Analyzing Propositions Underlying the Theory of Statistics

Nick J. Broers
Maastricht University

Journal of Statistics Education Volume 9, Number 3 (2001)

Copyright © 2001 by Nick J. Broers, 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 author and advance notification of the editor.

Key Words: Cognitive units; Conceptual understanding; Statistics education; Statistical knowledge.

Abstract

Conceptual understanding of statistics is usually considered one of several aspects of statistical knowledge. It refers to the ability of students to tie their knowledge of statistical ideas and concepts into a network of interrelated propositions. In this study an attempt was made to analyze the theory of descriptive regression analysis into its constituent propositions. Content analysis of the work of nine students revealed that these propositions were used by the students as cognitive units in their mental representation of the statistical theory. Suggestions for a use of constituent propositions as learning tools are discussed.

1. Introduction

It has long been known that people fail to make proper use of statistical rules in the context of everyday problems (see, for example, Tversky and Kahneman 1974; Einhorn and Hogarth 1981) and even academic researchers have been shown often to possess a confused and insufficient grip on statistical tools and concepts (Greer and Semrau 1984; Clayden and Croft 1990). The development of statistical knowledge seems to demand the adoption of rules and ideas that to many are counterintuitive and therefore difficult to master. In light of this, an extensive amount of research has over the years been devoted to the improvement and innovation of the teaching of statistics.

Nonetheless, until relatively recently, little attention was devoted to the precise learning objectives behind the statistics curricula. In what they describe as the "assessment revolution," Garfield and Gal (1999) point out that in recent years much more attention has been devoted to the question of what exactly it is we wish students to master and what particular assessment techniques are best suited to probe whether the learning objectives have been met. In the wake of this revolution, much more clarity and consensus have arisen over key concepts that define the meaning of statistical knowledge.
As currently defined, three key concepts of statistical knowledge -- conceptual understanding (Huberty, Dresden, and Byung-Gee 1993; Schau and Mattern 1997), statistical reasoning (Chervany, Collier, Fienberg, Johnson, and Neter 1977; 1980) and statistical thinking (Wild and Pfannkuch 1999; Chance 2000) -- are hierarchically related. To be able to reason statistically, a student needs to possess knowledge of an integrated body of concepts and ideas. To be able to think statistically is indicative of an overall mental habit, developed on the basis of a long experience in statistical reasoning with respect to independent problem situations.

A representation of the hierarchical development of statistical knowledge in a way that we find heuristically useful, is provided by schema theory. According to this theory from cognitive psychology, people organize knowledge in the form of propositions, that are usually defined as the smallest items of knowledge that can stand as an assertion (McNamarra 1994). In the context of statistics, we may think of simple definitions of concepts like "$r_{xy}$ is a measure of linear correlation," "the mean is a measure of central tendency," or of principles like "$-1 \leq r_{xy} \leq 1$" and "$r_{xy}^2$ gives the proportion of variation explained.” Such propositions are related to each other to form simple cognitive schemata. A schema is usually defined as a mental framework for representing knowledge, encompassing an array of interrelated concepts in a meaningful organization (Sternberg 1996). An example of a simple cognitive schema could be: "$r_{xy}$ is a measure of linear correlation, with values ranging from -1 to +1, whereby -1 or +1 denotes a strict linear relationship between two variables $X$ and $Y$, and a value of zero denotes no linear relationship at all." The simple schemata form cognitive units at a higher level of organization than the constituent propositions. Likewise, the simple schemata are connected to each other to form more complex knowledge structures (for a general overview of schema theory and its implications for instruction, see Van Merrienboer 1997).

Schau and Mattern (1997, p.91) observed "... a critical weakness in post-secondary students who have taken applied statistics courses: they lack understanding of the connections among the important, functional concepts in the discipline." As an aid towards visualizing a network of connected concepts, Schau and Mattern make use of a concept map (as developed by Novak and Gowin 1984). A concept map is a graph consisting of nodes containing concepts (visually depicted as ovals containing a description of the concept) and the connections between the concepts (visually displayed as arrows). As an example, Schau and Mattern (1997) show a concept map with separate arrows going from STATISTICS to DESCRIPTIVES and from STATISTICS to INFERENTIAL. Three separate arrows connect DESCRIPTIVES to CENTRAL TENDENCY, VARIABILITY, and CORRELATIONS. From VARIABILITY, separate arrows show links with $\sigma^2$ and $S^2$, and so on.

For any given course in statistics, a concept map can be constructed which reflects the conceptual network that the student is to master. Construction of such a concept map forces the instructor to make his educational aims explicit. Apart from this, construction of a concept map may be used as an instructional tool in at least two ways (Schau and Mattern 1997). First, a teacher may create a concept map to provide students with an overall view of the interconnected content of the theory they are required to master. Second, a teacher may present the students with an incomplete map and ask them to fill in the gaps. Alternatively, the students may be presented with cards containing concepts and asked to construct a concept map on the basis of these. Such an assignment may stimulate the connected understanding of the student and enhance the formation of networks of interrelated propositions.

While a concept map focusses directly on the way that a student organizes a complex body of statistical concepts into a coherent knowledge framework, it largely ignores the way that a student decomposes the statistical theory into propositions. Propositions do play a part in concept maps, but not in a very articulate sort of way; they appear as small statements commenting on the nature of the arrows that connect two concepts with each other. For instance, in the concept map discussed by Schau and Mattern
(1997), the arrow connecting the concept STATISTICS with DESCRIPTIVE is accompanied by the small statement "can be." So the proposition captured by this connection is "statistics can be descriptive." Concept maps focus on the big picture -- the concepts and their interrelationships -- and we believe that by doing so this visual technique pays scant attention to the way that students actually organize the statistical material. Based on our experience with oral examinations and on a previous study (Broers, in press), we believe that students learn a statistical theory by identifying a number of basic propositions that underlie the study material, and by subsequently trying to establish the links between these basic knowledge fragments. On the basis of this, we have been exploring a different method for monitoring the level of comprehension by the student, based on the identification of the propositions that underlie the theory of statistics that we wish to teach.

