ASSESSING STUDENTS' STATISTICAL COMPETENCE BY MEANS OF WRITTEN REPORTS AND PROJECT WORK

BIEHLER, Rolf
University of Kassel
Germany

As part of the assessment after an introductory statistics course, students had to do a small project and submit a written report describing their methods, results and conclusions.

We supported the report writing and the project work by several means. Among others, we developed an „exemplary project report" they were introduced to. This project report was written in two columns. In the first column the report about a question concerning a data set is with our best knowledge, in the accompanying second column, we reflect on the choices and options to be made in the respective stages of the report. The aim is to stimulate meta-cognitive activity and to help the students seeing the general in the particular of the exemplary report.

We got several dozens of project reports and analyzed them carefully. We developed a grading scheme with several dimensions, including the quality of introductory and concluding sections, the quality of method choice and the quality of analysis and conclusions. We did not only pay attention to statistical quality but also to questions of style of writing such as whether the project question is introduced in a motivating manner and whether clear and convincing conclusions are presented to the reader with good communicative means including adequate graphs.

The grading scheme was used to provide feed-back to the students. On the other hand we used this scheme for a systematic analysis of the available project reports. Weaknesses and strengths, most difficult areas for our students were identified and we were able to reflect on the adequacy and the shortcomings of our guiding "exemplary report" and our grading scheme.

1. INTRODUCTION

At the University of Kassel, a course "Elementary Stochastics" is compulsory for future secondary teachers. The semester long course comprises 4 hours lecture and 2 hours laboratory work per week. The topics are descriptive statistics, probability, and a little bit of inferential statistics. Since the year 2001, I restructured the course giving more emphasis on exploring data, modeling, and simulation (Biehler, 2003). The software FATHOM has been used as a student tool for data analysis and stochastic simulation. Since 2002, the students are required to submit a project report as part of the assessment in the course. As data base to be used in the projects we used a complex data set with 540 cases and about 50 variables that is based on a questionnaire concerning media use and leisure time of 540 11grade high school students: the so-called Muffins data (for details, see Biehler, 2003). The data are so rich that more than hundred different project themes have been worked on with this single data set. We have analyzed several dozens of student projects in three "generations" and developed a project guide for improving the quality of the projects (Biehler, 2005; Heckl, 2004).

2. EXAMPLES AND TYPES OF STUDENT PROJECTS

I sketch some of the project questions that our students introduced in their reports.

Overweight and going in for sports. Overweight is a recognized problem even among young people. We are interested how the body mass index is related to whether people are actively doing sports, and whether there is a difference between boys and girls.

Television watching. My interest is in how long people watch television, whether there is a difference between boys and girls. Moreover I am interested in how interesting people find watching news on the TV as compared to watching "tabloid like programs".

Computer and internet use. How long do people use the computer? Are there any differences between boys and girls, between those who have a computer and those who don't? For what purposes do people use the internet most?

Homework. I am interested in the amount of homework being done by the students. Is it enough? Is it true that boys on average do less homework than girls? Can it be that other activities such as jobbing can affect the amount of homework being done?

Typically as above, a student project involved several variables of the Muffins data set. Answering the questions usually required to do group comparisons. The students had learned descriptive and exploratory methods such as measures of spread and location, histograms, box plots, percentile plots, and bar charts. We also taught them a unit on (descriptive) group comparison. At the time of the project they had not had any instruction of inferential statistics. The Muffins data are a sample of 540 11graders in Germany in the year 2000 that however was not a representative one. We told our students that their exploratory findings will relate to this sample only and that findings cannot automatically be generalized to a larger population. Also we argued that it is difficult to justify causal claims from such a survey study: For instance, from a difference in TV watching between those who own a TV set and those who don't it cannot be concluded that having a TV set in their bedroom "causes" higher TV watching. It could be vice versa as well in that students are more likely to own a TV set of their own if their interest in TV watching is high.

