Directed Self-Explanation in the Study of Statistics

Nick J. Broers
Maastricht University
The Netherlands

Marieke C. Mur
Maastricht University
The Netherlands

Luc Budé
Maastricht University
The Netherlands

Abstract

Constructivist learning theory has suggested that students can only obtain conceptual understanding of a knowledge domain by actively trying to integrate new concepts and ideas into their existing knowledge framework. In practice, this means that students will have to explain novel ideas, concepts, and principles to themselves. Various methods have been developed that aim to stimulate the student to self-explain. In this study, two such methods were contrasted in a randomized experiment. In one condition students were stimulated to self-explain in an undirected way. In the other the stimulus to self-explain was directed. We examined whether the directive method leads to a greater level of conceptual understanding. To assess conceptual understanding we asked students to construct a concept map and to take a 10-item multiple-choice test. The results are somewhat contradictory but do suggest that the directive method may be of value. We discuss the possibility of integrating that method in the statistics curriculum.

Introduction

At the most basic level, statistics is comprised of a wealth of highly abstract concepts. In any elementary statistics course the student is immediately confronted with a vast collection of concepts, ideas, and principles that often have a mathematical connotation (like distribution, standard deviation and mean) and which often lack a clear referent in the experiential world of the student (e.g. multimodality, skewness, kurtosis). Moreover, statistical concepts are sometimes ambiguous, like the meaning of "mean" (see Hawkins, Jolliffe and Glickman, 1992) and counterintuitive (see, for example, the classical studies by Kahneman, Slovic and Tversky, 1982) on the various misconceptions regarding stochastics). Add to this the fact that many students who take statistics classes have little mathematical background and do so out of curricular demands, and it is not surprising that statistics is often approached with dislike and apprehension (see Gall and Ginsburg, 1990). Often, such a negative attitude results in a postponement of studying the material until one or two weeks before the exam, when many students resort to rote learning of concepts and ideas.

But rote learning of the important concepts and principles that make up the body of any statistical theory does not usually lead to the formation of an integrated knowledge network (i.e., to conceptual or connected understanding of statistics). It has frequently been found that a distinction should be made between knowledge of individual concepts of a knowledge domain and knowledge of the interrelationships between these concepts. The latter, more integrated type of knowledge, is variously referred to as connected or conceptual understanding (Huberty, Dresden and Bak, 1993; Kelly, Finbarr and Whittaker, 1997; Schau and Mattern, 1997), meaningful knowledge (Hiebert and LeFevre, 1986), or principled knowledge (Lampert, 1986). Within cognitive psychology, such integrated knowledge networks are often referred to as cognitive schemata – although this term actually has a broader meaning than conceptual understanding (Marshall, 1995).

Constructivist learning theory assumes that integrated knowledge structures cannot be simply transferred from teacher to student but have to be actively constructed by the learners on the basis of the learning material with which they are presented (e.g. Von Glasenfeld, 1987; Novak, 1998; Mintzes and Wandersee, 1998). In this process, the learners are not a blank slate filing appropriate knowledge, but they bring along a range of intuitions and conceptions with which they approach the study material. From

a scientific point of view, such preconceptions are often misconceptions. However, many of these misconceptions are not senseless ideas but rather are intuitively plausible constructions that have been shown to be valid on previous occasions. Such ideas become misconceptions because they are used beyond their natural limits of applicability (Smith, di Sessa, and Roschelle, 1993; Mevarech, 1983). When confronted with the over-generalized use of their preconceptions, learners may adopt scientifically more correct conceptions instead.

This perspective on learning makes clear that students will not be able to construct an integrated knowledge network by passive absorption of concepts and ideas that are presented in a lecture or outlined in a book. Rather, they need to reflect on what is presented to them, to experience that some of their intuitions are wrong and to actively try to comprehend just why they are wrong and how alternative conceptions will prove to be right.

The Importance of Self-Explanation

In other words, a key activity in learning is self-explanation. There is now a large body of research on the benefits of self-explanation and on ways to stimulate this activity. Chi, Bassok, Lewis, Reimann and Glaser (1989) conducted a pioneering study on self-explanation . They trained eight students in mechanics by providing them with a standard physics text. After this basic training, all eight students were given a number of worked-out problems on mechanics. They were asked to think aloud as they studied these worked-out examples. Subsequently, the researchers administered a posttest on mechanics problems. On the basis of this posttest, the researchers considered four students to be successful problem solvers and four students as poor ones. Subsequent analysis of the think aloud protocols revealed that the successful students had generated a far greater amount of self-explanation than the unsuccessful students.

In a subsequent study, Chi, DeLeeuw, Chiu and LaVancher (1994) tried to demonstrate a positive self-explanation effect by conducting an experiment. Twenty-four students were given a biology text on the human circulatory system. Of these, 14 students received a prompt to self-explain after reading each individual line of the text. The control group, consisting of the remaining 10 students, simply received the instruction to read the same text twice. Administration of a posttest showed that the self-explanation group had made greater progress than the controls and especially did better on questions that required knowledge inferences and use of common sense knowledge. Analysis of the content of the self-explanations showed that 30 percent of these were produced by integrating new information with prior knowledge by the student, and 41 percent of the self-explanations constituted the integration of new information with prior sentences. The better explainers, moreover, frequently integrated new information with preceding information pertaining to a slightly different topic in the same text (Chi et al., 1994).