A concept map aids an instructor in making educational aims explicit. In a similar way, so does the parcelling out of a statistical theory into its constituent propositions. When we teach a certain topic to a selected audience, like descriptive regression analysis, there are some aspects of the theory that we highlight and make explicit to our audience, and there are other aspects that we take for granted -- because we assume these parts of the theory to be already familiar to our students -- and therefore leave implicit. The propositions that we try to convey to our audience have a certain complexity, which is what makes them hard to grasp for those students who do not possess the necessary preliminary knowledge. For example, we may take some effort in explaining the idea of explained variance ($S^2_y$) in the context of descriptive regression analysis, and stress that $S^2_y / S^2$ gives us the proportion of variance explained in $Y$ on the basis of $X$. What we might leave implicit, is further explication of what we mean by the variance of $Y$, what we mean by the arithmetical mean of $Y$ (necessary in computing the variance), or what is meant by the square of $Y$. All this extra information could be made explicit in the form of additional propositions, but we have decided that this would be redundant for our present group of students.

The question remains of whether it is really possible to write out a finite list of constituent propositions that actually corresponds to the propositions that the students are likely to use. Alternatively, there might be an endless variety of ways that a student might decompose the study material into propositions, and maybe there are no two students that would come up with a similar list.

Assuming that the material can be parcellled out into a finite list of propositions that will correspond to the in-coming ability level of the students, then by making these propositions explicit we obtain an important instrument for assessing the quality and effect of our teaching. For example, if we present students with statistical problems pertaining to the subject material of our course, we may ascertain whether they make use of our propositions when reasoning on these problems. If certain propositions are frequently omitted by students this may alert us to the fact that we need to pay extra attention to these propositions in our course. On the other hand, it may be that students frequently make use of propositions that were left implicit by us. This may mean that our course will become more accessible if we were to make explicit use of these propositions in our teaching. Alternatively, some propositions that we have used may be consistently left implicit by our students in their reasoning, perhaps indicating that these propositions may be given less attention in our future courses. Furthermore, students will be prone to use erroneous propositions in their reasoning. Relating these errors to our list of propositions will alert us to possible weak links in our exposition of the material.

The goal of the present study was twofold. First, we wanted to examine whether it is possible to produce a list of propositions related to a statistical topic, that will recognizably correspond to the propositions that the students themselves use for encoding the information they derive from the course. Second, we wanted to examine if such a list can help us to gain insight into the strong and weak points of our teaching.
2. Method

2.1 Participants

Participants for the study were ten second-year psychology students (out of a class of about 100 students), who volunteered for participation in the study. Each student received renumeration for his or her efforts. All students had successfully completed a course in descriptive regression analysis (amongst other topics of descriptive statistics) during their first year of study. Of the ten students, one had performed poorly on this course, passing the exam only after repeated reexaminations, four had shown a mediocre performance, four were good and one was excellent. Of the good students, one participated in this study in an unmotivated manner, resulting in material that was of no use for further analysis. The analysis was therefore performed on the material of that of the remaining nine students.

2.2 Material

The material was comprised of five separate booklets, each containing an elaborate instruction of the task required, one multiple choice item with four alternatives, and approximately five blank sheets. The instruction stated that the objective was to gain insight into the thought process leading up to the qualification of a multiple choice alternative as either right or wrong. The instruction required students to put their thoughts on paper as explicitly as possible: it was stressed that whether they got the answers right or wrong was not important, what mattered was that students should give a clear exposition of their reasoning. Apart from arguing why one of the four alternatives was the correct answer to the problem, the student also had to argue why each of the remaining alternatives had to be wrong. If he or she had no idea why a given alternative might be either right or wrong, the student had to indicate this explicitly. As we noted in the previous subsection, the work of one student was omitted from the analysis: contrary to the instruction he made no effort to explicate his thought process, but instead gave short (one line) answers that did not contain any useful information.

An example of one of the items used is the following:

In the same study, two variables $X$ and $Y$ correlate $r_{xy} = 0.50$ with each other, whereas the same variable $X$ correlates $r_{xz} = 0.70$ with a third variable $Z$. Independently, two bivariate regression analyses are performed: $Y$ is regressed on $X$ and $Z$ is regressed on $X$. After analyzing the results it is shown that $S_{y'}^2/S_y^2$ (the explained variance of $Y$ divided by the total variance of $Y$) equals 0.55, whereas $S_{z'}^2/S_z^2$ equals 0.49. Which conclusion may be drawn with certainty?

a. A nonlinear model has been used for regression of $Y$ on $X$ (*)

b. A nonlinear model has been used for regression of $Z$ on $X$

c. The prediction of $Z$ based on $X$ is more accurate than the prediction of $Y$ based on $X$

d. The researcher has made computational errors, for the results are impossible

The asterisk marks the correct alternative. The students are expected to recognize this, as one of the learning objectives of the regression course was to make clear that the commonly used linear regression models are not the only possible ones and that alternative models may be better in terms of the proportion of variance that can be explained. At this level, the term "nonlinear" is meant to refer to a higher degree polynomial. Although such a model is linear in the parameters, the students have learned in high school
to call such a relationship "nonlinear," as opposed to a first degree or linear relationship. A list of all five items is presented in Appendix A. The example above figures as item number 3 in the list. Since the subject was required to reason on each of the alternatives, it could be said that item number 3 actually represents three subitems: items 3A, 3B and 3C. The last alternative is not independent of the first three: if a student thinks that computational errors have been made, than it follows logically that alternatives A, B and C cannot be correct. However, if a student thinks that a higher degree polynomial has been used for regression of $Y$ on $X$, this in itself does not logically imply that $Z$ has been regressed on $X$ with use of a linear model. Following a similar line of reasoning, there were 15 subitems in total: item 1, subitem 2A, 2B, 2C and 2D, subitems 3A, 3B and 3C, subitems 4A, 4B, 4C and 4D, and subitems 5A, 5B and 5C (see Appendix A).