3. HOW TO STRUCTURE A REPORT

We suggest that students structure their report into 3 parts: (1) Introduction (2) Analysis (3) Summary and Conclusions. In the first generation of students' project reports, we got many reports without any substantial section on introduction or on summary & conclusion. The problem is not that any decent report on a topic whatsoever has to have a section that is named "introduction" or "conclusions" but that substantial parts of what we think are part of statistical competence were missing. For instance, a student took "gender differences in newspaper reading" as her topic. His report had 8 paragraphs, each concerned with a different variable related to newspapers (interest in reading local news, sports …), all similarly structured with very detailed descriptions of distributional differences, with no introduction and summary.

Wild & Pfannkuch (1999) developed a process model for statistical thinking: their general framework is the PPDAC cycle: *problem →plan → data → analysis → conclusions*. This framework for statistical thinking can structure *statistical writing*, too. In essence, the above student did write on the analysis section leaving out the steps *problem →plan → data* and *conclusions.* Although we took the (Muffins) data set as given a similar process is needed. Students have to select those variables from the data set that fit to the problem they are going to analyze. Moreover, they have to reflect on the problem to which extend the available attributes are adequate to the problem at all. The conclusions section turned out to be difficult as well. Jambu (1991) coined the term "data synthesis" for the process, where the results of an exploration have to be ordered, compared, assessed according to importance, refined and presented to a potential audience in a convincing way. Data synthesis involves preparing an act of communication that may need specific means of communication, such as graphs and convincing arguments that anticipate possible criticism.

We developed an assessment scheme, partly based on other work on this topic such as Starkings (1997). We assigned different grades to the different parts of a project report. We divided the section "Analysis" in the two major parts "Section of statistical methods" and "interpretation". In order to improve the quality of reports and the underlying statistical thinking, we developed a so-called "project guidebook".

4. THE PROJECT GUIDE

How can we improve statistical thinking and report writing? Well, we should provide good worked-out examples. But how can students learn from examples, how can they learn to see the general in the particular? We intuitively chose the following "two-column" approach. In the first column we wrote an exemplary project report, which is structured into various sections. In the second column, we are commenting on what we do pointing out the general in the particular. We later discovered similarities to the approach in pedagogical psychology to improve students work with “worked-out examples” with “self-explanations”(Reiss & Renkl, 2002). Reiss & Renkl use this approach for supporting students learning to prove. We use also for the complex task of statistical project work, as compared to more simpler uses of this method for teaching and learning routine tasks.

Because of the limited space of this paper, we will pick out some aspects of our guide and relate this to difficulties in students’ reports in a previous generation.

4.1 GUIDING REPORT INTRODUCTIONS

Our exemplary project report was concerned with the attribute “doing homework”. With regard to the second column we distinguish meta-data, motivation and goals for the project, expectations and hypotheses, adequacy between questions and variables. Meta-data in our cases concerned information about the Muffins data set (which constrains possible conclusions). Explicating motivation and goals are to embed the project into a personal or societal interest background that will provide a perspective for the data analysis (e.g. “as a future teacher I am interested in how much homework students do” or “overweight of young people is a societal problem”). Formulating expectations and hypotheses also contributes to setting up a context, a “horizon”, for later data interpretation. For instance, we might expect that girls on average do homework about half an hour more per weekday than boys do. From this we expect a mean difference of about 3.5 hours if we include the weekend. Most important are what we call *distributional expectations*. A statistically educated student should be able to specify not only expectations concerning the average of a variable but also concerning its distribution. Activating personal knowledge about students at school level, one might expect a range between 5 and 15 hours for nearly most of the students.

Discussion the *adequacy* problem might consider that the students were asked to estimate their study time, no objective measurement was taken. The students were not specifically educated for good estimates. One of our project students had a specific hypothesis from a personal background, namely that different schools have different cultures with regard to homework: Therefore one might expect differences if we compare the different schools involved in the Muffins study according to home work.