Since the initial work of Chi et al. (1989), the beneficial effects of self-explanation have been demonstrated in various studies (see e.g. Pirolli and Recker, 1994; Ferguson-Hessler and de Jong, 1990); Webb, 1989). In view of the importance of self-explanation, attempts have been made to directly stimulate students to self-explain. In one experiment, Renkl (1995) had students study worked-out examples and led them to expect that they would later have to explain similar problems to a fellow student. He predicted this teaching expectancy to result in a greater amount of self-explanations, but, in fact, the effect of this teaching expectancy was negative rather than positive: the induced stress led to reduced motivation.

Stark (1998, quoted in Renkl, 1999) tried a different approach. He presented worked-out examples in which part of the solution was replaced by question marks. This way, it was believed, the student would be forced to self-explain. Indeed the number of self-explanations strongly increased in comparison to a control group, but the incomplete solutions introduced problematic gaps of comprehension and also self-explanations that were clearly incorrect but provided the student with an illusion of understanding.

A different technique for stimulating students to self-explain involves asking them to construct a concept map on the basis of a collection of concepts. One form of a concept map is a graph depicting ovals that are connected by arrows. The ovals contain concepts (such as "mean" and "random variable") connected by arrows that are accompanied by a short comment describing the relationship between two concepts (e.g. "is a" could be a comment next to the arrow connecting "mean" with "sample statistic."' By drawing a concept map, a student externalizes the links and relationships he or she perceives between a number of concepts. Bulmer (2002), Knypstra (1999) and by Schau and Mattern (1997) provide examples in which concept maps have been used for instructional purposes within statistics. Schau and Mattern presented students with a concept map in which a number of the ovals were empty, together with a list containing concepts from which the students had to select the relevant ones to be placed in the ovals. Schau and Mattern discussed this use of the concept map mainly as a procedure to assess connected understanding, but it is clear that the assessment method they used has the effect of stimulating students to self-explain.

Directing the Stimulus to Self-Explain

Most of the methods designed to stimulate self-explanation have in common that the stimulus to self-explain is undirected. Creating a concept map will have the student think about important connections between concepts, as will presenting with worked-out examples. However, different students may focus on entirely different connections or principles when working on concept maps or worked-out examples.

Broers (2002) has outlined a method to direct students in their self-explanation activity. Basically, this method is comprised of several steps. First, the instructor has to deconstruct the learning material of a given knowledge domain into a finite number of elementary propositions, which, together, cover all the relevant concepts and principles. For example, when presenting material on the theory underlying hypothesis testing, relevant propositions will state what we mean by a null hypothesis, by a significance level, by a sampling distribution, a test statistic, etc. Second, the learner is presented with a list of study questions that aim to have the student identify all the relevant propositions. For example, a question might read "What do we mean by the null hypothesis?" Another will ask what a test statistic is, and yet another question will ask about the sampling distribution. Each relevant proposition is translated into a study question. Third, the instructor decides which particular connections between concepts the student should learn. Finally, the instructor constructs a number of true-false questions that can only be properly answered by a student who comprehends the connections between the relevant concepts. The true-false statements are accompanied by a subset of study questions, all pertaining to the concepts the instructor wishes the student to relate. The student is then instructed to create an argument out of the answers to these study questions that logically shows the statement to be either true or false.

We decided to compare the efficacy of this directive method to foster self-explanation with a more traditional, undirected approach. The guiding question of our research was: Is there evidence that the directive method gives rise to a greater amount of connected understanding than the undirective method? In addition, we focussed on an explorative issue. Researchers often advocate that concept maps are appropriate tools for assessing conceptual understanding. However, others have raised critical questions regarding reliability and validity (see Ruiz-Primo and Shavelson, 1996). We wanted to compare the use of concept maps with a more conventional test for assessing conceptual understanding, to see if the former would be manifestly superior in detecting differences in conceptual understanding amongst students.

Method

Participants

Twenty-five psychology students volunteered to participate in the experiment. All volunteers were second-year students who had repeatedly failed to pass the elementary statistics exam and who were motivated to participate by the advertisement on the experiment, which had stated that the research project would provide students with extra training for their final re-exam.

Design

Participants were randomly allocated to one out of two groups, respectively called the *undirected self-explanation* group and the *directed self-explanation* group. We established that the students in the two groups did not know each other and thus were not likely to communicate their different experiences. There were six meetings, three of which were primarily meant as training sessions to acquaint students with the procedures and the material. These meetings focused on concepts related to estimation and will not be elaborated on in this paper. The final three meetings were treatment-related and, as such, were organized separately for groups 1 and 2. In the first of these meetings, students in both groups were provided with a questionnaire containing 18 questions, each question pertaining to an individual concept. Some examples of these questions are listed in Box 1.

1. What do we mean by the "null hypothesis"?
2. What is meant by a test statistic?
3. The value of the test statistic is reported with a corresponding p-value. What is meant by this p-value?
4. Why is this p-value a conditional probability?
5. How can we increase the power of a statistical test?
6. What do we mean by the sampling distribution of the mean?

Box 1. Some Examples of Study Questions on Individual Concepts

Subsequently, students in both groups were asked to construct a concept map illustrating the interrelationships between these 18 concepts. The 18 concepts involved are listed in Box 2.