Each subitem requires a number of propositions for its solution. Subitem 3A for example, requires the following propositions:

1. \( r_{xy}^2 \) gives the proportion of variance explained in $Y$ by linear regression on $X$
2. \( S_{y/}^2/S_y^2 \) gives the proportion of variance explained in $Y$ by regression on $X$
3. The proportion of variance explained in $Y$ by $X$ on the basis of a nonlinear relationship is equal to or higher than the proportion of variance explained in $Y$ by $X$ on the basis of a linear relationship
4. The greater the proportion of variance explained on the basis of a regression model, the better the predictions yielded by that model

The above four propositions formed some of the basic building blocks used for teaching the theory of descriptive regression analysis. Of course many other propositions are necessary for solving the above item (like the idea of a nonlinear relationship), but these were considered as required knowledge at the beginning of the lectures on descriptive regression analysis and therefore left implicit. What has been taught in the course but is not listed above, is a proposition like \( r_{xy} \) is a measure of the strength of linear correlation between $X$ and $Y'$. This proposition is not necessary for correctly solving the item. \( r_{xy} \) is given, and it is likely that many students when reasoning on the information given will directly focus on the fact that in the item \( r_{xy}^2 \) does not equal \( S_{y/}^2/S_y^2 \). The four propositions stated above are the propositions that we considered necessary to be able to teach the topic, and in this sense we consider this set of propositions to be necessary for the students to be able to solve the item.

An overview of the propositions that together form the basic building blocks of descriptive regression analysis, as taught by us to psychology students, is given in Appendix B. The teaching of the topic consists of two lectures, complemented with two workgroups in which the students, organized in small numbers and headed by a senior student, work on a number of exercises related to the topics that were discussed in the lectures. We used as a text Moore and McCabe (1999). Not included in Appendix B are propositions on the use of residual plots and on the role of outliers in the determination of the best line of fit.

2.3 Analysis

The individual sentence formed the unit of analysis. Each sentence was scanned for the presence of one or more propositions. Not all sentences contained propositions. Some sentences merely restated the information given in the text, like "\( r_{xy} \) equals 0.50." Other sentences contained statements like "This is a
tough question," and "I am not sure whether I have enough knowledge to solve this item." Many sentences contained logical conclusions like: "In view of the preceding, this result seems contradictory."

Where sentences did contain propositions, they were classified in one of the following four categories:

1. Explicit

   When a student made use of a proposition that corresponded with a proposition that we had in mind as one of the basic building blocks this was categorized as an explicitly mentioned proposition. For instance, one student wrote down the following, whilst looking at alternative 3C:

   "Is the prediction of Z better than that of Y? No that’s not right. Y on X gives 0.55, whereas Z on X gives only 0.49. So more of Y is explained. Alternative C is therefore incorrect."

   Although this student does not literally state the proposition we have in mind (the greater the proportion of variance explained, the better the prediction) he does unambiguously refer to it, so it is registered as an instance of an "explicitly stated proposition."

2. Implicit

   A proposition was registered as implicitly referred to, when the students reasoning on a problem was sound and the conclusion correct, yet at no point did she overtly refer to the proposition concerned. An example is given by a student who wrote:

   "\( r_{xz} \) gives a measure of linear association. \( r_{xz}^2 = 0.49. S_x^2/S_z^2 = 0.49. \) This corresponds. So alternative b is incorrect."

   Since this student at another point in her text explicitly stated that \( S_{y'}^2/S_y^2 \) gives the proportion of variance explained, we notice that in the excerpt above she implicitly stated that in the case of a linear model, \( r^2 \) gives the proportion of variance explained (although it does not follow that she realizes that this equivalence only holds when the least squares criterion is met).

3. Additional

   A proposition was classified as additional when it concerned a valid proposition that was not included in our list. Such propositions were typically more specific than the propositions listed by us (for example, "A Z-transformation expresses the scores as standard deviations from the mean," which forms a piece of detailed information that was considered by us as preexistent knowledge and therefore not taken up as a learning objective) or conversely were more general than the propositions in our list (for example, "the Z-transformation leaves the proportion of variance of Y explained by \( X \) unchanged," instead of stating the two propositions "A Z-transformation is a linear transformation" and "\( r_{xy} \) is invariant under linear transformation of \( X \) and \( Y' \)").

4. Erroneous

   A proposition was categorized as erroneous in all those instances where an assertion was made that was clearly incorrect. For example: "A Z-transformation is a nonlinear transformation."

To explore whether students make use of our list of propositions, or whether instead they make use of
more idiosyncratic lists, we determined the number of times a given proposition was explicitly or implicitly mentioned. In addition, we compiled a list of additional and erroneous propositions. We expected that our list of propositions would be the propositions that the students actually used in the context of solving statistical problems. We expected relatively few additional propositions.

With regard to our second research question, we looked for propositions that stand out as relatively unfamiliar to our students, as evidenced by the fact that relatively few students make use of these propositions. In conjunction with this frequency analysis, we will examine the list of erroneous propositions to see whether this provides clues to possible weak spots in our teaching.