At the end of the introduction a *research plan* of which variables are going to be analyzed should be formulated. The exemplary report, in the spirit of Exploratory Data Analysis, explains that this is a framework and that the student researcher should be open to new questions and unexpected results that will come up during his/her data analysis. Examples for such “further questioning” are shown in the report.

4.2 GUIDING DATA ANALYSIS

Students encounter four different types of analysis problems: distributional analysis of a numerical attribute or a categorical attribute; group comparisons of one or more numerical or categorical attributes.

When we started asking students to do project work, the lecture had provided all the tools for distributional analysis such as histograms, box plots, and percentile plots, measures of location, center and spread. Moreover we taught concepts of types of distributions symmetric, skewed to the left and right, U – shaped and so on. This is what most textbooks on descriptive statistics do. Moreover, we discussed at length the relative merits and draw-back of each tool. For instance, students should draw several possible histograms by hand that can be compatible with a given box plot and vice versa in order to learn the “diagnostic properties of the tool”. Median and mean were related to distributional shape; it was pointed out what we can see in a percentile plot that can be hidden in a histogram or box plot. We discussed the sophisticated details of box plots and warned students that there may be more than 50% data equal or below the median if we have ties in the data (Bakker, Biehler, & Konold, 2005). They may wish to check this by calculating these frequencies directly from the data and check whether saying that about 50% of the data are below or equal the median is a reasonable approximation in the respective case.

We showed some examples in the course where the tools were used to make a distributional analysis and left it to the students to select from the tools and combine them according to their needs.

Well, how did the students take this approach in the first generation of reports? Not surprising, there was a large variation. I try to characterize types of students.

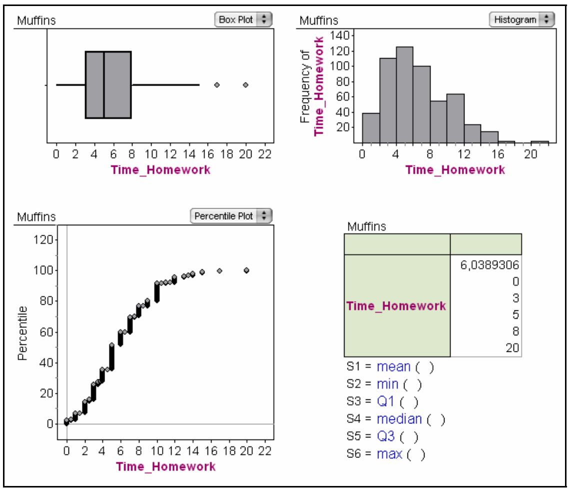
Fallback to averages. Although the major message of the course was "averages are not enough" there was still a group who essentially only used averages for group comparisons or "distributional analyses". When graphs appeared, they were more ornamental in function than being essential.

Personal and context-related styles. A number of students seem to have developed personal styles for distributional analysis and group comparisons partly related to the context, which may favor one tool over another. The personal styles include priorities such as that some students do not "like" percentile diagrams, or box plots or the median and just do not use them voluntarily. Some always add summary values to histograms, mean values to box plots and so on. Some have developed styles and priorities that would certainly be accepted by most experts. Others have developed quite wrong schemes such as one group used the difference between mean and median as an indicator of spread (and not as an indicator of skewness).

Distributional overflow. This group made no priority choice at all, but systematically used all displays and all summary measures they knew and put them in the report. Every distribution was analyzed with great care, paying attention to popular values, outliers, frequencies in certain intervals and so on. Within this group, we can distinguish students who carefully integrate information across displays from those who just itemize and collect the different features from the various displays without integration and mutual comparisons. So we can distinguish *distributional integrators* from *distributional itemizer*s as two extremes. Students of course differ with regard to the quality concerning using information in individual displays and in integrating information across displays. A problem of this approach is, of course, that the students get lost in details and may loose contact to the subject matter goal of the whole analysis.

