that college-level statistics courses are in need of restructuring. Hogg (1992) gives his impression of the state of statistics education at that specific moment in time. One of Hogg's main objections concerns the passive role that was assigned to students in traditional statistical curricula. Instead of a learning environment dominated by exposure teaching, in which students are required to remember the information handed out to them by a teacher, the statistical learning environment should encourage students to actively construct their knowledge.

Many researchers (including Shuell, 1986; Kilpatrick 1987; Kintsch 1988; de Corte 1990; Wheatley 1991; Cobb 1994; Von Glasersfeld 1995) have stressed the constructive nature of learning. It is assumed that learners bring into the learning environment their own distinct body of knowledge. Upon confrontation with newly offered information, learners activate relevant pieces of prior knowledge, in which the recently provided information can be embedded. The integration process requires a considerable effort on the part of the learner; not only should pieces of new information be linked to already existing knowledge structures, but previously held erroneous conceptions should be erased and replaced with more appropriate ones. Eventually, this process of active learning will lead to the development of sophisticated and elaborate mental knowledge structures.

The constructive nature of learning should be taken into consideration while designing a statistical learning environment. It is essential that statistics teachers create courses that encourage constructive activities within students. One way of achieving this goal is by providing students with small-group learning opportunities. Garfield (1993) recommended small-group learning as an optimal learning environment for the teaching of statistics. In small-group learning, students work actively together to solve a problem, complete a task, or accomplish a common goal. According to Garfield the characteristics of small group social interaction are peer teaching, the exchange and the critical evaluation of different opinions on the problem solution, and a strong bond between the group members. These factors are all assumed to contribute to the active construction of sophisticated knowledge structures.

The hypothesis that small-group learning has beneficial effects on the teaching of college statistics courses was experimentally tested (Giraud 1997). A total number of 95 students enrolled in an introductory statistics course participated in the experiment. Classes were assigned to either an experimental (N = 44) or to a conventional lecture-based class (N = 51). In the experimental condition, students were randomly assigned to small-learning groups of five students, hence creating heterogeneous groups in terms of statistical

abilities. This was assumed to evoke the opportunity for scaffolding among the group-members. The experiment took place over a complete one-semester course. During the experiment both the small group and the lecture class met twice a week for 75 minutes. The students in the small-group condition discussed the relevant statistical concepts. The students in the lecture condition, on the other hand, attended lectures on the same topics, their input being constrained to questions and brief statements. Students in both conditions were required to hand in practical assignments. Students in the lecture classes worked on the projects individually, whereas the students in the small-group classes cooperatively worked on the assignments. At the end of the experiment, both groups were compared in terms of the achievement on a final test, consisting of 27 multiple-choice items and one constructed response item. The analysis of the test scores showed that those with low pretest scores especially benefited from small-group learning.

Magel (1998) also successfully introduced small-group learning in an introductory statistics course. A total number of 195 students participated in a one-semester course. For several class sessions during the course, students were randomly divided into learning teams consisting of up to five people. Members of the learning teams were instructed to cooperatively solve a practical assignment. The purpose of the small-group activities was to provide students with self-generated, concrete elaboration on a number of statistical concepts (such as random variables, probability distributions and the mean and variance of a random variable) which had been explicated in courses prior to the experimental course. On the final test, the overall performance of the students was reported to be slightly better than the overall achievement of the students who had been taking the course one year before the introduction of small-group learning.

Roberts (1992), Keeler and Steinhorst (1995), and Smith (1998) all report positive effects derived from project-driven, small-group activities. Typically, the innovative style of teaching increased the overall achievement on the final examination.

The aforementioned studies, despite their substantial contribution to the innovation of the statistical learning environment, were largely exploratory in nature and as a consequence they are not entirely flawless. In most of the studies, the innovative learning environment was assessed immediately after its first introduction. The motivations of the students to actively engage in the learning of statistics were enhanced upon the introduction of a non-conventional statistical learning environment, leading to an increased performance on the final examination. Thus, the reported positive effects of an innovative small-group learning environment might be attributed to its novelty, rather than to educational theory. In order to obtain a more valid account of the effect of a small-group learning environment, it would be preferable to postpone the assessment of the learning environment until students have become familiar with the new learning environment.

Another important, point of criticism lies in the coarse measurement used to determine the effects of the small-group statistical learning environment. Although the scores on a knowledge test provide us with insight into the effects of a learning environment, this measurement does not reveal at a detailed level the mental representations of statistical knowledge students have constructed. Consider a course on the topic of analysis of

variance that is conducted within an innovative learning environment. Ideally, conceptual understanding of the topic would result, as reflected in the structure of the knowledge representations. Knowledge representations of students who have acquired conceptual understanding of statistics may be assumed to be coherent structures incorporating statistical terms, formulas, arithmetic procedures, the conditions for application of the learned knowledge, interpretations of the outcomes of mathematical calculations, and theoretical background knowledge. In contrast, the knowledge representations of students who have failed to achieve conceptual understanding are less complete and may contain misconceptions. Ideally, the assessment of curricular restructuring involves a measure that taps directly into these knowledge representations. Unfortunately, the test scores, and in particular the scores on a multiple-choice test, do not meet this criterion because through such a test the content of the relevant knowledge representations can only be derived in an indirect manner.

In the next section the method of free recall will be put forward as a valuable evaluation tool. Through this method, more direct information about the quantitative and the qualitative structure of a mental representation of meaningful information can be obtained.

2. Free Recall as an Evaluation Tool of Knowledge Structures

In a study conducted by McNamara, Kintsch, Songer-Butler and Kintsch (1996) a 683-word text explaining the mechanism underlying a heart disease was given to 6th- and 8th-grade students. On the basis of a pre-test assessing the participants' biological knowledge, the participants were divided into high- and low-knowledge groups. After they had read the text, participants were required to write down whatever they could remember of the text they had just read. The analysis of the recall protocols revealed that high-knowledge participants produced more extensive recall protocols than low-knowledge participants. The explanation of this finding is fairly straightforward and completely in line with important theories on text processing (see, for example, Kintsch and van Dijk 1978; Kintsch 1988). Upon confrontation with a meaningful text people construct a coherent mental representation by means of trying to integrate the presented information with relevant prior knowledge. Because high-knowledge readers have acquired more domain-specific knowledge than low-knowledge readers, they are conjectured to construct more elaborate text representations and hence to produce more extensive recall protocols. Results similar to those obtained by McNamara et al. (1996), were found in other text studies as well (see Ausubel and Youssef 1963; Chiesi, Spilich, and Voss 1979; Spilich, Vesonder, Chiesi, and Voss 1979; McNamara and Kintsch 1996). From these studies it can be deduced that free recall protocols may reflect the quantitative nature of a mental representation.