3. Results

Per item, each student was required to judge each of the alternatives as either correct or incorrect, and to provide a motivation for that qualification. A response was qualified as correct when the student correctly identified an alternative as either right or wrong, giving a sound argument for the response. When a false alternative was judged to be false on the basis of an irrelevant or incorrect line of argument, the response was qualified as incorrect. Table 1 gives an overview of the number of correct responses for each student. These results correspond closely to the performance that the students had shown on the actual course exam, with the good students obtaining sum scores of 10 or 11 and the weakest student scoring only two correct responses.

| Student ID | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 |
|------------|----|----|----|----|----|----|----|----|----|
| Number correct | 8  | 7  | 3  | 11 | 8  | 10 | 11 | 2  | 10 |

The subitems used clearly differed in level of difficulty. Table 2 provides an overview of these differences. Since a total of nine students worked on these 15 subitems, the maximum score was 9. No subitem was answered correctly by all of the students, but the results indicate that subitems 2A, 2D and 4B were relatively easy, whereas subitems 1, 2C, 4C, 5B and 5C were clearly difficult.

To get an impression of the extent to which the learning objectives of the course have been met, i.e. to see whether the propositions are indeed imbedded as cognitive units in a mental organisation of the course material, we checked the number of students that mentioned (either explicitly or implicitly) a given proposition at least once. When all of the propositions are at any time used by each of the students, we can conclude that our objective to provide the students with a mental framework representing the statistical theory has been at least partially fullfilled. Table 3 gives the relevant frequencies.

| Subitem | 1 | 2A | 2B | 2C | 2D | 3A | 3B | 3C | 4A | 4B | 4C | 4D | 5A | 5B | 5C |
|---------|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|

Page 7 of 18
As is clear from the table, the majority of propositions are at any time used by the students in their attempts at solving the statistical problems. In the remainder of this article, we will refer to propositions that were mentioned by at least six different subjects as "familiar," and to propositions that were mentioned by fewer than six different subjects as "unfamiliar." As we can see, on the basis of this definition only six out of 23 propositions could be considered as unfamiliar.

To get another impression of the extent to which the students have mastered the course material, we checked how many different propositions were used (either explicitly or implicitly) by each of the nine students. The results are shown in Table 4.

Table 3. Number of Students who Mentioned a Proposition at Least Once (propositions presented in order of decreasing familiarity)

| Prop | Freq | Prop | Freq | Prop | Freq |
|------|------|------|------|------|------|
| P7   | 8    | P3   | 7    | P17  | 6    |
| P8   | 8    | P4   | 7    | P18  | 5    |
| P9   | 8    | P5   | 7    | P21  | 5    |
| P10  | 8    | P23  | 7    | P15  | 1    |
| P11  | 8    | P1   | 6    | P16  | 1    |
| P12  | 8    | P6   | 6    | P20  | 1    |
| P22  | 8    | P13  | 6    | P19  | 0    |
| P2   | 7    | P14  | 6    |      |      |

As is clear from the table, the majority of propositions are at any time used by the students in their attempts at solving the statistical problems. In the remainder of this article, we will refer to propositions that were mentioned by at least six different subjects as "familiar," and to propositions that were mentioned by fewer than six different subjects as "unfamiliar." As we can see, on the basis of this definition only six out of 23 propositions could be considered as unfamiliar.

To get another impression of the extent to which the students have mastered the course material, we checked how many different propositions were used (either explicitly or implicitly) by each of the nine students. The results are shown in Table 4.

Table 4. Number of Different Propositions Recalled by the Nine Students

| Student ID | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 |
|------------|----|----|----|----|----|----|----|----|----|
| Number of different propositions recalled | 18 | 14 | 10 | 17 | 16 | 21 | 20 | 0  | 19 |

Not one student actually used all 23 propositions, although the student with excellent results on the
course exam came very close to this, with 21 propositions recalled and used in the process of problem solving. The weakest student had not internalized a single proposition. Note, with reference to Table 1, that she had correctly solved two of the 15 subitems, using a correct line of reasoning. This implies that there are arguments and knowledge items, different from the ones held in mind by the teacher, that can lead to a correct interpretation and solution of the problems presented. However, most of the correct solutions were based on our 23 propositions. The results in Table 4 correlate 0.91 with the results in Table 1, indicating that 81% of the variance in the number of correctly provided solutions can be explained by use of the intended propositions.

As we stated before, we wanted to discriminate between propositions that were explicitly mentioned and those that were used implicitly. By definition, a proposition was considered implicit when it was used at least twice as much implicitly as it was mentioned explicitly. Eight propositions were identified as implicit. They are listed in Table 5. Note that in this table we counted the total number of times that a proposition was mentioned, including the multiple times that a proposition was mentioned by the same subject. With the exception of the last two propositions, each of the propositions listed is a familiar one. Referring to the content of the listed familiar propositions (see Appendix B), it seems likely that these are no longer consciously reflected upon.

| Proposition | Explicitly Mentioned | Implicitly mentioned |
|-------------|---------------------|---------------------|
| P1          | 0                   | 6                   |
| P7          | 0                   | 8                   |
| P8          | 1                   | 7                   |
| P12         | 2                   | 11                  |
| P13         | 0                   | 6                   |
| P14         | 1                   | 5                   |
| P15         | 0                   | 3                   |
| P16         | 0                   | 3                   |

Apart from the propositions that we wished to convey as learning objectives, students sometimes made use of other propositions. A list of these, that were explicitly mentioned by at least two students, is provided in Table 6. For convenience, we have given these propositions numbers that are consecutive to our own list.

| Proposition | Explicitly Mentioned | Implicitly mentioned |
|-------------|---------------------|---------------------|
| P1          | 0                   | 6                   |
| P7          | 0                   | 8                   |
| P8          | 1                   | 7                   |
| P12         | 2                   | 11                  |
| P13         | 0                   | 6                   |
| P14         | 1                   | 5                   |
| P15         | 0                   | 3                   |
| P16         | 0                   | 3                   |

Table 5. Implicit Propositions

Table 6. Additional Propositions Mentioned by More than One Student
Upon reviewing this list, it became clear to us that P24 and P25 should actually have been taken up in our own list of fundamental propositions. The range and meaning of the values of $r_{xy}$ are so fundamental to a correct understanding of regression analysis that any course, including that of our own, will devote explicit attention to these propositions. We have done so, but overlooked these two propositions when compiling the list of Appendix B.