4.3 SOME PROBLEMS RELATED TO DISTRIBUTIONAL ANALYSIS

Let me start with the example of the variable *Time_Homework* (weekly hours that a student does homework). We show the standard set of displays that students have learned to use in the lecture.



Muffins

Time_Homework	6,0389306 0 3 5 8 20

S1 = mean ( )
S2 = min ( )
S3 = Q1 ( )
S4 = median ( )
S5 = Q3 ( )
S6 = max ( )

Figure 1 Data on weekly hours of doing homework from the Muffins data set, screen capture from FATHOM

I like to expose the following problems we observed and some remedies we put into the project guide

- Inconsistencies between inferences from different displays
- Addition of information across displays instead of integration of information
- Conflicts between intuitive criteria for distributional analysis and statistical concepts
- Discrepancy between richness of distributional information and poorness of subject matter interests

Inconsistences. A number of students tend to translate statements about a median of $\tilde{x}$ into the proposition (exactly) 50% of the data are less and 50% of the data are higher than $\tilde{x}$, and use similar statements with regard to the other quartiles. Because of ties, this is generally not true. At most 50% are less than the median and at least 50% are less or equal the median. A suggested remedy was either not to translate the median back to frequency statements, or to formulate as "about 50%" or to use the software for calculating the exact percentage, especially if one expects large deviations from 50%.

An unexpected side effect was that some students calculate these exact frequencies without having a clear need to do that.

Adding and integrating information across displays. We have recommended distributional integration in the following sense. We consider the box plot as the major summary together with the summary table that provides the exact values of the statistics displayed in the box plot and in addition the mean value. Histogram and percentile plot are considered to have a secondary function in order to check whether the impression the box plot shows has to be corrected or supplemented. For this purpose students learned, among others, to see the different expressions of data density in the different displays (high density: box plot: small distance of quartiles, histogram: high columns, percentile plot: large "gradient". Moreover, they learned how popular values can be hidden in box plots and sometimes in histograms.

An unexpected side effect was that some students describe on all kinds of details such as "the density left from the median is higher than the density right from the median". Although this is true in our display of the homework data, this information as such is hardly relevant to any subject matter question.

Conflicts between intuitive approaches, everyday language and statistical concepts. Consistent with what others have found (Konold, 2002) some students translated the box – information of the box plot (the interval between Q1 and Q3) into the wording: the majority of the students did homework between 3 and 8 hours. As a remedy, we generalized the statistical concepts and introduced the "middle $\alpha\%$", say $\alpha = 90$, as the interval between the two quantiles Q(5%) and Q(95%). We further observed the tendency of some students to repeat the frequency information of a histogram interval by interval. The box plot is supposed to be a summary of the distribution, but is not based on choosing intervals first and than calculating the frequency. So we suggested that students may wish to transform a numerical attribute into a categorical one with classes they choose themselves if they wish to do a different histogram summary. Own criteria of what is "high", "medium", or "low" can be used.

Whereas some students just ignored these recommendations, others were very keen in calculating all these options without there being a real interest stemming from the subject matter context.

Distributional richness and subject matter poorness.

The students were prepared with many sophisticated tools for distributional analysis. To which extent did they feel a need to use these tools in their project depending on the questions they posed?