Null Hypothesis	Test Statistic	Parameter(s)	Power	z
Alternative Hypothesis	Critical Value	Sample Size	P-value	s_x
Sampling Distribution	Type I Error	Sample	t	
Significance Level	Type II Error	Decision	σ_x	

Box 2. The 18 Concepts Presented to Students

At the end of this first meeting, students were provided with homework. For both groups, this consisted of filling out the same questionnaire again, this time not by heart but after consulting the relevant learning material (i.e., literature and lecture notes). The literature consisted of the 4th edition of Moore and McCabe's *Introduction to the Practice of Statistics* (2003).

At the second meeting, all students handed over their completed questionnaires, and we found that each individual had now more or less correctly answered each of the study questions. At this second meeting, new homework was provided. The undirected group was provided with the instruction to actively think about possible interrelationships between the 18 concepts. They were to do this by constructing a new concept map at home. The students had already charted the relevant propositions in the study material concerning hypothesis testing, so they had acquired knowledge about the 18 concepts

in isolation. Now they were asked to study the material in order to comprehend the interrelations amongst these concepts. This stimulus to self-explain was undirected because students could choose to focus on some relations and to ignore others. Which particular relations they focussed on and which relations they ignored was entirely a matter of their personal consideration.

The directed group received 10 true-false statements. Each of these 10 statements was accompanied by a subset (about six) of the study questions the students had answered during the previous homework assignment. The 18 study questions, it should be remembered, pertained to each of the 18 individual concepts. For each of these 10 true-false statements, the students were instructed to construct an argument in which they made use of the answers to the subset of questions that accompanied the statements. This way, they were guided towards self-explanation activity pertaining to the relationships among the pre-selected subset of concepts. They knew the meaning of the individual concepts. Now they had to use their sense of logic to ponder the question of how this set of concepts together determined the validity of the statement that was given. An example of a true-false statement, together with the accompanying set of study questions, is given in Box 3.

Statement: If we make a Type II error, this implies that the probability distribution we used to determine the p-value of our test statistic was not an adequate model of the empirical reality

Instruction: Construct an argument that shows the above statement to be either true or false, and use the answers to each of the following questions in your argument:

Study questions to be used:

- What is a test statistic?
- What conditional probability distribution are we working with, when testing a null hypothesis?
- The value of the test statistic is reported with a corresponding p-value. What is meant by this p-value?
- Why is this p-value a conditional probability?
- What is a significance level?
- What is a Type II error?

Box 3. Example of a True/False Question (with Accompanying Study Questions)

At the final meeting, students of both groups were once more asked to construct a concept map, showing the interrelationships among the 18 concepts. In addition, they were given a 10-item multiple choice test in which they had to apply conceptual or connected understanding in order to identify the correct alternatives.

Material

Concept Maps

During the first meeting, participants were provided with an example of a concept map on meaningful learning (taken from Novak and Gowin, 1984), with some additional oral instruction on how to construct such a map. Next, they were presented with a list of the 18 concepts that were presented in Box 2. They were asked to construct a concept map using each of these 18 concepts that showed the interrelationships among the concepts by drawing an appropriate arrow and writing a small comment alongside it. During the final session, students were again provided with the list of 18 concepts and asked to construct a new concept map depicting as many interrelationships as they could think of. Appendix 1 gives an example of an incomplete and partially incorrect concept map that was constructed by a student during the first session. We scored the concept map in two ways: first, by counting the number of correct relationships specified in the map (i.e., by counting the number of appropriate arrows with correct

comments); second, by comparing the constructed map with an expert map (constructed by the principal author) and counting the number of correct comments on arrows that corresponded with the arrows in the expert map.

Test for Conceptual Understanding

Next, all participants were presented with 10 multiple choice items pertaining to hypothesis testing. The items could not be answered on the basis of isolated knowledge of concepts but required connected understanding to do so. A correct response was scored with 1, an incorrect response with 0. The sumscore on this 10-item test was taken as an alternative measure of conceptual understanding. Some examples of items that were used are shown in Appendix 2.

Results and Discussion

The two methods of scoring the quality of concept maps yielded rather similar scores. For the concept maps created on the final session, the mean score obtained by counting the number of correct and appropriately commented arrows was 11.3 (s = 3.9), while the mean score determined by the number of corresponding arrows with the expert map was 8.4 (s = 3.4). The correlation between these two measurements was equal to 0.80. A possible reason for this high correlation is that the expert who constructed the concept map also judged whether a relationship was meaningful or not. In an effort to obtain a measure with higher reliability, we decided to take the average of the scores. Using these averaged scores, we found the means and standard deviations for the two groups and for the two sessions, reported in Table 1.

Table 1.
Means (and Standard Deviations) of Scores on Concept Map Activities

	Group		
	Undirected (n = 12)	Directed (n = 13)	All students (N = 25)
Session			
First	7.38 (3.3)	8.31 (3.3)	7.86 (3.3)
Last	9.21 (3.4)	10.5 (3.5)	9.88 (3.5)

As would have been expected, a paired analysis of mean scores on the dependent variable indicated a significant improvement between the quality of the concept maps constructed at the first and those constructed at the second (or last) meetings ($t(24) = -3.34$, $p < .01$). Looking only at the results for the concept maps constructed at the last session, an independent samples t-test showed that the directed group did not significantly outperform the undirected group, although the former group did specify slightly more relationships than the latter ($t(23) = -.93$, $p = .36$). Our second dependent variable was the score on the multiple choice test for conceptual understanding. This test was only administered during the final session. Table 2 shows means and standard deviations for both groups on this variable. Both the undirected and the directed groups have a relatively low mean score, probably reflecting the fact that we were working with a selective group of students that had a record of poor performances on previous statistics exams. The means differ significantly from each other at the 5% level ($t(23) = -2.42$, $p < .05$).