Lastly, students frequently made use of erroneous propositions. In most cases, these errors were idiosyncratic, but some recurred a number of times. These are listed in Table 7.

Table 7. Erroneous Propositions (F) Mentioned by More Than One Student

| Item 2           |  
|------------------|
| P24: $-1 \leq r_{xy} \leq +1$ (2 students)  
| P25: $|r_{xy}|$ of 1 denotes a strict linear relationship (2 students)  
| P26: The larger the value of $r_{xy}$, the better the prediction of $Y$ on the basis of $X$ (4 students)  
| P27: If $X$ gives a good description of variation in $Y$, than $S_{y'}$, should be approximately equal to $S_{y}$ (2 students)  

| Item 4           |  
|------------------|
| P28: The $Z$-transformation changes the scale values of $X$ and $Y$ (3 students)  
| P29: After $Z$-transformation the scores are expressed as standard deviations from the mean (2 students)  
| P30: The relative amount of variance explained in $Y$ by $X$ does not change after transformation of scale values (4 students)  
| P31: $z$-transformation leaves the value of $r_{xy}$ unchanged (3 students)  

| Item 3           |  
|------------------|
| F1: Irrespective of the chosen loss function, in case of a linear regression model, $r_{xy}^2 = S_{y'}^2/S_y^2$ (2 students)  
| F2: Irrespective of the chosen regression model, when the least squares criterion is met, $r_{xy}^2 = S_{y'}^2/S_y^2$ (6 students)  
| F3: If the least squares criterion has not been used, than the regression model that has been used cannot be linear (2 students)  

Upon reviewing this list, it became clear to us that P24 and P25 should actually have been taken up in our own list of fundamental propositions. The range and meaning of the values of $r_{xy}$ are so fundamental to a correct understanding of regression analysis that any course, including that of our own, will devote explicit attention to these propositions. We have done so, but overlooked these two propositions when compiling the list of Appendix B.

Lastly, students frequently made use of erroneous propositions. In most cases, these errors were idiosyncratic, but some recurred a number of times. These are listed in Table 7.

Table 7. Erroneous Propositions (F) Mentioned by More Than One Student
4. Discussion

The main research question guiding this explorative study was whether it would be possible to reduce a body of complex theoretical knowledge -- in this case descriptive regression analysis -- to a set of constituent propositions. A set of propositions that would, upon successful mastery of the course material, be used by our students to represent statistical problems and to construct solutions of these problems. Although the study made use of only nine students, these did represent the whole spectrum of statistical aptitude (from very poor to very good) and their data yielded support for the importance and key role of the propositions that we initially compiled. Of the 23 propositions, only six were shown to be relatively unfamiliar to our students, which by definition meant that five or fewer students at any time made use of these propositions in their attempts at problem solving. On the other hand, students did make use of propositions that were not in our list, but most of the time such propositions were idiosyncratic. They were used by only a single student. Few alternative propositions were used by more than one student and those that were, were usually used by two or sometimes three students.

What this suggests is that it is possible to deconstruct a statistical theory into its constituent propositions in a nonarbitrary way. Given our learning objectives and the way we present our material in a statistics class, there is seemingly a limited number of ways that the learning material can be decomposed into constituent elements. The list that we compiled as meaningful and elementary will likely correspond to the list that a good student will compile for him or herself.

The value of constructing a list of constituent propositions, in advance of the actual process of teaching, is that this helps us clarify our learning objectives to ourselves. Our goal is to convey these propositions to our students and to show the way these propositions relate to each other. The list can also be of help in the assessment phase. The total set of relevant propositions can be checked on the extent to which they are covered by the exam questions to ensure good content validity. Also, we can manipulate the complexity of exam questions by constructing items that will require the students to relate a number of propositions to each other.

With regard to our second research question, our data show that the deconstruction of a statistical theory into constituent propositions may also help us identify which parts of our teaching have not come across sufficiently well. A first indication of possible weak spots can be derived from an examination of the number of times that a given proposition was used by our students. Two propositions were mentioned by only five students, four other propositions were mentioned only once or not at all. Of these unfamiliar

| Item 4 |  |
| --- | --- |
| F4 | After Z-transformation all ratio’s between quantities remain the same (2 students) |
| F5 | Two students concocted an incorrect formula for $S_e^2$ |

| Item 5 |  |
| --- | --- |
| F6 | In case of a linear regression model, $r_{xy}^2 = S_{y'}^2/S_y^2$ (3 students) |
| F7 | $r_{xy}^2$ gives the maximum proportion of variance in $Y$ that can be explained by $X$ on the basis of any possible regression model, linear or otherwise (2 students) |
propositions, three pertain to the role of the least squares criterion in regression analysis, and two to the transformation of scores into Z-scores.