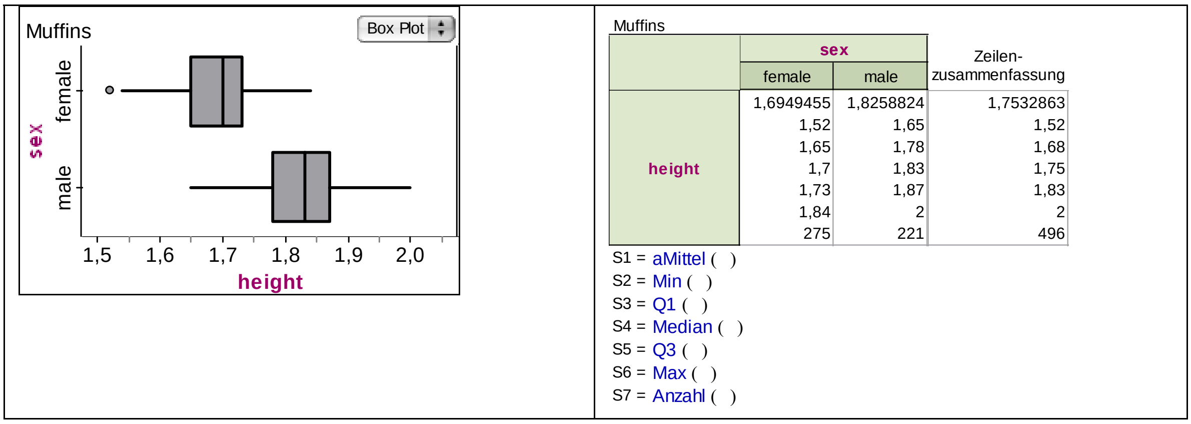
We can grade the sophistication with which the students did their distributional analyses. A subgroup of students spent a lot of time of doing sophisticated distributional analyses but the results were just collected by them and were difficult to interpret in the subject matter context. Statistical experts might do this also sometimes: the statistician provides a rich data analysis but it is not responsible for the detailed interpretation. We wonder, however, whether we should accept such an approach in our projects.

Two factors may have favored such an approach. First, our students' have expected that they would be judged by the sophistication of their tool use. Therefore they felt to be on the certain side, if they used as many tools as possible and read off as many details as they see in the displays. This problem is generated by the didactical situation. The students did not do a "genuine" project for a client, but they new of course that is was part of our assessment.

A second source could be the kind of descriptive stance of the questions some students chose, such as "How much homework do students do?" Or, "what is the difference with regard to homework between males and females?" In particular, when no subject matter based expectations and hypotheses were formulated, it is quite tempting to describe everything you "see". No criteria for selecting results were at hand. We tried to improve on this in asking students to relate questions to a personal and societal context and to develop expectations and hypotheses.

4.4 SOME PROBLEMS RELATED TO COMPARING DISTRIBUTIONS

Initially we underestimated the problems students have with comparing distributions. Let us start with a simple example.



Muffins

	sex		Zeilen-zusammenfassung
	female	male	
height	1,6949455 1,52 1,65 1,7 1,73 1,84 275	1,8258824 1,65 1,78 1,83 1,87 2 221	1,7532863 1,52 1,68 1,75 1,83 2 496

S1 = aMittel ( )
S2 = Min ( )
S3 = Q1 ( )
S4 = Median ( )
S5 = Q3 ( )
S6 = Max ( )
S7 = Anzahl ( )

Figure 2. Height of 11graders (Muffins data)

Students generally assume that males tend to be taller than females before the data analysis. What do they conclude from such box plots? We give some idealized responses:

1. The hypothesis that males tend to be larger is confirmed. This can be seen from the means or medians.
2. The hypothesis that males tend to be larger is confirmed by the min, Q1, median, mean, Q3, max. All these values are higher for the males than for the females.
3. On average, males are 13 cm larger than females.

Answer 1 can be characterized as "fall-back to averages". Answer 2 is often given in a way that the fact that 5 summary values of males are higher than the corresponding of females increases the "evidence" for the hypothesis that males are taller. The comparison between all these values is not regarded as a richer description of what the difference between 2 distributions is.