Furthermore, free recall protocols provide information about the qualitative structure of a mental representation. To illuminate this point it might be useful to consider studies on medical problem solving (see Schmidt and Boshuizen 1993; Van de Wiel, Schmidt, and

Boshuizen 2000). In these studies, participants of different levels of medical expertise were presented with a written version of a clinical case, describing the signs and symptoms displayed by the patient. After reading the text, participants were required to formulate a diagnosis. Subsequently, they are asked to provide an explanation of the pathophysiology underlying the presented case. In order to arrive at this explanation, participants had to revert to their mental case representation. Analyses of the recall protocols consistently showed that experts applied clinical knowledge in explaining the case whereas non-experts mainly used biomedical knowledge. In addition, the protocols of experts were more condensed than those of less experienced participants. The experimenters also compared the recall protocols to a model explanation of the case. This model explanation can be considered as an evaluation standard that consists of a minimal but sufficient set of biomedical and clinical knowledge, which causally explains all signs and symptoms in the case (Van de Wiel, et al. 2000). The model explanation reflects the case representation that the participants should have constructed. Typically, the recall protocols of medical experts had more concepts in common with the concepts that were presented in the model explanation. Thus, experts produced less elaborate but qualitatively superior case representations.

3. Free Recall Protocols in Statistics Education

The purpose of this report is to introduce the free recall method to evaluate the effects of statistics education in a constructive statistical learning environment at the faculty of Health Sciences, Maastricht University. The reframing of the statistical learning environment at the faculty of Health Sciences started at the beginning of the 1990's. In order to encourage active and constructive learning within the students, an emphasis was put on the interpretation of data from real-life, such as from a health science, problem-solving context. The intervention was in line with the recommendations put forward in previous articles on the reform of statistical education (see Lock and Moore 1992; Scheaffer 1992; Tanner and Wardrop 1992). In accordance with the problem-based-learning system implemented at all faculties of Maastricht University, statistics courses were designed in such a way that the core learning activities took place in small collaborative groups. From this perspective, the learning environment at the faculty of Health Sciences is largely compatible with other small group statistical learning environments such as those described by Giraud (1997) and Magel (1998). In the following section the statistical learning environment will be explicated.

Statistics education in the Health Sciences curriculum comprises a number of statistics courses each covering a set of related statistical concepts in a four week instructional cycle. A cycle starts with an introductory lecture on concept to be covered (for instance, analysis of variance) in which students are provided with an outline of the important aspects of the concept. After the lecture, one week is reserved for individual study of relevant chapters from the course book. In the second week, the students meet in a two hour tutorial group to discuss the studied literature under the guidance of a tutor. The tutor is either a staff member of the Department of Methodology and Statistics, or an advanced undergraduate student. Typically, the tutor initiates the discussion by prompting the students to collaboratively generate a summary of the concept under study.

As the summary is being constructed, poorly understood aspects of the concept are quickly identified and the group attempts to clarify these aspects. The tutor fulfills a monitoring role and does not intervene in the group process unless this is strictly necessary. For instance, if the group members have failed to mention an important aspect of the concept the tutor gives a hint to provoke a discussion about this aspect. Furthermore, when the group members do not succeed in the clarification of an issue, the tutor provides extra support by means of questioning and eventually explaining. At the end of the meeting, practical assignments are handed out to the students. Students are given one week time to use SPSS® 8.0 to solve individually a set of problems usually based on real-life data sets. In the third week, students meet again with their tutorial groups in order to discuss the solutions to the practical problems. The solutions usually take the form of relevant SPSS® output. During the two-hour group meeting all problems are dealt with in a sequential order. The students largely control the discussion of the problems. The tutor will only intervene if students are not capable of handling a complication that arises by themselves. In such a situation, the tutor will ask the students relevant questions in order to let them elicit the statistical concepts, which are needed to solve the problem at hand. Finally, in the fourth week the cycle ends with a lecture. The purpose of this lecture is to provide students with the opportunity to pose questions and to get some additional explication on aspects of the subject that remain unclear.

4. Method

Participants

Participants were 107 first year health science students, who took part in an introductory statistics instructional cycle focusing on the basic principles of statistical inference. The students taking part in the study formed 27% of the whole population, meaning that 291 students did not participate in the study. In order to test the comparability of the study group and the non-study group in terms of statistical competence, the mean scores of the two groups were compared on a test administered in a previous statistics course. A total of 101 out of 107 students participating in the study had taken this test. The test was scored on a 10-point scale. Analysis by means of two-tailed *t*-test for unequal variance, independent samples showed that the mean test score (M = 7.40, SD = 1.52) of these 101 students did differ significantly from the mean score (M = 6.97, SD =1.74) obtained by the 297 students who were not engaged in the study ($t = 2.21$, p-value = 0.03). However, the absolute difference of 0.43 units between the mean scores of the two groups is not particularly large. Considering the number of participants in this study it might be very well possible that the significant difference between the mean scores of the two groups is the result of the power of the *t*-test used rather than a reflection of a relevant difference in statistical competence.

Materials

The Dutch-language textbook for the course was Imbos, Janssen, and Berger (1996). Concept maps were created for each of the three concepts described in Chapter 5 of the textbook: confidence intervals, one-sample *z*- and *t*-tests, and errors in statistical

inference. The concept maps are similar to the previously mentioned model case explanations (Van de Wiel, et al. 2000) because they depict the contents of an ideal knowledge representation; in this case the representation students should have acquired after having studied a particular concept. The concept maps were formulated as summary like structures consisting of six pre-defined superordinate slots that referred to statistical terms formulas, arithmetic procedures associated with the formulas, interpretation, theoretical background or conditions of application. These slots are listed in Appendix A. A slot can be filled with a variable number of information elements dependent on the specific concept. Across the six slots a difference exists in the character of an atomic element. For example, in the "statistical-terms" slot an element corresponds to a single statistical term, such as "variance" or "standard error." In the "interpretation" slot, on the other hand, an element relates to a single idea, such as the conception that the null hypothesis should be rejected if the *p*-value is less than 0.05. In order to fill the six slots for each concept, the most important information elements per concept were identified by three statistics experts from the faculty of Health Sciences, Maastricht University. For an overview of the concept maps, see Appendix B.