Additional information on problematic points can be gained from a look at erroneous propositions that were used by the students. Of the six different errors that were made by at least two of the students (Table 7 lists seven errors, but two of these are the same, appearing in two different subitems), three directly or indirectly pertain to the least squares criterion and one concerns the effects of the Z-score transformation. The erroneous propositions bearing on the least squares criterion are all associated with P21: "$r_{xy}^2$ gives the proportion of variance explained in $Y$ by $X$ on the basis of a linear regression model when the least squares criterion is used." P21 was identified as an unfamiliar proposition, and it is evident from the errors that most students do realize that $r_{xy}^2$ equals the proportion of variance explained when a linear model is used and that they do often realize that this equivalence holds only when the least squares criterion is used, as is commonly the case. What they failed to consider -- and what they should have observed in order to correctly solve subitems 3A, 5A and 5B -- was that both the linearity of the model and the least squares criterion for the best fitting line are necessary preconditions for $r_{xy}^2$ to be equal to the proportion of variance explained. If we want to stress that the least squares criterion is only one among many criteria for choosing a best line of fit, than this weak spot should be given extra attention in a future edition of the course.

Of course, the fact that 17 propositions were mentioned by most of our subjects in the course of problem solving does not constitute direct evidence that the students have mastered conceptual understanding of the material. As we indicated earlier (Chi, Bassok, Lewis, Reimann, and Glaser 1989; Huberty et al. 1993; Schau and Mattern 1997), there is a difference between propositional knowledge and conceptual understanding, and what we have established here is the existence of propositional knowledge. However, since the propositions were mentioned and used in the course of problem solving, and since 81% of the variation in number of correct solutions to problems was explained by the number of different propositions that students mentioned, this does suggest to us the existence of appropriate schemata.

Looking at the additional propositions suggests that students sometimes have a primarily superficial grasp on the theory. For example, three students used the proposition that a Z-transformation leaves $r_{xy}$ unchanged. Our learning objective is that students should understand that a Z-transformation is a linear transformation and that $r_{xy}$ is invariant under linear transformation of $X$ and $Y$. From this it follows that $r_{xy}$ will be left unchanged after a Z-transformation. The use of this additional proposition, in conjunction with the fact that P15 and P16 were shown to be unfamiliar, suggests that this objective has not been met. A second example is somewhat more problematic. Some students used the additional proposition "The larger the value of $r_{xy}$, the better the prediction of $Y$ on the basis of $X'". This begs the question of why this should be the case. This they should explain by making use of P12, P13 and P14, respectively: $r_{xy}$ is a measure of linear relationship, the less variation there is of points about the best fitted straight line in a scatterplot, the better the prediction of $Y$ on the basis of $X$ will be, and the less variation there is of points about the best fitted straight line in a scatterplot, the higher $r_{xy}$ will be. But whether the use of the additional proposition indicates a superficial grasp of the theory or, conversely, whether this suggests that the students have made P12, P13 and P14 implicit, indicating that they have moved beyond the status of pure novice to a status where some of the important propositions have become integrated into a schema, we cannot decide conclusively on the basis of our data. The fact that P12, P13 and P14 were all shown to be familiar propositions, could be advanced in support of the latter interpretation. Both of the above examples indicate that it is difficult to make an objective distinction -- that is, instead of one that is merely intuitively plausible -- between propositions that are used implicitly as part of more general schemata, and propositions that are not mentioned because they were never properly understood. A firm
distinction could only be based on an interview of the student, explicitly inquiring after his or her interpretation.

As we showed in Table 6 and discussed in the Results section, two or more students mentioned propositions that were not added to our list but the fact that they were mentioned leads us to reconsider this. This shows that the compilation of fundamental propositions basic to a topic of elementary statistics is not a one way process, but rather materializes in interaction with the students. It is worth noting, however, that few propositions have been suggested as worth adding to our list. Although repetition of the study with another group of students and another collection of items is likely to yield some additional fundamental propositions, it is to be expected that the number of additions will be limited.

5. Conclusion

We believe that any topic of elementary statistics can be parcelled out into constituent propositions and feel that as teachers we should start by compiling an initial list. This list contains the propositions that students are likely to use as cognitive units from which to built gradually more elaborate schemata. By probing the work of a limited but representative sample of students, the list can be used by the teacher to check for himself whether he has explicitly covered all the relevant points in his exposition of the material.

Various ways suggest themselves in which the use of lists of propositions, representing the content of a particular course, can help the students in their efforts to build up their knowledge and understanding of statistics. For instance, at the beginning of a course students may be handed out a list containing all the propositions that they need to master, with detailed reference to the literature. The list may serve to them as a checklist to find out whether they possess the elementary knowledge required to build up a more sophisticated understanding of the statistical theory. One might object that such a task will become complicated in the case of a course that covers a wider range of topics than the introductory course on regression analysis that we discussed. In fact, our experience shows otherwise. We have compiled a list of fundamental propositions covering a second year course on analysis of variance. This extensive course, that covers topics as diverse as factorial designs, data from nonorthogonal designs, ANCOVA, and the use of dummy variables in a regression model, was decomposed into approximately 140 (about 25 for each separate topic) different propositions. This initial list was handed out to the students as a checklist of important points to know. Rather than experiencing the multitude of propositions as disheartening, they actually welcomed the fact that they were alerted to the relevant knowledge elements in the material. After all, in the normal case of events, the students have to identify all these elements themselves, and subsequently try to establish relevant links between these propositions. By providing students with a checklist, our experience suggests that we actually reduce the complexity of the task they are faced with.

A variant of the above use of propositions that requires more activity from the part of the students, is to present the list of propositions in the form of questions like "What measure is used to summarize the strength of linear association between two variables?" and "What is the range of values that this measure can take up?", etc., with a separate question for each individual proposition. This way, the students actively have to establish the various propositions themselves, but they are put on the right track by us.