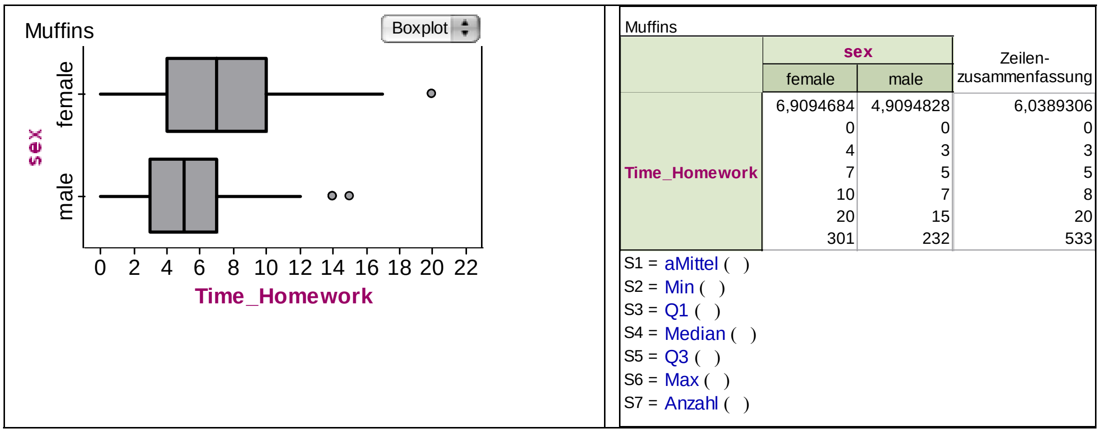
In our project guide, we suggested the following improvements:

1. Do not just confirm/reject hypotheses but try to make quantitative statements concerning the difference of summary values such as statement 3 above.
2. Check basic summary values of group A against those of group B: If all are larger in group B, we can say that our attribute is "statistically larger" in group B than in group A. Consider this being a more rich information than just looking at the difference in means or medians, instead of considering this as a cumulating evidence for the "is larger" hypothesis.

3. Check whether the difference between the summary values is nearly equal. In the above example this is approximately the case: All values are shifted to the right by about 13 cm. We speak of a *uniform (additive) shift* of 13 cm of the whole distribution.
4. Describe differences in distribution - if possible - as deviations from the “shift model” or the “uniform shift model”. Consider, whether a shift can be multiplicative (all summary values are multiplied by approximately the same factor).

The shift model or the uniform shift model often is a model assumption in inference statistics in two sample problems. Such models help to see data under certain perspectives. In our case it helps students to concentrate on comparing distributions as a whole instead of concentrating on individual summary values.

As part of the guidance, various examples for group comparisons seem helpful.



Muffins

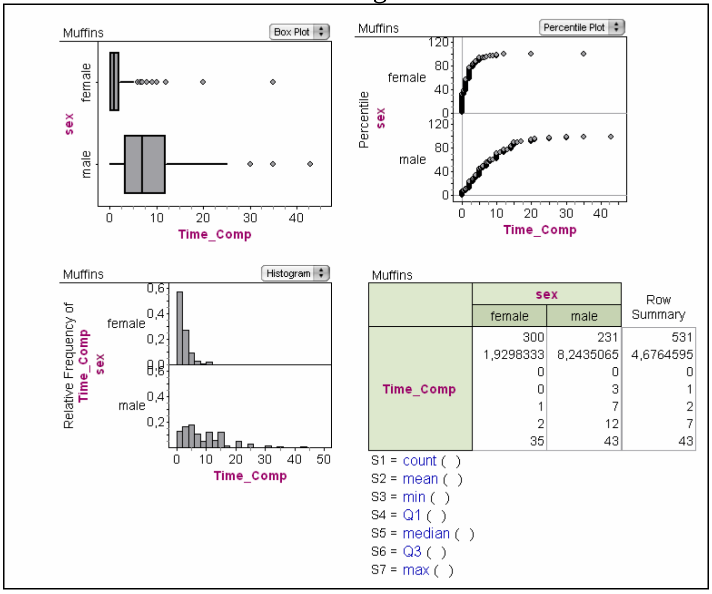
	sex		
	female	male	Zeilen-zusammenfassung
Time_Homework	6,9094684 0 4 7 10 20 301	4,9094828 0 3 5 7 15 232	6,0389306 0 3 5 8 20 533

