10 Conclusions and discussion

Statistics is a little arithmetic and a lot of thinking
Eighth-grade student

The purpose of the present research is to contribute to an empirically grounded instruction theory for early statistics education. An instruction theory, in short, is a theory of how students can be supported in learning a specific topic, in our case the concept of distribution in relation to other statistical key concepts and graphical representations. The contribution of the present study to such an instruction theory is summarized in this final chapter, which consists of a reflective and a prospective component. In the reflective component, we summarize the answers to the research questions (10.1) and other more general results relevant to an instruction theory (10.2). In the discussion of the results, several topics are addressed: methodology, heuristics for Realistic Mathematics Education, computer tools, and symbolizing; and a comparison is made with the Nashville research (10.3). In the prospective component, we make suggestions for a statistics curriculum at the middle school level (10.4). The last section offers recommendations for teaching, instructional design, and future research (10.5).

10.1 Answers to the research questions

The research questions of the present study are:

1. *How can students with little statistical background develop a notion of distribution?*

2. *How does the process of symbolizing evolve when students learn to reason about distribution?*

The first question is answered by summarizing a reconstruction of the hypothetical learning trajectory (HLT) on the basis of what has been learned from the study. This implies an omission of the activities that were not so fruitful within the trajectory (Section 10.1.1). As shown in Chapters 8 and 9, the process of symbolizing can be understood as embedded in the process of diagrammatic reasoning. The second research question is therefore answered by summarizing the key steps in students' diagrammatic reasoning about distribution aspects, and about shape in particular (Section 10.1.2). During the teaching experiments in grade 7, the idea emerged of using the activity of 'growing samples' to support students' development of the notion of distribution. In retrospect, it proved possible to frame the HLT for grade 8 as progressive diagrammatic reasoning about distribution aspects in relation to growing samples. Due to this formulation, the first and second research questions became strongly related. In grade 8, we therefore answer the two research questions simultaneously by answering the following integrated research question:

How can students with little statistical background develop a notion of distribution by diagrammatic reasoning about growing samples?

The answer to this question is summarized in Section 10.1.3.
As a background to the answers, we summarize relevant information about the design of the HLTs. The design of the first HLT was prepared by a historical and a didactical phenomenology (Chapters 4 and 5). The historical phenomenology was carried out to gain ideas for instructional activities and to formulate hypotheses about students' statistical learning. One of the hypotheses was that estimation tasks could support an implicit use of the mean. As part of the didactical phenomenology, we then translated the historical contexts into modern ones so as to be useful for education. In the didactical phenomenology, we identified the core goal of the HLT as extending students' case-oriented views with aggregate views of data. Building on ideas of the Nashville team (Cobb, McClain, & Gravemeijer, 2003), we assumed that a notion of distribution ('shape') could offer a conceptual structure with which students could come to reason about aggregate features of data sets.
The one-sentence summary of the HLT was to challenge students to reason about distribution aspects with increasingly sophisticated notions and diagrams. The initial ideas for the HLT were elaborated in several teaching experiments in grade 7 (Chapter 6) and tested in the last seventh-grade teaching experiment (class 1B; Chapter 7). In the HLT for 1B we initially focused on center and spread as the most important distribution aspects and then factored everything together by discussing growing samples and by making shape a topic of discussion. In class 1B, students came to characterize shapes as bumps and hills, but their reasoning was not as sophisticated as in class 1E (Chapter 8). Because one instructional activity, called growing samples, turned out to be particularly fruitful to foster coherent reasoning about distribution aspects, we decided to take this activity as the recurring theme in grade 8 (Chapter 9).

10.1.1 Answer to the first research question

This section provides a reconstruction of the HLT tested in the last teaching experiment in grade 7. For a schematic overview of the HLT reconstruction see Figure 10.1. The claims have to be taken as anticipatory conjectures that can be tested and revised in practice (cf. Section 3.1).

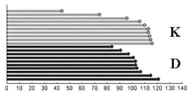
1. Estimate a large number or objects in a picture. When estimating the size of a herd of elephants for instance, students need to find a way to use part of the picture to find the total number. One of the possible strategies is to make a grid and count the elephants in an 'average box', which is the box in the grid representing the other boxes such that the total is accurate. In discussing what an average box is, students learn to look at how values are distributed. The average box is also a sample that is chosen because its size helps to say something about the total. The issue of representativeness can thus be dealt with in an intuitively clear manner. It is possible that students use a midrange strategy of averaging the fullest and emptiest box in the grid. Skewed distributions in other grids can then be used to challenge the midrange as a measure of center and focus students' attention to how values are distributed in relation to the center.