Table 2
Descriptive Statistics on the Test for Conceptual Understanding

	Minimum	*Maximum*	*Mean*	*Standard Deviation*
Undirected group (n=12)	2	6	4	1.13
Directed group (n=13)	2	8	5.4	1.66

The fact that the concept map suggests that the two groups do not differ in conceptual understanding, whereas the multiple choice test suggests that they do constitutes a contradictory result that requires critical consideration. A possible explanation may be that the sort of reasoning required by the multiple choice test corresponds directly with the training that the directed group had received. Considering a multiple choice item, the student has to infer a set of relevant propositions on concepts that bear on the statement given and by logical deduction conclude that one of the provided alternatives is correct, whereas the three others are false. It is this process of reasoning that was explicitly prescribed in the directed condition of our experiment. Perhaps, therefore, the two groups do not actually differ in their level of connected understanding but only in their familiarity with using logical reasoning on true-false items.

On the other hand, previous research does not establish whether multiple choice tests and concept maps measure the same aspects of knowledge. Several studies showed consistent correlations between these two types of measurements, while others failed to find such correlations (see the overview by Ruiz-Primo and Shavelson, 1996). The results of this study could be taken to suggest that multiple choice tests do tap a different aspect of knowledge than concept maps do. Maybe the concept maps were too easy, in the sense that some of the relationships could just be learned without real understanding of the subject matter. In this way, the understanding of the undirected group may have been overestimated. In Ruiz-Primo and Shavelson (1996), a study of Baxter, Glaser and Raghavan (1993) is described, which resulted in this conclusion.

Still another factor may have obscured differences in conceptual understanding between our groups. In our experiment, we instructed students to create a concept map on the basis of 18 concepts. This approach may have facilitated the task and thereby hidden the effect of the directed training. Even though the training concerned the learning of connections between concepts and not the concepts themselves, learning of connections can improve knowledge of a concept itself. Conceivably the students in the undirected group might have produced less concepts (and links between them) than the students who received the directed training if they had come up with the concepts themselves (see Ruiz-Primo and Shavelson, 1996). A further critical point concerns the reliability of the concept map scores. Few studies have examined this aspect, but the ones that did, showed low reliability of scores (Ruiz-Primo and Shavelson, 1996).

A final factor that may account for the lack of difference in performance between the two groups on the concept maps is the following: Our instruction for drawing a concept map meant that students started with whichever of the 18 concepts they cared to choose. This meant that some students started with concepts like “hypothesis”, “'parameter” and “test statistic”, which would therefore appear somewhere in the center of their paper, with other concepts like “Type I error”, “power” and “sample” appearing somewhere in the periphery of the map. Conceivably, students are more likely to link concepts that appear spatially close to each other than concepts that appear spatially remote from each other. So in the above example, links between “hypothesis”, “parameter” and “test statistic” would then be more likely than links between “hypothesis”, “test statistic”, “sample” and “power”, although the student may well be able to meaningfully provide such links when pressed to do so.

Apart from the above quantitative analyses and ensuing interpretations, we decided to also take a look at qualitative aspects of our data. For example, did the answers to the study questions (see examples in Box 1) and the arguments that were constructed by the directed students suggest that the assignments had been carried out conscientiously and meaningfully? We had some worries in this respect, due to the selective nature of our participants. The students in our experiment were students who had performed very badly on previous statistics exams, either because they lacked the necessary skills, or the necessary motivation, or both. The possibility that our selected group of students would lack the motivation to invest enough time and effort in the assignments was an issue of concern to us. If students did not seriously attempt to construct an argument on the basis of the provided material, the whole exercise would be meaningless as no self-explanatory activity would result.

Both in the answers to the study questions and in the constructed arguments we found evidence that supported our concern. In the use of the study questions, for instance, students sometimes failed to

answer a question, simply putting down a question mark or nothing at all. Other questions showed answers that were very similar across students, indicating that these answers had simply been copied from the text in Moore and McCabe. Both the failure to respond and the mindless copying of text are indications of insufficient effort. A question like, "What is meant by a test statistic?" is direct and unambiguous enough for the student to think it over and to give a meaningful answer in his or her own words. Open spaces and copied text suggest minimal effort in the completion of this task. Although we encountered various instances in which the above was the case, many other students had considered the questions seriously and responded in meaningful and semantically idiosyncratic ways.

The construction of the arguments also showed a wide variation in the seriousness with which this task was carried out. On the one hand, we encountered students who had made a serious effort to construct a sound argument that included all of the answers to the study questions that were meant to be used. On the other hand, there were students who sometimes did not construct any argument at all but had simply written "Correct", meaning that in their view the given statement was true. Most of the students fell somewhere in between these two extremes, constructing arguments that were incomplete (i.e., did not contain all the answers to the selected study questions) or that contained the answers to the study questions in a mindless or meaningless way (e.g., instead of constructing an argument, a student sometimes wrote down all the answers to the study questions selected and than concluded with "the statement is therefore false."

Overall, we found that our students had put in at least some effort to complete all the assignments. But the effort ranged from very minimal to exemplary. We feel it is noteworthy that even in this case, in which lack of motivation does seem to have played an impeding role, the test for conceptual understanding showed a significant difference in favor of the directed students.