S1 = aMittel ( )
S2 = Min ( )
S3 = Q1 ( )
S4 = Median ( )
S5 = Q3 ( )
S6 = Max ( )
S7 = Anzahl ( )

Figure 3. Homework (weekly hours) according to gender, Muffins data

For instance, in the case of gender differences in doing homework, we see that a multiplicative shift model seems to be pretty adequate. Females do about 30 to 40% more homework than males, this applies to the whole distribution approximately. As a side effect, the spread increases by this factor, too. It is clear that this multiplicative comparison is not uncommon in advanced statistical practice, where we often find situations where spread increases with level.

Being prepared by these model situations students may try to make comparisons in less clear-cut situations such as the following:



Muffins

	sex		
	female	male	Row Summary
Time_Comp	300 1,9298333 0 0 1 2 35	231 8,2435065 0 3 7 12 43	531 4,6764595 0 1 2 7 43

S1 = count ( )
S2 = mean ( )
S3 = min ( )
S4 = Q1 ( )
S5 = median ( )
S6 = Q3 ( )
S7 = max ( )

Figure 4 Group comparison, weekly time for computer use, according to gender, Muffins data (collected in spring, 2000)

In addition to recognizing a non-uniform shift, the attention can go the minimum of the female distribution, where we see from the box plot that at least 25% of the female students do not use the computer at all. From, the percentile plot we can estimate this proportion as about 35% (Muffins data were collected in 2007). We may also recognize the different types of outliers that the Tukey box plot shows. A female is shown as an outlier from 6 hours per week already whereas a male has to use computers more than 25 hours to become displayed as an outlier.

Group comparison shows already various problems at the analysis level, which makes clear that interpretation and data synthesis is even more difficult.

5. CONCLUSIONS

After some experience with assessing students' statistical competence by means of project reports, we consider the problem to be much more complicated than it looked to us at the beginning. Making explicit the criteria for judging students' work forced us to think much deeper about elementary statistical reasoning with distributions and comparing distributions than has been done in the statistics education community so far.

SOFTWARE

FATHOM^{TM} http://www.keypress.com/fathom/
or German version: http://www.mathematik.uni-kassel.de/~fathom

REFERENCES

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 28 June-3 July 2004. [www.stat.auckland.ac.nz/~iase/publications.php]* (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (2003). *Interrelated learning and working environments for supporting the use of computer tools in introductory courses.* Paper presented at the IASE Satellite Conference on Teaching Statistics and the Internet (CD-ROM Proceedings) [also: http://www.stat.auckland.ac.nz/~iase/publications/6/Biehler.pdf].

Biehler, R. (2005). Strength and weaknesses in students' project work in exploratory data analysis. In M. Bosch (Ed.), *Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education, Sant Feliu de Guíxols, Spain – 17 - 21 February 2005 [http://ermeweb.free.fr/CERME4/CERME4_WG5.pdf]* (pp. 580-590).

Heckl, R. (2004). *Die Bewertung von Projektarbeiten zur Explorativen Datenanalyse in der schulischen und universitären Ausbildung* (Zulassungsarbeit Erste Staatsprüfung (Master thesis)). Kassel: University of Kassel.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. London: Academic Press.

Konold, C., et al.,. (2002). Students' use of modal clumps to summarize data. In *Proceedings of ICOTS 6 [www.stat.auckland.ac.nz/~iase/publications/1/8b2_kono.pdf]*.

Reiss, K., & Renkl, A. (2002). Learning to Prove - The Idea of Heuristic Examples. *Zentralblatt für Didaktik der Mathematik, 34*(1), 29-35.

Starkings, S. (1997). Assessing Students Projects. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education.* Amsterdam: IOS Press. *[www.stat.auckland.ac.nz/~iase/publications/assessbk/chapter11.pdf]* (pp. 139-152).

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-265.