In addition to this exercise, we may present the student with a large amount of worked-out problems, in which the solution is written out in terms of the necessary propositions. A variant on this worked-out problem strategy is to present a number of problems, coupled to the propositions that we believe are necessary for their solution. We may then ask the students to construct an argument on the basis of the provided propositions, that leads up to the correct solution of the problem.
The latter strategy forces students to think of relationships between the various propositions. It is particularly when used in this form that we believe the use of propositions may have potential as a learning tool for stimulating conceptual understanding. Further research is necessary to see whether some of this potential can indeed be realized in a practical way.

Appendix A

The following five multiple-choice items were used in the study. In all cases, the asterisk marks the correct alternative.

1. We have collected data on two variables $X$ and $Y$. We decide to regress $Y$ linearly on $X$. The results show that the error variance $S_e^2 = 1$ and is smaller than $S_y^2$, the amount of variance explained by $X$, and $S_y^2$, the total variance in $Y$. Are these results possible, or has there been some sort of an error in the data analysis?

   a. No, these results cannot possibly be right
   b. Yes, these results are possible in case variable $X$ explains less than 50% of the variance in $Y$
   c. Yes, these results are possible in case variable $X$ explains more than 50% of the variance in $Y$ (*)
   d. Yes, these results are possible irrespective of the percentage of variance explained in $Y$ by $X$

2. In a study, 30 subjects obtained a score on two variables $X$ and $Y$. Regression of $Y$ on $X$ yields the following linear model: $Y' = 0.90 X + 1.5$. In addition, the analysis shows the mean on $Y$ to equal 6.3 and the variance to equal 3. The correlation between $X$ and $Y$, expressed in Pearson’s $r_{xy}$, equals 0.91. The variance of the predicted $Y$-values, $S_{y'}^2$, equals 2.5 and the sum of squares $\sum (y - y')^2 = 15$. Which of the following statements is incorrect?

   a. In conjunction with the other data reviewed above, the value of $S_{y'}^2$ suggests that the linear model describes the relationship between $X$ and $Y$ fairly well
   b. The value of the correlation coefficient $r_{xy}$ suggests that the chosen regression model allows for an accurate prediction of $Y$, based on $X$
   c. The data suggest that no other regression model will yield a value of $\sum (y - y')^2$ smaller than 15
   d. The value of the regression coefficient shows that $Y$ can be accurately predicted from $X$ (*)

3. In the same study, two variables $X$ and $Y$ correlate 0.50 with each other, whereas the same variable $X$ correlates 0.70 with a third variable $Z$. Independently, two bivariate regression analyses are
performed: \( Y \) is regressed on \( X \) and \( Z \) is regressed on \( X \). After analyzing the results it is shown that 
\[ \frac{S_{y'}^2}{S_y^2} \] (the explained variance of \( Y \) divided by the total variance of \( Y \)) equals 0.55, whereas 
\[ \frac{S_{z'}^2}{S_z^2} \] equals 0.49. Which conclusion may be drawn with certainty?

a. A non-linear model has been used for regression of \( Y \) on \( X \) (*)
b. A non-linear model has been used for regression of \( Z \) on \( X \)
c. The prediction of \( Z \) based on \( X \) is more accurate than the prediction of \( Y \) based on \( X \)
d. The researcher has made computational errors, for the results are impossible

4. We regress \( Y \) on \( X \) using a linear equation \( Y' = bX + a \). Suppose that we convert \( X \) and \( Y \) into standard scores \( Z_x \) and \( Z_y \) and then regress \( Z_y \) linearly on \( Z_x \). Which of the following statistical quantities will, in comparison with the original regression of \( Y \) on \( X \), have remained unchanged?

a. The value of the error variance
b. The intercept of the regression line
c. The slope of the regression line
d. The proportion of variance in \( Y \) explained by \( X \) (*)

5. Two variables \( X \) and \( Y \) have a correlation of \( r_{xy} = 0.80 \). Amongst other results, regression of \( Y \) on \( X \) yields the following: 
\[ \frac{S_e^2}{S_y^2} = 0.50 \] Which of the following conclusions can now be drawn?

a. This results indicates that use was made of a nonlinear regression model
b. The regression equation that was used does not satisfy the least squares criterion (*)
c. It is possible to come to a better prediction of \( Y \) based on \( X \), but only in case a nonlinear regression model will be used
d. Each of the alternatives listed above is correct

Appendix B

Overview of the 23 propositions that were considered to be the cognitive units that together make up the theory of descriptive regression analysis, as taught to first year psychology students at Maastricht University. Omitted in the list are propositions on residual plots and on the role of outliers, which did not figure in any of the five items above. The students were not required to derive any of the definitional formulae, so these are also not included as propositions.

1. The intercept gives the value of \( Y' \) when \( X \) is zero
2. The regression coefficient \( b \) pertains to the slope of the best fitted straight line
3. The regression coefficient $b$ does not give any information on the amount of variation of points about the best fitted straight line.

4. $S_{y'}^2 = \text{variance in } Y$, explained by $X$.

5. $S_e^2 = \text{variance in } Y$, not explained by $X$ (error variance).

6. $S_y^2 = S_{y'}^2 + S_e^2$.

7. $0 \leq S_y^2 < \infty$.

8. $0 \leq S_e^2 \leq S_y^2$.

9. $S_{y'}^2 / S_y^2$ gives the proportion of variance in $Y$ explained by $X$ on the basis of the chosen regression model.

10. $S_e^2 / S_y^2$ gives the proportion of variance in $Y$ left unexplained by $X$.

11. $S_{y'}^2 / S_y^2 + S_e^2 / S_y^2 = 1$.