Perspectives for Implementation in the Curriculum

We have presented the results of an experimental study in which we compared two fairly small groups. Apart from the sample size, the homogeneous population that we studied (all very weak students in terms of their record on previous statistics exams), and the inconsistent results of the study also combine to discourage us from making sweeping assertions and conclusions. However, results from previous studies (Broers, 2001; Broers & Imbos, in press) as well as our personal experience with this teaching method lead us to believe that the method advocated has the potential to enrich the statistics curriculum and, indeed, the curriculum of any complex subject in which *connected* understanding is a goal.

Stressing the utility of a method for achieving conceptual understanding naturally raises the question of its usefulness for stimulating other types of statistical understanding. In the literature on statistics education lots of references can be found regarding the distinction between statistical literacy, statistical reasoning and statistical thinking. Somewhat surprisingly, the term conceptual understanding is not often used in a discussion on the dimensions of statistical knowledge. As delMas (2002) noted, the terms statistical literacy, statistical reasoning and statistical thinking are often used interchangeably and sometimes used in different ways by different people.

According to Garfield (2002), "Statistical reasoning may be defined as the way people reason with statistical ideas and make sense of statistical information." This involves making interpretations based on sets of data, graphical representations, and statistical summaries. Much of statistical reasoning combines ideas about data and chance, which leads to making inferences and interpreting statistical results. Underlying this reasoning is a conceptual understanding of important ideas, such as distribution, center, spread, association, uncertainty, randomness, and sampling. Note as a matter of interest that Garfield does make a distinction between statistical reasoning and conceptual understanding, without going into a definition of the latter.

Reviewing a number of articles on statistical literacy, Rumsey (2002) encounters various definitions of and references to statistical literacy (or, as it is alternatively called, statistical competency).

It is sometimes defined as an individual's "...ability to interpret and critically evaluate statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant (...) their ability to discuss or communicate their reactions to such statistical information" (Gal, 2002, p. 2-3). It is alternatively defined as "being able to interpret graphs and tables. Being able to read and make sense of statistics in the news, media, polls, etc." (Garfield, 1999). Snell (1999) defined statistical literacy as "the ability to understand statistical concepts and reason at the most basic level." Based on definitions such as these, Rumsey (2002) concludes her overview of statistical literacy with the statement that literacy primarily reflects data awareness: the ability to understand how data are used to make a decision.

The final important component of statistical knowledge — statistical thinking — has much in common with both literacy and reasoning. Yet, as Chance (2002) puts it, "While literacy can be narrowly viewed as understanding and interpreting statistical information presented, for example in the media, and reasoning can be narrowly viewed as working through the tools and concepts learned in the course, the statistical thinker is able to move beyond what is taught in the course, to spontaneously question and investigate the issues and data involved in a specific context."

What becomes clear from this small overview of statistical literacy, reasoning, and thinking is not only that these categories tend to overlap with each other but also with our idea of conceptual understanding. Yet, whereas attempts to clearly demarcate the other three categories from each other remain somewhat elusive (see delMas, 2002, who discusses a number of different possible demarcations with the help of Venn diagrams), the idea of conceptual understanding as the perception of links between statistical concepts seems less controversial. Why this is so can be discussed on the basis of an observation that we made on questioning students about results of a one-way ANOVA, in which the F-statistic yielded a value smaller than 1. When we asked students if they could make a decision on the acceptance or rejection of the null hypothesis, based on this F-value, but without getting to see the accompanying p-value, only a minority of our better students came to a conclusion somewhat like this: The numerator (MS Between) and denominator (MS Within) of the F ratio are estimators. As such they form random variables with an expected value. Under the null hypothesis, both estimators have the same expected value, equaling the population error variance. If the null hypothesis is not true, MS (Between) has an expected value that is larger than MS (Within). Since both estimators are random variables, MS (Between) can be equal to, smaller than or larger than MS (Within) regardless of whether the null hypothesis is true or not. However, only if F is sufficiently larger than 1 can there be reason to suspect that the null hypothesis maybe false.

The interesting point about this process of reasoning is that it can be taken to be indicative of statistical literacy, reasoning, or thinking, without being able to say that it is primarily the one and not the other. However, per definition, what we see here is a demonstration of conceptual understanding. The students that offer this line of reasoning show that they understand the links between concepts such as estimators, random variables, expected value, and probably between test statistic, null hypothesis, and conditional probability as well. It is by demonstrating conceptual understanding that they reflect statistical literacy and reasoning. What we hope to suggest here, is that conceptual understanding may be an aspect of statistical knowledge that is more basic than statistical literacy, reasoning, and thinking, and because of this, is both easier to define as well as easier to assess.

Apart from being possibly more basic, there is another conspicuous difference between conceptual understanding on the one hand and the other components of statistical knowledge on the other. Considering the examples discussed in Garfield (2002), Chance (2002), Rumsey (2002) and delMas (2002), it seems that where literacy, reasoning, and thinking do not overlap with conceptual understanding, they primarily reflect what cognitive psychologists call procedural knowledge of statistics. Knowing how to make use of your data in order to come to proper inference reflects statistics knowledge as a skill. Underlying this skill is a body of declarative knowledge, the knowledge of a body of concepts and principles and an understanding of the way these various concepts relate to each other. Statistical literacy – reasoning and thinking – as discussed above, all encompass conceptual understanding but involve much more than that. We believe the conceptual understanding that underlies the three

dimensions of statistical knowledge forms the declarative part of statistical understanding, whereas the surplus meaning of literacy – reasoning and thinking – primarily pertains to the procedural part. So, understanding the F-statistic as forming the ratio between two estimators with similar expected values, if the null hypothesis holds, would typically reflect conceptual understanding. Recognizing the relevance and validity of the F distribution for your own dataset as well as for data from different but related research designs reflects a reasoning process that includes but transcends conceptual understanding in demonstrating procedural understanding.