2. Compare different distribution aspects in a value-bar graph. To avoid the so-called 'mean distractor effect', distribution aspects other than the mean have to be addressed (cf. Kelly & Beamer, 1986). The two battery data sets were chosen based on similar means, but different distributions: one is normal and one skewed. In this way, students are stimulated to describe other features of the data sets than the mean. They are likely to reason about extreme values, values that are close together, and the reliability of the brands. In other words, they start to reason about spread issues.

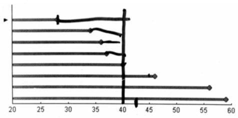
3. Explain what an average box is in a value-bar graph and reinvent a compensation strategy for visually estimating the mean. The first estimation activity, together with students' experience with value-bar graphs, forms the basis for visually estimating means. In this way, students can further develop qualitative and conceptual aspects of the mean such as intermediacy (somewhere in the middle), balancing and compensation (influence by all data values), and representativeness (it says something about the total, it accounts for all data values, and it can be used to compare two data sets).

4. Invent data according to aggregate features and coordination of center and spread in a meaningful context. Inventing data according to aggregate features of battery brands, or anything else, can stimulate students to develop an aggregate view of data. By answering questions such as what an unreliable brand with a long life span looks like, students learn to coordinate center and spread aspects. Students probably use predicates such as 'reliable' in different ways. Some may find brand D more reliable because its range is smaller, but others may find brand K more reliable because it has more of the same value. This reflects two different views on spread.

1. Estimate a large number in a picture. Implicit use of the mean. Discussion of 'average box', mid-range, median in other grids too.

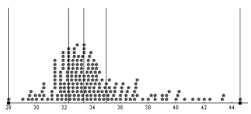


2. Compare different distribution aspects in value-bar graph. Informally discuss center, spread, and skewness aspects in suitable context (long life span, reliability, extreme values, position of the majority).



3. Explain what an average box is in a value-bar graph. towards a compensation strategy. Means represents and accounts for all values.

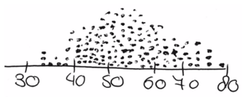
4. Invent data according to aggregate features. Coordination of center and spread in a context (e.g. unreliable brand with a long life span).



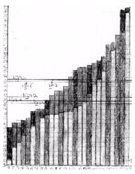
5. Compare distribution aspects of one data set in a value-bar graph and a dot plot. Necessity of concepts and structures underlying both representations. Discussion of spread: range and density views of spread. How are data spread out?

6. Estimate numbers with the mean and discuss sampling methods (e.g. balloon activity). Preparation to growing samples.

7. Characterize a larger sample. Coordination of distribution aspects: mean more precise, more spread out, more around average, few low and high values, shape.



8. Represent various distribution aspects from a story.



9. Compare student graphs of one data set and reason about shape as an object. What happened to bump in the left graph? What happens with bump if sample grows? And if older students are measured?

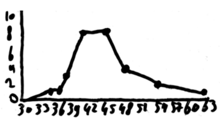


Figure 10.1: Schematic overview of the reconstructed HLT in grade 7

The first is a range view on spread and the second is a local view on spread that we characterize as a density view. In student language, this density view can be expressed as follows, in this example related to a dot plot organized into four equal groups: “Here the dots are spread out, but there they are close together.” Students also combine the two views, for example when they explain how they examine a data set: “I first look at the highest and lowest, and then in between.” It is probably necessary to introduce the notion of range to avoid situations in which students use the term ‘spread’ for the range only. Students are not likely to view spread as dispersion from a measure of center in such contexts as we have chosen, but this could be different in a context of repeated measurements of a ‘true value’ (see 10.4).

5. Compare spread of data sets in a value-bar graph and a dot plot. Describe how data are spread out with respect to organized dot plots. When describing how values or dots are spread out, students probably describe how the data values are distributed. This means that if the instructional design and teaching stimulate students to describe how data are spread out, they can also develop a sense of how data are distributed in relation to the meaning of the context. For instance, “the majority is more to the left” or “in the beginning the dots are less spread out.” Students gradually develop a language in which they can express how data points are distributed, for example in reference to brand K: “in the beginning the steps are large and at the end they are small.” The grouping options in Minitool 2 help students to express differences in density, especially in relation to four equal groups and fixed group sizes (for example in a problem situation such as the speed sign). After a few lessons with Minitool 2, students may be able to explain features such as the following: if the vertical lines of four equal groups or fixed group size are close together, the dots are bunched up or close together, even when data are hidden. This refers to a group of data with similar values. The implication is that they learn to see how the dots are distributed through the abstract diagrams that stem from grouping options in several ways and from hiding the data.