Mastery of the basic principles of statistical inference was assessed by means of a final test, covering a variety of statistical topics including the principles of statistical inference. The test consisted of 20 multiple-choice items and was eventually scored on a 10-point scale. The pass-fail threshold was set at 5.5. It was delivered to the students two months after they had been dealing with the basic principles of statistical inference. Of the 107 students who had participated in the study a total number of 95 students attended the final test.

Procedure

At the final lecture of the four-week instructional cycle in which the basic principles of statistical inference were covered, students were given the opportunity to get additional explication on poorly understood aspects of the topics discussed in the course. Before the lecture started, the lecturer handed out blank sheets to the students and asked them to write down everything they had learned during the instructional cycle. Furthermore, he told the students that the recall protocols would serve an important role in evaluating the course. The students were informed that statistical concepts which were in need of further explication could be identified by means of analyzing the recall protocols and that these topics might be given extra attention in next year's course. It is important to note that students had not anticipated participation in the free recall study. Because it is impossible for students to have prepared themselves for the task, for example by rote learning important concepts, the recall protocols were supposed to give an indication of the structure and content of the students' knowledge structures with respect to the basic principles of statistical inference. Although there was a time constraint of 30 minutes, students took a maximum of 10 minutes to write down everything they could remember of the subject matter. Of the 107 students who attended the lecture, 101 filled out the recall protocols. After the recall protocols were collected, the actual lecture started.

Analysis of the recall protocols

In order to analyze the students' knowledge with respect to the basic principles of statistical inference, a three-step analysis was carried out for each of the following concepts: confidence intervals, one-sample *z*- and *t*-tests, and errors in statistical inference. First the recalled elements were mapped onto the slot-elements of the appropriate concept map. Dependent on the quality of mapping, each recalled element was defined as correct, incomplete or incorrect. Incompletely recalled elements should be considered as rather vague statements that indicate a rudimentary, but superficial understanding of the subject matter. Incorrectly recalled informational elements, on the other hand, were reflections of profound misunderstanding. The weights attached to these three qualifications were respectively, 1, 1/2, and 0. Therefore, the recall protocol for one participant could be quantified into a total recall score by means of the formula:

(Number of correctly recalled propositions) + (Number of incompletely recalled propositions * ½)

For the concept maps all slot-elements had a weight equal to 1. In order to determine whether students have met the learning objectives for a concept, the total recall score was compared to the maximum number of slot-elements in the concept map.

Second, the distribution of the total recall score across the six superordinate slots of the appropriate concept map was examined. Recalled elements that did not match with any of the six slot-categories were placed into a rest category. Conceptual understanding for a single concept was assigned to each participant whose total recall protocol comprised elements from four different slots, provided that at least one of these recall elements reflected an interpretation or theoretical background knowledge. Conceptual understanding was assessed on a dichotomous scale; it was either absent or present. Students who had, at a minimum, obtained conceptual understanding for the concept of one-sample *z*- and *t*-tests were assumed to have acquired overall conceptual understanding on the topic of basic principles of statistical inference. The rationale for this criterion came from the fact that the concept of one-sample *z*- and *t*-tests, as reflected in the size of the concept map, formed the core of Chapter 5 of the textbook.

Finally, the recall protocols were examined for the presence of misconceptions. As outlined before, the total recall score was based on correct, incomplete, or incorrect information elements. The incompletely and incorrectly recalled elements were considered to be misconceptions. Prototypical examples of the incompletely and incorrectly recalled elements were identified, thereby illustrating the misconceptions students hold with respect to a concept.

After the three concepts - confidence intervals, one-sample *z*- and t-tests, and errors in statistical inference - had been submitted to the previously described three-step examination, an additional analysis was carried out. The mean final test score of students with overall conceptual understanding of the topic of the basic principles of statistical inference was compared to the mean score of students without overall conceptual understanding.

Statistics

For the comparison of the total recall scores with the maximum number of slot-elements in the concept maps, one sample, one-tail, 0.05 level *t*-tests were conducted. To determine whether students with overall conceptual understanding performed better on the final test than students without overall conceptual understanding a one-tail, two-sample *t*-test was used. A subset of the free recall protocols was scored by the first author and an independent rater. For every free recall protocol, the total recall scores of the three aforementioned concepts were summed up for each rater. The sum scores of the two raters were correlated to yield an interrater agreement of 0.93.

5. Results

Knowledge representation for the concept "confidence intervals"

For the concept of confidence intervals, students produced on average 0.69 correct, 0.10 incomplete and 0.07 incorrect informational elements. The mean total recall score was calculated to be 0.74 (with standard error = 0.08) for this concept. This score reflects the elaborateness of the average mental representation on the topic of confidence intervals. A histogram of the total recall scores is shown in Figure 1. The comparison of the total recall score to the maximum number of slot-elements in the concept map ($n = 11$) revealed a significant difference ($t = -128.55$, p-value < 0.005), showing that the total recall score was considerably smaller than the maximum number of slot-elements in the concept map.