Saying that statistical reasoning includes but transcends conceptual understanding seems to imply a hierarchical relationship between these two types of statistical understanding. Of course, if we were to see the ability to reason statistically as being indicative of both conceptual and procedural knowledge of statistics, this would depict statistical reasoning as a process that is richer than mere conceptual understanding. At the same time, it is possible for a student to possess procedural knowledge without being able to reason properly. This would be the case whenever we have a student who is capable of carrying out a number of necessary steps in performing a statistical analysis, without really understanding the rationale of what he or she is doing. Such a person demonstrates procedural knowledge without the necessary conceptual understanding to underpin it. Such a student has learned recipes for doing statistics.

If we accept conceptual understanding to be a prime target of statistics education, the important question arises of exactly what concepts we want the student to learn and also what particular links between concepts we want them to learn. This is, of course, the sole responsibility of the statistics teacher. The instructor decides the level of abstraction at which the study material is to be taught and also which particular concepts and links between concepts may be taken for granted. To become fully aware of the concepts and interrelationships between these that the teacher wishes the student to learn, it may be helpful to write out a list of all the relevant propositions the instructor wishes to convey in any particular lecture (see Broers, 2001).

Helping students to develop conceptual understanding (i.e., to perceive the interrelationships between relevant concepts) can be done in various ways. However, all methods require a conscious effort by the student to integrate the newly taught material into his or her existing knowledge framework, where necessary by reforming older conceptions and beliefs. We believe that having students explain the material to each other can be valuable in stimulating them to reflect on what they currently understand and where their knowledge falters. Such an activity can be assigned in a very unstructured or in a more structured way. For instance, we could assign two students the tasks of explaining to each other the logic of hypothesis testing, but we could also give them a true-false statement to discuss, like "A p-value of 0.03 demonstrates the falsehood of the null hypothesis." Such a training method, although somewhat intensive and therefore demanding, may yield very interesting data. For example, we might wish to investigate why students either did or did not accept certain arguments brought forward by their fellow students.

The method we have advocated is, of course, very akin to such group activity. Although we have presented it as an assignment to be carried out by individual students in isolation, it can of course be modified to be presented in some sort of group activity form. The main feature or defining characteristic of the method we have advocated, however, is the presentation by the teacher of a subset of propositions that we want the students to learn and to understand. By coupling a particular true-false statement to a particular subset of propositions (knowledge fragments in which concepts are defined or in which relations between concepts are specified), we intentionally direct the student to reflect on a particular body of interrelated concepts. Moreover, we force the student to make use of a particular argument. Without deliberately instructing the student to make use of a prescribed subset of propositions, we would end up with very different arguments on the same true-false statement. All of these arguments may be basically correct, but some students may leave a lot of premises unmentioned because they do not know them, whereas others may leave them implicit as unnecessary to articulate. It would be reasonable to assume that the same assignment would yield different arguments, depending on the audience the student feels he or she has to convince. In theory, our method would yield similar arguments regardless of intended audience as we have prescribed the collection of premises that should be included in the

argument. It should be reiterated though, that in the research we have reported in this paper, students often did not follow the instructions and left several of the prescribed propositions out of the argument.

What role can our method play in the statistics curriculum? Just as worked-out-examples seem to be effective for the development of procedural knowledge in mathematics education, we believe that our method could be of value for the development of conceptual understanding of more verbally articulated theories. Although statistics is a branch of mathematics, many of its knowledge components, especially at the elementary level, can be made intuitively plausible by a verbal exposition (e.g., the central limit theorem and its useful applications). We believe that our method can help students in their efforts to explore the logical structure of these verbal theories. Since many students of elementary statistics have a non-mathematical background, helping them focus on the important interrelations between elementary concepts seems both worthwhile and necessary. If students fail to obtain an intuitive grasp of basic statistical theory, the more ambitious goals of training them to think or reason statistically cannot be attained. So our method could play a role in the statistics curriculum very much like worked-out-examples play in the mathematics curriculum.

In our own university, we have been using a variation of the method discussed in this paper. Working with small groups that are guided by a senior student, we have implemented the method as follows. All students are provided with a list of true-false statements relating to the statistical topic at hand. The senior student is provided with a list of study questions to put forward, coupled to a particular true-false statement. The students read out the true-false statement, after which the senior student presents the study questions one by one. The students answer these questions in turn and try to relate these answers to the true-false statement. When all of the study questions have been put forward by the senior student, the students have enough information to infer whether the statement as given is either true or false. The discussions that this approach yields have invariably been considered by the students as valuable and to be a clear help in their attempts to come to an understanding of the interrelationships between the statistical concepts.

Alternatively, the written form of the method, as discussed in this paper, could easily be handed out to students as homework. We have only recently begun to study the potential benefits of this method, and further research is needed to ascertain whether it holds enough promise and if so, how it could best be implemented in the statistics curriculum.

References

Baxter, G. P.; Glaser, R. & Raghavan, K. (1993). *Analysis of cognitive demand in selected alternative science assessments*. Report for the Center for Research on Evaluation, Standards and Student Testing. Westwood, CA: UCLA Graduate School of Education.