6. Estimate numbers using a notion of average, and discuss sampling methods. The balloon activity is a variant of the elephant estimation problem, but the sampling issue is more complex. In the elephant task, the population is the whole herd visible in the picture, but in the weight context, students need to use their contextual knowledge or simple sampling techniques to find out about students’ and adults’ weights. Again they may use average values as representative values and tools in their reasoning. The balloon activity forms the starting point of the growing samples activity in which students’ initial guesses of weight averages are challenged. One disadvantage of the balloon activity is that it may contribute to the mean distractor effect.

7. Characterize a larger sample. After focusing on center and spread, the HLT aims at reasoning about shape as standing for the distribution of a data set as a whole. During the activity of growing samples, students predict the shapes of large samples (or even populations). In comparing their predictions with diagrams of real data sets, students are expected to reason about aggregate features of the samples and about shape. In Section 10.1.3 we elaborate on the growing samples activity and the importance of distinguishing between 'spiky' shapes of real data and idealized shapes with which experts model data.

8. Represent various distribution aspects from a story. The following example gives an impression of the extent to which seventh graders learned to reason about distribution. In the second assessment task of 1B, students had to draw diagrams that matched with a story on running trainings. From this item we concluded that most students were well able to symbolize case-oriented and many aggregate features of the story into the diagrams. Students can describe how data points are spread out or distributed, but reasoning about shape is probably difficult to accomplish in only twelve lessons.

9. Compare student graphs of one data set and reason about shape as an object. If students make various graphs of one data set they know well, they are likely to understand that these graphs are different representations of the same structure. Using such questions as mentioned in Chapter 8, teachers can support students in reasoning about shapes as objects. However, such reasoning is probably quite challenging to accomplish in grade 7 (see Section 10.1.2).

In short, we claim that students can learn to reason about distribution if an HLT similar to the one reconstructed above is used. This HLT offers an empirically grounded theory of how students may learn to reason about distribution. As mentioned before, an HLT always needs to be adjusted to local circumstances (Barab & Kirshner, 2001). It is likely, for instance, that more lessons are needed than we took, to compensate for the effect of the mini-interviews (see Section 10.3.1) and to avoid the pitfall of addressing too many topics without clearly defining what these topics are (10.5.1). Developing a well-defined terminology took more time than we had anticipated. Although the battery life span and speed sign activities had their merits in supporting particular types of reasoning, it is worthwhile to try out other contexts as well, because these appear to have disadvantages: in the battery context, students do not always expect variation and in the speed sign activity, students may focus on the speed limit as a cut point. Furthermore, the role of the median in the HLT has to be revised (see Sections 10.2.1 and 10.4).

10.1.2 Answer to the second research question

As motivated in Chapters 2 and 8, we used semiotics for analyzing the symbolizing process when students learned to reason about distribution. The semiotic theory most viable for this purpose turned out to be Peirce's semiotics. Compared to the theory of chains of signification, for instance, one major advantage of Peirce's theory of signs is non-linearity. Before summarizing an answer to the second research question, we briefly repeat the semiotic notions that are most important to the analysis.

In Peirce's theory, a sign stands in a triadic relation to an object and an interpretant. The interpretant is the response of an interpreter to the sign. Signs can have different functions depending on how they are interpreted. A sign is a symbol if its relation to its object and its interpretant is formed by habit or rule (and not by similarity for instance). A diagram is an icon representing relations. *Diagrammatic reasoning* involves three steps: constructing a diagram, experimenting with it, and reflecting upon the results. Anything that is thought or talked about is an object in Peirce's theory, and this object is mediated by a sign. From an educational point of view, it is therefore important that the topics of discussion are clear. The process of describing qualities of those topics or objects can be called *predication*. Next, *hypostatic abstraction* is one of the forms of abstraction that Peirce distinguished: a predicate becomes an object in itself that can have characteristics. This is linguistically reflected in the transition from a predicate (e.g. most, lying out) to a noun (majority, outlier). Symbolizing, within this theory, involves making a sign that is interpreted as a symbol, but generally also requires forming the object for which it stands: the symbol of a hill has to stand for an object, which students in general still have to develop (a notion of distribution). The notion of diagrammatic reasoning in combination with that of hypostatic abstraction offers a framework for analyzing the symbolizing process: symbolizing involves both the step of making a diagram (diagrammatization) and forming an abstract object such as distribution (e.g. by hypostatic abstraction). One advantage of using a differentiated notion of sign is that we can analyze students' difficulties with graphs in differentiated ways: we cannot simply say that histograms are difficult. Semiotically, interpreting elements of a sign such as reading off values from a plot is not so difficult, but to interpret the plot as a diagram standing for relations between data or even as a symbol standing for a frequency distribution requires much more conceptual understanding.