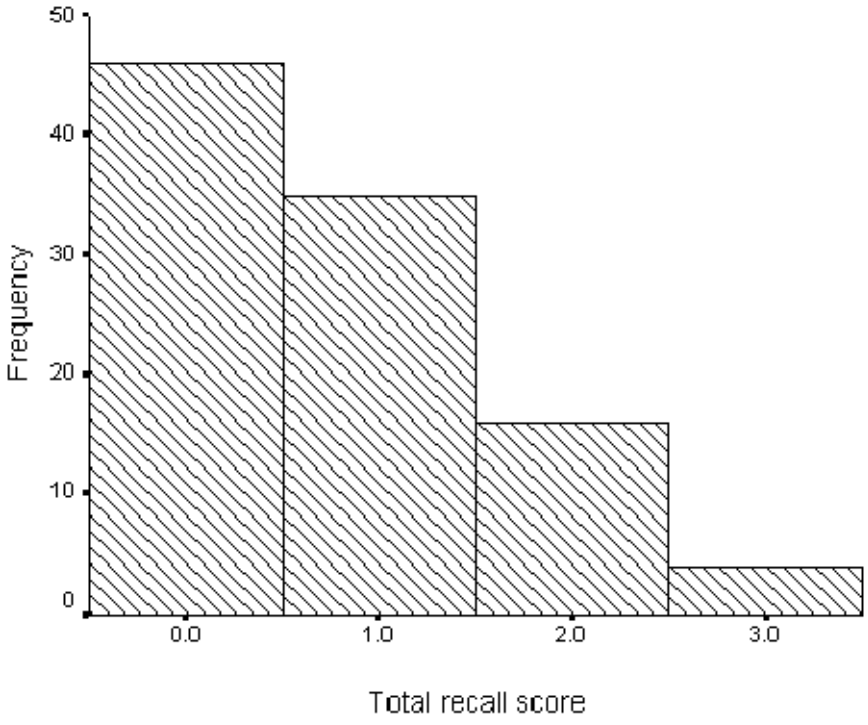


Figure 1

Figure 1. A histogram of the total recall scores for the concept of confidence intervals.

In Table 1, the contribution of each of the seven categories to the total recall score is depicted. The decompositions of the score revealed that students predominantly recalled statistical terms. The remaining part of the recall protocols consisted of formulas and interpretations. On the basis of the data depicted in Table 1 it can be inferred that none of the students had attained conceptual understanding. Remember that conceptual understanding for a single concept was assigned to each participant whose total recall protocol included elements from four different slots, provided that at least one of these recall elements reflected an interpretation or theoretical background knowledge.

Table 1. Contribution of each category to the total recall score for the concept of confidence intervals.

	Mean score	Standard error	Proportional contribution (%) in the concept map	Proportional contribution (%) in the student responses
Statistical term	0.58	0.06	5.27	79

Formula	0.07	0.02	0.64	9
Procedure	0.00	0.00	0.00	0
Condition	0.00	0.00	0.00	0
Interpretation	0.09	0.03	0.81	12
Theoretical background	0.00	0.00	0.00	0
Rest category	0.00	0.00	0.00	0
Total	0.74		6.72	100

Prototypical examples of incompletely and incorrectly recalled informational elements, with respect to confidence intervals, are provided in Table 2. Interestingly, all of the incorrect prototypical examples contained the misconception of equating α with the confidence interval.

Table 2. Prototypical examples of incompletely and incorrectly recalled elements for the concept of confidence intervals.

Incomplete	Incorrect
If the outcome lies within the acceptance area, the null hypothesis is not rejected.	... A confidence interval is chosen: α.
A confidence interval can be used to determine the plausibility of the null hypothesis.	The level of significance depends on the confidence interval.
If you have determined the confidence interval it is possible to see whether or not the null hypothesis should be rejected.	½ α = confidence interval

Knowledge representation for the concept "one-sample *z*-and *t*-tests"

For the concepts of one-sample *z*- and *t*-tests on average, 8.91 correct, 1.14 incomplete, and 0.38 incorrect recall elements were generated. Therefore, the mean total recall score on this concept was 9.48 (with standard error 0.47). A histogram of the total recall scores

is depicted in Figure 2. Analysis revealed a significant difference (t = -43.83, p-value < 0.005) between the elaborateness of the total recall score and the maximum number of slot-elements in the concept map (n = 30).

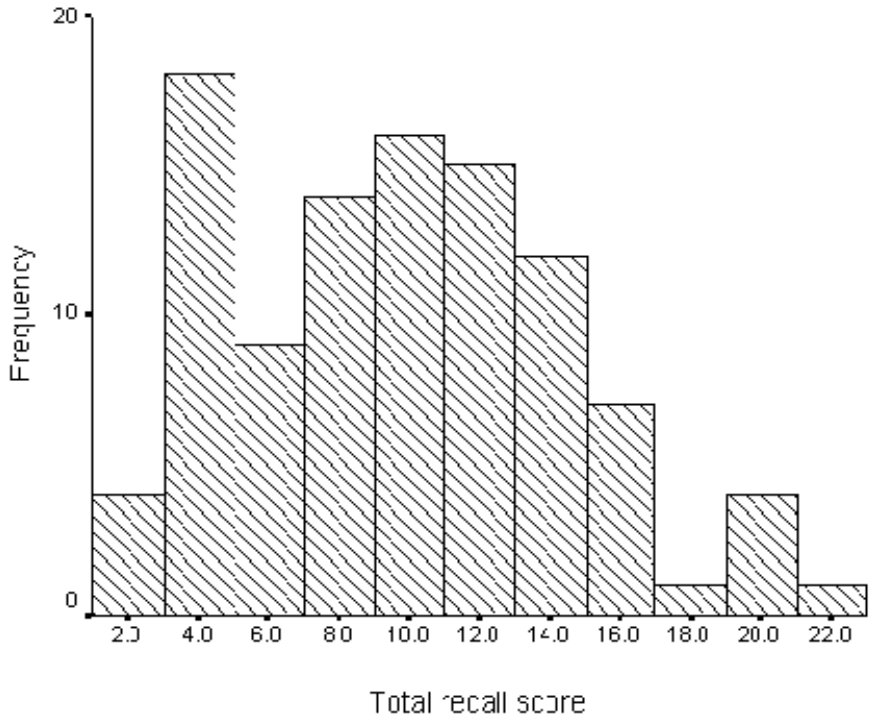


Figure 2

Figure 2. A histogram of the total recall scores for the concepts of one-sample z- and t-tests.

In Table 3, the contribution of each category to the total recall score is presented. Students mainly recalled statistical terms and formulas. Furthermore, the remaining part of the recall protocols consisted of elements referring to the procedure of executing a z- or a t-test, the interpretation of the outcome of a z- or a t-test or to the theoretical background associated with this concept. On the basis of the data in Table 3 conceptual understanding was assigned to 37 students.

Table 3. Contribution of each category to the total recall score for the concept of one-sample z- and t-tests.