Broers, N. J. (2001). Analyzing propositions underlying the theory of statistics. *Journal of Statistics Education*, *9*(3).*(* www.amstat.org/publications/jse/v9n3/broers.html).

Broers, N. J. (2002). Learning statistics by manipulating propositions. *Proceedings of the Sixth International Conference on Teaching Statistics*, Capetown, South Africa.

Broers, N. J. & Imbos, Tj. (in press). Charting and manipulating propositions as methods to promote self-explanation in the study of statistics. Accepted by *Learning and Instruction*.

Bulmer, M. (2002). A narrated concept map for statistics. *Proceedings of the Sixth International Conference on the Teaching of Statistics*, Capetown, South Africa.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3). (www.amstat.org/publications/jse/v10n3/chance.html).

Chi, M. T. H.; Bassok, M.; Lewis, M. W.; Reimann, P. & Glaser, R. (1989). Self explanations: how students study and use examples in learning to solve problems. *Cognitive Science, 18*: 145-182.

Chi, M. T. H.; DeLeeuw, N.; Chiu, M. H. & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*: 439-477.

DelMas, R. (2002). Statistical literacy, reasoning and learning: a commentary. *Journal of Statistics Education*, *10*(3). (www.amstat.org/publications/jse/v10n3/delmas_intro.html).

Ferguson-Hessler, M. G. M. & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction, 7*: 41-54.

Gal, I. (2002). Adults' statistical literacy: meanings, components, and responsibilities. *International Statistical Review*, Vol. 70 (1): 1-25.

Gal, I. & Ginsburg, L. (1994). The role of beliefs and attitudes in learning statistics: towards an assessment framework. *Journal of Statistics Education* [Online], *2*(2).

Garfield, J. (1999). Thinking about statistical reasoning, thinking, and literacy. Paper presented at First Annual Roundtable on Statistical Thinking, Reasoning, and Literacy (STRL-1).

Garfield, J. (2002). The challenge of statistical reasoning. *Journal of Statistics Education*, *10*(3).(www.amstat.org/publications/jse/v10n3/garfield.html).

Hawkins, A.; Jolliffe, F. & Glickman, L. (1992). *Teaching statistical concepts*. London: Longman.

Hiebert, J. & LeFevre, P. (1986). Conceptual and procedural knowledge in mathematics: an introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: the case of mathematics* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Huberty, C. J.; Dresden, J. & Bak, B. (1993). Relations among dimensions of statistical knowledge. *Educational and Psychological Measurement*, *53*: 523-532.

Kahneman, D.; Slovic, P. & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kelly, A.E.; Finbarr, S. & 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.

Knypstra, S. (1999). Inference as a dynamic concept map. *Kwantitatieve Methoden, 60*: 71-80.

Lampert, M. (1986). Knowing, doing, and teaching multiplication. *Cognition and Instruction, 3*: 305-342.

Marshall, S. (1995). *Schemas in problem solving*. Cambridge: Cambridge University Press.

Mevarech, Z. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*: 415-429.

Mintzes, J. J. & Wandersee, J. H. (1998). Research in science teaching and learning: a human constructivist view. In: J.J.Mintzes, J.H. Wandersee & J.D. Novak (Eds). *Teaching science for understanding: A human constructivist view*. San Diego: Academic Press.

Moore, D. S., & McCabe, G. P. (2003). *Introduction to the Practice of Statistics (4th ed.)*. New York: Freeman.

Novak, J. D. and Gowin, D. B. (1984). *Learning how to learn*. New York: Cambridge University Press.

Novak, J. D. (1998). *Learning, creating and using knowledge: concept maps as facilitative tools in schools and corporations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pirolli, P. L. & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*: 235-275.

Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction, 5*: 21-36.

Renkl, A. (1999). Learning mathematics from worked-out examples: analyzing and fostering self-explanations. *European Journal of Psychology of Education, 14*, 477-488.

Ruiz-Primo, M. A., & Shavelson, R .J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6): 569-600.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10 (3)*.(www.amstat.org/publications/jse/v10n3/rumsey2.html).

Schau, C. & 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.

Smith, J. P.; di Sessa, A. A. & Roschelle, J. (1993). Misconceptions reconceived: a constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2): 115 – 163.

Snell, L. (1999), "Using *Chance* media to Promote Statistical Literacy," Paper presented at the 1999 Joint Statistical Meetings, Dallas, TX. (www.dartmouth.edu/~chance/course/Articles/JSM99talk.html).