12. The less variation there is of points about the best fitted straight line in a scatterplot, the better the prediction of $Y$ on the basis of $X$ will be.

13. The less variation there is of points about the best fitted straight line in a scatterplot, the higher $r_{xy}$ will be.

14. $r_{xy}$ is a measure of linear correlation.

15. $r_{xy}$ is invariant under linear transformation of $X$ and $Y$.

16. A $Z$-transformation is a linear scale transformation.

17. $z$-scores have a mean of zero.

18. $z$-scores have a standard deviation of one.

19. The criterion most often used to decide which straight line best describes the relationship between $X$ and $Y$ in a scatterplot is the least squares criterion.

20. The least squares criterion makes sure that the error variation $\sum (y - y')^2$ is as small as possible.

21. When the least squares criterion is met, $r_{xy}^2$ gives the proportion of variance explained in $Y$ by $X$ on the basis of a linear model.
22. The greater the proportion of variance explained by a regression model, the better the prediction on the basis of that model.

23. The proportion of variance explained in \( Y \) by \( X \) on the basis of a nonlinear model is greater than or equal to the amount of variance explained in \( Y \) by \( X \) on the basis of a linear model.

References

Broers, N. J. (in press), "Selection and Use of Propositional Knowledge in Statistical Problem Solving," *Learning and Instruction*.

Bromage, B. K. and Mayer, R.E. (1981), "Relationship Between What is Remembered and Creative Problem-Solving Performance in Science Learning," *Journal of Educational Psychology*, 73 (4), 451-461.

Chance, B. L. (2000), "Components of Statistical Thinking and Implications for Instruction and Assessment," *Proceedings of the American Educational Research Association*.

Chervany, N., Collier, R., Fienberg, S., Johnson, P., and Neter, J. (1977), "A Framework for the Development of Measurement Instruments for Evaluating the Introductory Statistics Course," *The American Statistician*, 31, 17-23.

Chervany, N., Benson, P. G., and Iyer, R. (1980), "The Planning Stage in Statistical Reasoning," *The American Statistician*, 34, 222-226.

Chi, M., Feltovich, P. J., and Glaser, R. (1981), "Categorization and Representation of Physics Problems by Experts and Novices," *Cognitive Science*, 5, 121-152.

Chi, M., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989), "Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems," *Cognitive Science*, 18, 145-182.

Clayden, A. D., and Croft, M. R. (1990), "Statistical Consultation - Who’s the Expert?," *Proceedings of AI and Statistics* 2, 65-76.

Cobb, G. (1993), "Reconsidering Statistics Education: A National Science Foundation Conference," *Journal of Statistics Education* [Online], 1(1). (http://ww2.amstat.org/publications/jse/v1n1/cobb.html)

Einhorn, H. J., and Hogarth, R. M. (1981), "Behavioral Decision Theory: Processes of Judgment and Choice," *Annual Review of Psychology*, 32, 53-88.

Garfield, J. B., and Gal, I. (1999), "Assessment and Statistics Education: Current Challenges and Directions," *International Statistical Review*, 67, 1-12.

Greer, B., and Semrau, G. (1984), "Investigating Psychology Students’ Conceptual Problems in Relation to Learning Statistics," *Bulletin of the British Psychological Society*, 37, 123-125.

Hinsley, D., Hayes, J., and Simon, H. (1980), "From Words to Equations: Meaning and Representation in Algebra Word Problems," in P. Carpenter and M. Just (eds.), *Cognitive Processes in Comprehension*. 

Page 17 of 18
Hillsdale, NJ: Lawrence Erlbaum.

Huberty, C. J., Dresden, J., and Byung-Gee, B. (1993), "Relations Among Dimensions of Statistical Knowledge," Educational and Psychological Measurement, 53, 523-532.

Kelly, A. E., Finbarr, S., and Whittaker, A. (1997), "Simple Approaches to Assessing Underlying Understanding of Statistical Concepts," in I. Gal and J. B. Garfield (eds.), The Assessment Challenge in Statistics Education, Amsterdam: IOS Press.

McNamara, T. P. (1994), "Knowledge Representation," in R. J. Sternberg (ed.), Thinking and Problem Solving, San Diego: Academic Press.

Moore, D. S., and McCabe, G. P. (1999), Introduction to the Practice of Statistics, (3rd ed.) New York: Freeman.

Newell, A., and Simon, H. A. (1972), Human Problem Solving, Englewood Cliffs, NJ: Prentice-Hall.

Novak, J. D., and Gowin, D. B. (1984), Learning How to Learn, New York: Cambridge University Press.

Schau, C., and Mattern, N. (1997), "Assessing Students’ Connected Understanding of Statistical Relationships," in I. Gal and J. B. Garfield (eds.), The Assessment Challenge in Statistics Education, Amsterdam: IOS Press.

Sternberg, R. J. (1996), Cognitive Psychology, Fort Worth, TX: Harcourt Brace.

Tversky, A., and Kahneman, D. (1974), "Judgment Under Uncertainty: Heuristics and Biases," Science, 185, 1124-1131.

Van Merrienboer, J. J. G. (1997), Training Complex Cognitive Skills, Englewood Cliffs, NJ: Educational Technology Publications.

Weiser, M., and Shertz, J. (1983), "Programming Problem Representation in Novice and Expert Programmers," International Journal of Man-Machine Studies, 19, 391-398.

Wild, C. J., and Pfannkuch, M. (1999), "Statistical Thinking in Empirical Enquiry," International Statistical Review, 67, 223-265.

Nick J. Broers
Department of Methodology and Statistics
Maastricht University
P.O. Box 616
6200 MD
Maastricht, The Netherlands
nick.broers@stat.unimaas.nl