	Mean	Standard	Proportional	Proportional

	score	error	contribution (%) in the concept map	contribution (%) in the student responses
Statistical term	5.16	0.23	17.20	54
Formula	2.43	0.20	8.10	26
Procedure	0.33	0.06	1.10	3
Condition	0.63	0.06	2.10	7
Interpretation	0.88	0.11	2.93	9
Theoretical background	0.00	0.00	0.00	0
Rest category	0.05	0.02	0.17	1
Total	9.48		31.60	100

Prototypical examples of incompletely and incorrectly recalled elements are shown in Table 4. Most of the prototypical examples of incorrectly recalled elements again reveal the misinterpretation of α. Some reveal misconceptions concerning the ideas underlying hypothesis testing.

Table 4. Prototypical examples of incompletely and incorrectly recalled elements for the concepts of one-sample *z*- and *t*-tests.

Incomplete	Incorrect
The *t*-test is used if the standard deviation of the population is unknown.	Reject the null hypothesis if the *p*-value = α.
If the calculated value falls within the critical area, the null hypothesis should be rejected.	The alternative hypothesis is the assumption that is tested.
T-tests and *z*-test are used to determine the reliability of the null hypothesis.	α = the probability that the null hypothesis is rejected when the null hypothesis is wrong.

Knowledge representation for the concept "errors in statistical inference"

For the concept of errors in statistical inference, students produced 2.13 correct, 0.12 incomplete, and 0.23 incorrect information elements. Therefore, the total recall-score on this concept was 2.19 (standard error = 0.19). A histogram of the total recall scores is shown in Figure 3. Analysis indicated a significant difference (t = -61.14, p-value < 0.005) between the number of elements in the recall protocols and the maximum number of slot elements in the concept map (n = 14).

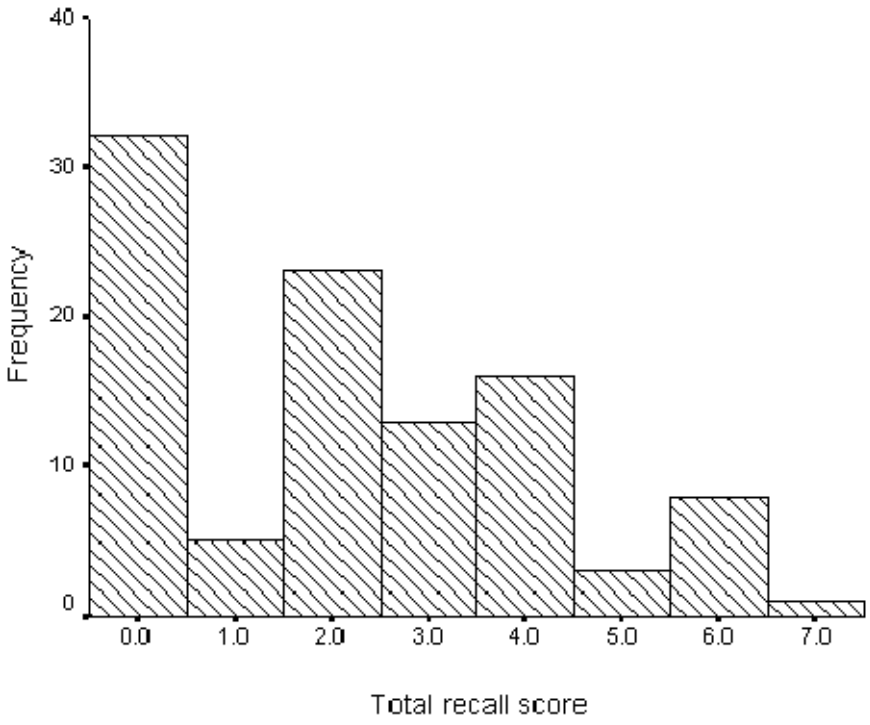


Figure 3

Figure 3. A histogram for the total recall scores for the concept errors in statistical inference.

Table 5 shows the contribution of each category to the total recall score. Students mainly recalled statistical terms and formulas. The remaining part of the recall protocols consisted of interpretations. Based on the data in Table 5 conceptual understanding could not be assigned to any student.

Table 5. Contribution of each category to the total recall score for the concept errors in statistical inference.

	Mean score	Standard error	Proportional contribution (%) in the concept map	Proportional contribution (%) in the student responses
Statistical term	1.37	0.12	9.79	63
Formula	0.46	0.07	3.29	21
Procedure	0.03	0.03	0.21	1
Condition	0.00	0.00	0.00	0
Interpretation	0.33	0.07	2.36	15
Theoretical background	0.00	0.00	0.00	0
Rest category	0.00	0.00	0.00	0
Total	2.19		15.65	100

Prototypical examples of incomplete and incorrect informational elements, as found in the recall protocols, are presented in Table 6. A glance at the prototypical examples of incorrectly recalled elements shows that the non-adequate notion of type 1 and type 2 errors was the most frequently occurring misconception.

Table 6. Prototypical examples of incompletely and incorrectly recalled informational elements for the concept of errors in statistical inference.

Incomplete	Incorrect
α: the null hypothesis is rejected, although the null hypothesis is correct.	Type 1 error: reject the null hypothesis when the null hypothesis is incorrect.
β: the null hypothesis is not rejected although the null hypothesis is incorrect.	Type 2 error: reject the null hypothesis when the null hypothesis is correct.
Power, 1 - β, is somehow related to the alternative hypothesis.	One kind of error occurs in case of a disproportional large α. This will result in a faulty decision.

The relationship between conceptual understanding and performance on the final test

To compare the mean final test scores of students with or without overall conceptual understanding, a one-tail, one-sample *t*-test was conducted. Of the 107 students who were involved in the experiment, 95 took part in the test. Their mean score, expressed on a 10-point scale, was 6.40 (with standard deviation = 1.52). Analysis revealed a significant difference (t = -2.08, p-value = 0.04), suggesting that students with overall conceptual understanding (N= 37, M = 6.80, SD = 1.34) performed better than students without overall conceptual understanding (N = 58, M = 6.20, SD = 1.57). However, the absolute difference between the groups can be considered to be marginal. Possibly, the reason for the emergence of a significant effect is not relevant and might attributed to the considerable power of the test.