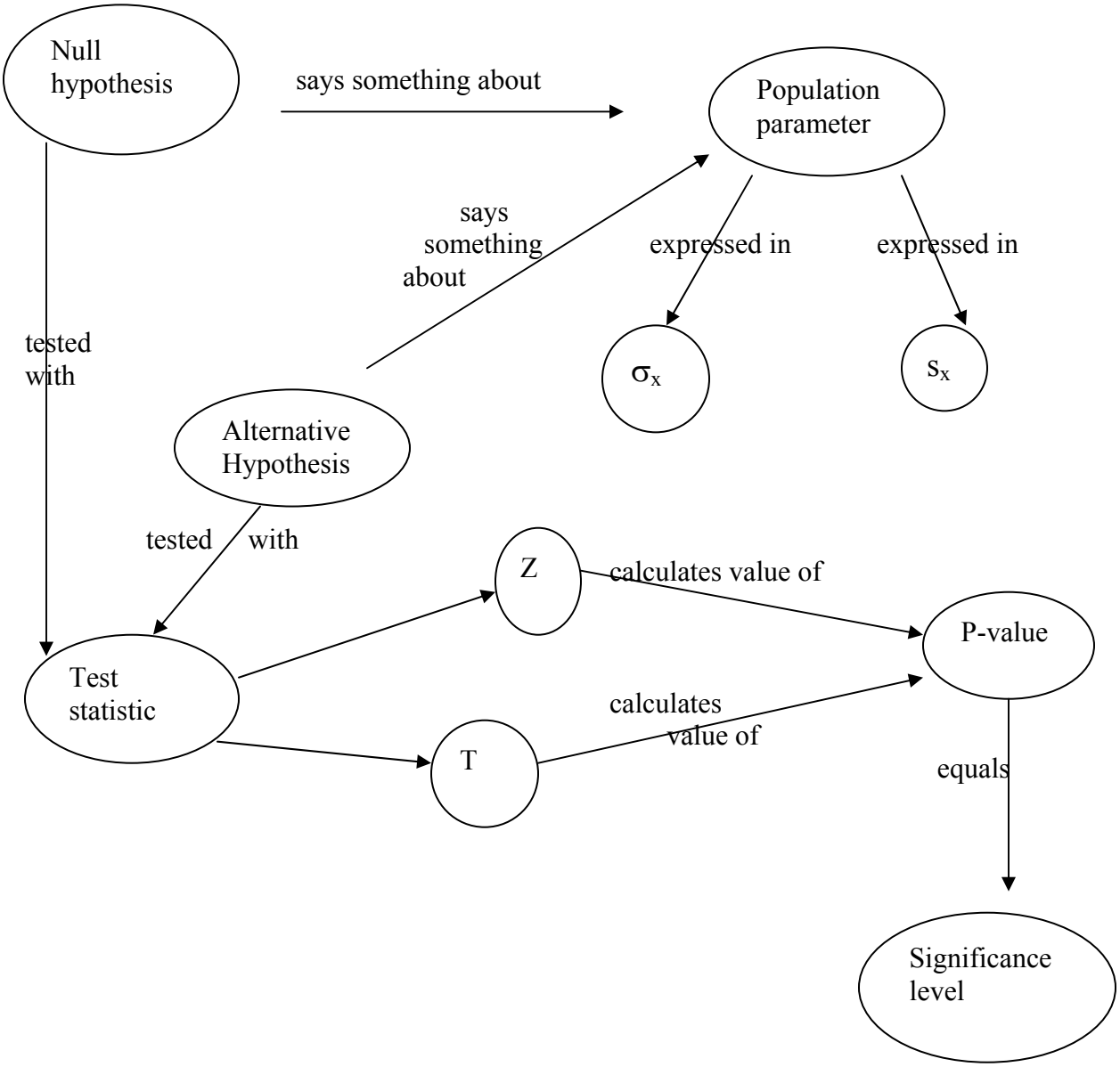
Stark, R. (1998). *Lernen mit Lösungsbeidspielen. Der Einfluss unvollständiger Lösungsschritte auf Beispielelaboration, Motivation und Lernerfolg* [Learning by worked-out examples. The impact of incomplete solution steps on example elaboration, motivation and learning outcomes]. Unpublished dissertation. University of Munich.

von Glasenfeld (1987). Learning as a constructive activity. In: *Problems in therepresentation in the teaching and learning of mathematics*. C. Janvier (Ed.). pp. 3-17. Hillsdale, NJ: Lawrence Erlbaum Associates.

Watson, J. (1997), "Assessing Statistical Thinking Using the Media," in *The Assessment Challenge in Statistics Education*, eds. I. Gal and J. Garfield, Amsterdam: IOS Press and International Statistical Institute.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Education Research, 13*: 21-39.

Paper Appendix 1: Example of a Concept Map (constructed by a participant and not necessarily complete or fully accurate)



Paper Appendix 2: Some Examples of Items from the Test for Measuring Conceptual Understanding (Note that the test will be modified and upgraded before each use to reflect concerns about interpretation.)

Background: The national mathematics exam in 2002 showed an average score of 7 (measured on a 10-point scale). The distribution of exam scores was slightly skewed to the right. Many teachers believe that the 2003 exam was clearly more difficult than that of the previous year. The null hypothesis that the population mean equals 7 is tested against the one-sided alternative that the population mean is smaller than 7. A random sample of 1000 examinees who participated in the 2003 exam is drawn out of the population of students who took the mathematics exam in that year.

Item 3
The value of the test statistic had a corresponding (one-sided) p-value of 0.07. We may conclude that

A the sample mean equaled 7
B the sample mean was greater than 7
C the sample mean was smaller than 7 *
D no conclusions can be derived about the value of the sample mean

Item 6
The researchers constructed a 90% confidence interval for the population mean. Considering the information in item 3, which of the intervals specified below could the researchers have found?

A [7.0 ; 7.1]
B [5.9 ; 6.0]
C [6.9 ; 7.0] *
D [8.0 ; 8.1]

Item 8
Suppose that the researchers had not used a one-sided but a two-sided test of significance. In that case the reported p-value

A would also equal 0.07
B would be smaller than 0.07
C would be greater than 0.07 *
D the information given does not allow a conclusion concerning the resulting p-value

.