6. Discussion

In the article of Van Boxtel, Van der Linden, and Kanselaar (2000), a distinction is made between collaborative learning as a process and as a learning environment for a collaborative task. Collaborative learning as a process refers to the learning activities students potentially engage in while socially interacting with peers: verbalization of their understanding of theoretical issues, collaborative reasoning with scientific concepts, the critical and constructive engagement with the ideas of the group members, solving of theoretical controversies, and the collaborative generation of sophisticated theoretical frameworks with respect to a scientific subject. These laborious activities and the accompanying high-quality discourse will eventually lead to the development of highly sophisticated knowledge representations of the subject under study and thus to conceptual understanding.

On the other hand, the learning environment consists of a collaborative learning task, such as making a poster or doing a project. In order to induce high-quality discourse a collaborative task should have certain characteristic aspects. For example, it is important that students experience positive interdependence, meaning they share the same activities to achieve a common goal, and that students possess a comparable prior knowledge level.

At the faculty of Health Sciences, statistics educators introduced a collaborative learning task, that is, the discussion of the statistical topic at hand, which had a number of high-quality discourse promoting characteristics. For example, the free-exchange of ideas concerning statistical topics was thought to be elicited by means of creating small-collaborative groups consisting of individuals with similar levels of prior statistical knowledge. In addition positive interdependence was assumed to be established by providing the group members with a common goal, in this context, conceptual statistical understanding. It was thus reasonable to assume that the constructive statistical learning environment would stimulate the students to link statistical concepts together into rich knowledge representations, comprising not only terms and formulas, but also the theoretical foundations of formulas, procedural knowledge, interpretations of

mathematical calculations, and a notion of when to apply the learned statistical knowledge.

In this study, the method of free recall was used to obtain insight into the students' knowledge representations with respect to the basic principles of statistical inference. The analyses of the free recall protocols, in which both the quantitative and qualitative aspects of the knowledge representations were considered, provided information on the effects of a prototypical constructive learning environment. These analyses showed somewhat disappointing results regarding the predicted effect of the constructive learning environment.

First, the comparison of the total recall scores with the maximum number of slot-elements in the concept maps revealed that after a four week period of supposedly constructive learning on the basic principles of statistical inference, the average size of the knowledge representations for confidence intervals, one-sample *z*-tests and *t*-tests, and errors in statistical inference proved to be significantly smaller in comparison to the size of the concept maps.

Second, the qualitative analyses of the recall protocols also showed disappointingly low levels of conceptual understanding. The distribution of the total recall score across the seven categories was roughly the same for the three concepts. Statistical terms and formulas took a disproportionately large share of the total recall score while interpretations and background knowledge were hardly mentioned. In addition, the examples of incorrectly recalled elements contained some serious misconceptions. For instance, the idea that the null hypothesis should be rejected in case of a *p*-value = α, does not reflect a particularly good understanding of the subject matter.

Finally, it was demonstrated that merely 40% of the participants who took the final test had acquired overall conceptual understanding on the basic principles of statistical inference at the end of the instructional cycle. This is an especially low percentage considering the constructivist nature of the learning environment.

In sum, the findings obtained in the present study suggest that, despite the integration of a well-designed collaborative learning task, the statistical learning environment was not effective in promoting adequate knowledge representations of some important statistical concepts. An elaboration on these peculiar findings will now be given.

An explanation for the findings reported in the present study might be found in the assessment of statistical competence by means of multiple-choice tests. It is well-known that learning activities of students are, to a large extent, driven by the test format (Frederiksen 1984). In a constructive learning environment, students are encouraged to engage in laborious cognitive activities; they are stimulated to explore and to critically assess the newly offered information and to incorporate the new information into already existing knowledge representations. In contrast, multiple-choice items are particularly directed at lower-level cognitive processes, such as the memorization of simple factual knowledge. Therefore, the engagement in constructive and elaborate learning activities

will not be a necessary requirement for passing the test and, consequently, the learning activities of the students will be characterized by a superficial dealing with the statistical materials under study. Thus, it is conceivable that the format of the test was devastating for a well-implemented collaborative learning task.

Alternative explanations for the findings of the present study are elusive and additional research seems to be required to arrive at a complete understanding of the findings. Future research should explore the complex relationship between the characteristics of a collaborative learning task and the collaborative learning process. In our research, the method of free recall could be used to determine the quality of the knowledge representations, which have developed as a result of the collaborative learning processes triggered by a certain collaborative task intervention. The suggested clarification of the relationship between the characteristics of collaborative a learning task and high-quality discourse is of crucial importance to statistics educators in general, as the obtained knowledge would be very helpful in increasing the effectiveness of curricular innovations.

As a last point in the discussion, it is necessary to deal with the inconsistency between the quantitative and qualitative analyses of the free recall protocols and the mean score of the participants on the final test. How could this satisfactory score match with the poor knowledge representations as reflected in the recall protocols? First, the observed discrepancy can be attributed in part to the structure of the items in the final test. As discussed above, multiple-choice items largely relate to the memorization of simple facts rather than to the desired analytical processing of knowledge associated with the development of rich knowledge representations (see, for example, Levine, McGuire, and Natress 1970; Frederiksen 1984; Birenbaum 1996). Therefore, even participants with low-quality knowledge representations might have been able to arrive at a satisfactory score on the final test.

The emergence of knowledge representations of acceptable quality is the second explanation we would like to put forward with respect to the discrepancy between the final test score and the poor knowledge representations. Remember that there was a considerable time interval of two months, between the conducting of the course of the basic principles of statistical inference and the final test. During this time period students participated in a number of different statistics courses dealing with topics related to the basic principles of statistical inference. Through the repeated exposure to the basic principles of statistical inference, students could have developed some conceptual understanding with respect to this statistical subject.

Finally, it might have been possible that students had in fact obtained adequate knowledge representations of the three concepts, but that they, for whatever reason, were not willing to write down everything they remembered. Perhaps the students had the feeling that a considerable amount of time would be consumed by the free recall task, time which would normally be spent on the explication of difficult statistical concepts. As a result the motivation of the students to engage seriously in the free recall task might have dropped. This presumption is supported by the fact that students had only spent 10

minutes on the free recall task. In order to ensure the appropriate level of motivation, the free recall task could be best administered independent of the other course activities. However, considering the outcomes of both the quantitative and qualitative analysis of the free recall protocols it seems rather unlikely that the students had failed to construct sophisticated knowledge representations with respect to the basic principles of statistical inference solely due to a lack of motivation.

7. Conclusion

A constructive statistical curriculum was designed to elicit active learning. It was designed to motivate the students to elaborate intensively on the statistical topics, thereby creating rich knowledge representations. However, the characteristics of the collaborative task implemented in the statistical learning environment at the faculty of Health Sciences did not demonstrate the expected effect. The findings reported in the present study lead to the conclusion that students have not formed particularly sophisticated representations of the basic principles of statistical inference. Apparently, students were not very eager to explore the statistical topics in a deep, elaborate fashion. It could be argued that this was the consequence of the specific way in which the mastery of the statistical materials is tested in the Health Sciences curriculum. Therefore, it might be advisable to experiment with alternative forms of assessment, such as open-ended questions or portfolios. It should be noted that alternative forms of testing are usually very time-consuming when it comes to grading the students.

In order to completely clarify the findings of the present study, future research at specifying the one-on-one relationship between the characteristic aspects of a collaborative task and the collaborative learning process is needed. Within this research, the method of free recall could be used to determine the effect of the introduction of a collaborative task.

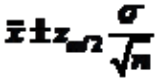
Before exploiting the free recall method, an explicit validation of the method is required. Research could be directed at comparing the knowledge representations of statistical experts to the representations of novices in the domain of statistics. But it might also be conceivable to introduce variations on the free recall method in statistics education reform. For example, a more directed form of free recall, in which individuals are asked to explain statistical concepts, might be equally useful in statistics education reform. Hopefully, the validation procedure will result in the establishment of the free recall method as a valuable tool for everyone who is interested in research the effect of statistics education.

Appendix A - List of the superordinate slots in a concept map.

- Statistical term: This category includes all non-formulaic elements mentioned in chapter 5 of Methodology and Statistics, part1.
- Formula: This category contains all formulaic expressions mentioned in chapter 5 of Methodology and Statistics part 1.
- Procedure: This category refers to the execution of the arithmetic procedures.
- Condition: This category refers to the conditions in which a statistical procedure has to be applied.
- Interpretation: This category contains the interpretations of procedural outcomes and the explications of statistical terms and formulas.
- Theoretical background: Theoretical explications on terms and formulas.

Appendix B - Concept maps for confidence intervals, one-sample *z*- and *t*-tests, and errors in statistical inference.

Confidence intervals.

Terms	Formula(s)	Procedure	Conditions of application	Interpretation	Theoretical Background
Confidence interval (95%) Significance level (α) Population mean Sample mean *z*-score Standard error		Arithmetic procedure for calculating the confidence interval formula.	A confidence interval can be used if one would like to determine, with an a-priori stated reliability, the range of the population mean on the basis of a sample mean.	There is a 95% chance that the population parameter falls within the given confidence interval. or Of 100 samples with sample-size *n*, 95 sample-means will fall within the confidence interval.	The systematic behavior of sample size means obtained from samples with a large enough sample size.

Total number of slot elements:

6 terms + 1 formula + 1 procedure + 1 condition + 1 interpretation + 1 theoretical background = 11 slot-elements.

One sample *z*-test and *t*-tests.

Terms	Formula(s)	Procedure	Conditions of application	Interpretation	Theoretical Background
Null hypothesis Alternative Hypothesis *z*-test *t*-test One-tailed testing Two-tailed testing Population mean Sample mean Standard error *p*-value Significance level Acceptance area	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ H0 H1 α n -1	Arithmetic procedures for calculating the *z*- and *t*-test.	If one would like to see whether or not a sample mean differs significantly from a known or a pre-supposed population mean. If both the population mean and the standard deviation are known, then use the *z*-test. If the population mean and the population standard deviation are unknown then a *t*-test should be used. If $n > 30$ then the *t*-test can	Null hypothesis: The assumption that there is no difference between the population and the sample mean. Alternative hypothesis: This hypothesis states that there is a difference between the population and the sample mean. The *p*-value is the probability of finding the given or a more extreme outcome, provided that the null hypothesis is	The systematic behavior of sample size means obtained from samples with a large enough sample size.

Critical area Degrees of freedom			be approached as a *z*-test.	correct. If the *p*-value = α, then reject the null hypothesis. If the *p*-value = α, then do not reject the null hypothesis.	

Total number of slot-elements:

14 terms + 6 formulas + 1 procedure + 3 conditions + 5 interpretations + 1 theoretical background = 30 slot-elements.

Errors associated with statistical inference.

Terms	Formula(s)	Procedure	Conditions of application	Interpretation	Theoretical Background
Alpha (significance level) Beta Power Type 1 error Type 2 error	α β 1 - β	Arithmetic procedure for calculating the power.		Type 1 error: Reject the null hypothesis when the null hypothesis was in fact correct. Type 2 error: Failing to reject the null hypothesis when the null hypothesis was incorrect. α: The probability of making a Type 1	

				error. β: The probability of making a Type 2 error. Power: The power of a test, that is, the probability of indicating a significant difference when there is a difference.	

Total number of slot-elements:

5 terms + 3 formulas + 1 procedure + 5 interpretations = 14 slot-elements.

References

Ausubel, D. P., and Youssef, M. (1963), "Role of Discriminability in Meaningful Parallel Learning", *Journal of Educational Psychology*, 54, 331-336.

Birenbaum, M. (1996), "Assessment 2000: Towards a Pluralistic Approach to Assessment," in *Alternatives in assessment of achievements, learning processes and prior knowledge*, eds. M. Birenbaum and F. J. R. C. Dochy, Boston: Kluwer Academic Publishers, 3-30.

Chiesi, H. L., Spilich, G. J., and Voss, J. F. (1979), "Acquisition of Domain-related Information in Relation to High and Low Domain Knowledge," *Journal of Verbal Learning and Verbal Behavior*, 18, 257-273.

Cobb, P. (1994), "Constructivism and Learning," in *International Encyclopedia of Education* (2nd ed.), eds. T. Husen and T. N. Postlewhaite, Oxford: Pergamon, 1040-1052.

De Corte, E. (1990), "Acquiring and Teaching Cognitive Skills: A State-of-the-art of Theory and Research," in *European perspectives in psychology* (Vol. 1), eds. P. J. D. Drenth, J. A. Sergeant, and R. J. Takens, London: John Wiley and Sons, 237-263.

Frederiksen, N. (1984), "The Real Test Bias: Influences of Testing on Teaching and Learning," *American Psychologist*, 39, 193-202.

Garfield, J. B. (1993), "Teaching Statistics Using Small-Group Cooperative Learning," *Journal of Statistics Education* [Online], 1(1). (www.amstat.org/publications/jse/v1n1/garfield.html)

Giraud, G. (1997), "Cooperative Learning and Statistics Instruction," *Journal of Statistics Education* [Online], 5(3). (www.amstat.org/publications/jse/v5n3/giraud.html)

Hogg, R. V. (1992), "Towards Lean and Lively Courses in Statistics," in *Statistics for the Twenty-First Century*, eds. F. Gordon and S. Gordon, MAA Notes No. 26, Washington, DC: Mathematical Association of America, 3-13.

Imbos, T., Janssen, M. P. E., and Berger, M. P. F. (1996), *Methodologie en Statistiek 1*, Maastricht: Universitaire Pers Maastricht.

Keeler, C. M., and Steinhorst, R. K. (1995), "Using Small Groups to Promote Active Learning in the Introductory Statistics Course: A Report from the Field," *Journal of Statistics Education* [Online], 3(2). (www.amstat.org/publications/jse/v3n2/keeler.html)

Kilpatrick, J. (1987), "What Constructivism Might be in Mathematics Education," in *Proceedings of the Eleventh Conference of the International Group for the Psychology of Mathematics Education*, eds. J. C. Bergeron, N. Herscovics, and C. Kieran, Montreal, 2-27.

Kintsch, W. (1988), "The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model," *Psychological Review*, 95, 163-182.

Kintsch, W., and van Dijk, T. A. (1978), "Towards a Model of Text Comprehension and Production," *Psychological Review*, 85, 363-394.

Levine, A. G., McGuire, C. H., and Natress, L. W. (1970), "The Validity of Multiple-Choice Achievement Tests as a Measure of Competence in Medicine," *American Educational Research Journal*, 7, 69-82.

Lock, R. H., and Moore, T. L. (1992), "Low-Tech Ideas for Teaching Statistics," in *Statistics for the Twenty-First Century*, eds. F. Gordon and S. Gordon, MAA Notes No. 26, Washington, DC: Mathematical Association of America, 99-108.

Magel, R. C. (1998), "Using Cooperative Learning in a Large Introductory Statistics Class," *Journal of Statistics Education* [Online], 6(3). (www.amstat.org/publications/jse/v6n3/magel.html)

McNamara, D. S., Kintsch, E., Songer-Butler, N., and Kintsch, W. (1996), "Are Good Texts Always Better? Interactions of Text Coherence, Background knowledge, and Levels of Understanding in Learning From Text," *Cognition and Instruction*, 14, 1-43.

McNamara, D. S., and Kintsch, W. (1996), "Learning From Texts: Effects of Prior Knowledge and Text Coherence," *Discourse Processes*, 22, 247-288.

Roberts. H. V. (1992), "Student-Conducted Projects in Introductory Statistics Courses," in *Statistics for the Twenty-First Century*, eds. F. Gordon and S. Gordon, MAA Notes No. 26, Washington, DC: Mathematical Association of America, 109-121.

Scheaffer, R. L. (1992), "Data, Discernment and Decisions: An Empirical Approach to Introductory Statistics," in *Statistics for the Twenty-First Century*, eds. F. Gordon and S. Gordon, MAA Notes No. 26, Washington, DC: Mathematical Association of America, 69-82.

Schmidt, H. G., and Boshuizen, H. P. A. (1993), "On the Origin of Intermediate Effects in Clinical Case Recall," *Memory and Cognition*, 21, 338-351.

Shuell, T. J. (1986), "Cognitive Conceptions of Learning," *Review of Educational Research*, 56, 411-436.

Smith, G. (1998), "Learning Statistics by Doing Statistics," *Journal of Statistics Education* [Online], 6(3). (www.amstat.org/publications/jse/v6n3/smith.html)

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., and Voss, J. F. (1979), "Text Processing of Domain-related Information for Individuals With High and Low Domain Knowledge," *Journal of Verbal Learning and Verbal Behavior*, 18, 275-290.

Tanner, M. A., and Wardrop, R. (1992), "Hands-on Activities in Introductory Statistics," in *Statistics for the Twenty-First Century*, eds. F. Gordon and S. Gordon, MAA Notes No. 26, Washington, DC: Mathematical Association of America, 122-128.

Van Boxtel, C., Van der Linden, J., and Kanselaar, G. (2000), "Collaborative Learning Tasks and the Elaboration of Conceptual Knowledge," *Learning and Instruction*, 10, 311-330.

Van de Wiel, M. W. J., Schmidt, H. G., and Boshuizen, H. P. A. (2000), "Knowledge Restructuring in Expertise Development: Evidence From Pathophysiological Representations of Clinical Cases by Students and Physicians," *European Journal of Cognitive Psychology*, 12, 323-355.

Von Glasersfeld, E. (1995), "A Constructivist Approach to Teaching," in *Constructivism in Education*, eds. L. P. Steffe and J. Gale, Hillsdale, NJ: Lawrence Erlbaum Associates, 3-16.

Wheatley, G. H. (1991), "Constructivist Perspectives on Science and Mathematics Learning," *Science Education*, 75, 9-21.

P. P. J. L. Verkoeijen
Institute of Psychology
Faculty of Social Sciences
Erasmus University of Rotterdam
The Netherlands
Verkoeijen@fsw.eur.nl

Tj. Imbos
Department of Methodology and Statistics
Faculty of Health Sciences
Maastricht University
The Netherlands
Tjaart.Imbos@stat.unimaas.nl

M. W. J. van de Wiel
Department of Experimental Psychology
Faculty of Psychology
Maastricht University
The Netherlands
M.vandeWiel@psychology.unimaas.nl

M. P. F. Berger
Department of Methodology and Statistics
Faculty of Health Sciences
Maastricht University
The Netherlands
martijn.berger@stat.unimaas.nl

H. G. Schmidt
Institute of Psychology
Faculty of Social Sciences
Erasmus University of Rotterdam
The Netherlands
schmidt@fsw.eur.nl