Earlier in this chapter we mention that the one-sentence summary of the HLT was to challenge students to reason about distribution aspects with increasingly sophisticated notions and diagrams. We can also conceive this semiotically as progressive diagrammatic reasoning about distribution aspects. In the remainder of this section, we highlight the most important ingredients of the symbolizing process when students learned to reason about distribution as an answer to the second research question.

Predication is a prerequisite for hypostatic abstraction. In all experiments, the estimation, battery, speed sign, and other activities were used to foster a process of predication: describing aggregate features of diagrams and what they represent. The activities aided students in describing features of the data sets with adjectives such as 'average', 'most', 'reliable', 'spread out', and 'low and high values' with respect to signs such as value-bar graphs and dot plots. The objects students talked about were mostly the bars and dots that stood for data values. The most common way of grouping data was into low, average, or high values (C1 in the Appendix). The instructional activities also offered opportunities for hypostatic abstraction, for instance of the average, majority, reliability, spread, and outliers. These notions have thus become distribution aspects with an object character, but these objects were still under development ('immediate objects', in Peirce's terms, not 'final objects').

Coordinating the steps of diagrammatic reasoning and diagrammatization according to aggregate predicates that are mean or spread related. One useful experimentation experience was using the value tool for estimating means in value-bar graphs (as could be done in Figure 10.2) and reflecting upon why this compensation strategy worked. Furthermore, student experimentation with the data sets in Minitool 1 and reflection on the features of the battery brands form the basis for diagrammatization of aggregate features such as "brand A has a long life span but is unreliable." In this way, students learn to think in center and spread-related terms with respect to diagrams, and to extend their case-oriented view with an aggregate view of data.

Diagrammatic reasoning about the bump. According to the HLT, the shape of a distribution had to become an object with aggregate properties. These properties are to represent features of the distributions, which in turn represent properties of the problem situation. To analyze this process and answer the second research question, we focused on the students' reasoning with bumps in class 1E.

The semiotic analysis in Chapter 8 shows that *diagrammatization* can involve multiple actions. Mike, for example, informally grouped the weight data values, used dots at certain positions to signify these groups and the frequencies, and connected the dots to one shape (Figure 10.1 right of item 9). Each of these actions has a history, either in the teaching experiment (e.g., grouping data and using dots) or in mathematics lessons (e.g., line graph).

During the *reflection* on the diagrams, the teacher used the term 'bump' to make this shape the topic of discussion. Students might have first interpreted the bump as an icon (a visual image of a bump), but the analysis shows that the meaning of the bump notion was not just iconic, and even changed from the eleventh to the thirteenth lesson, for several students at least. In the eleventh lesson, students used the bump notion to refer to a group of values close together in the middle part of the graphs. The interesting thing is that the same data set looked so different in various student graphs. By relating the bump in Mike's graph to Emily's graph (Figure 10.1, left of item 9), the teacher stimulated students to specify what the object was which looked like a bump in Mike's graph and a straight line in Emily's. This object, a group of values that were close together, can be seen as a hypostatic abstraction. We can also speak of it as the common conceptual structure underlying aspects of both graphs. To conceive such a structure, however, students need to have a history of reasoning with such plots and similar phenomena (Bakker & Gravemeijer, 2003).

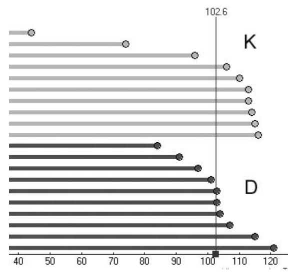


Figure 10.2: Reasoning with the 'bump' in a value-bar graph
("The bump of brand K is higher")

In the twelfth lesson, several students used the term 'bump' for a group of data even when there was no visual bump, for instance when they referred to the vertical straight part in value-bar graphs as a bump (Figure 10.2). This implies that the bump was not just a visual characteristic, but had become a conceptual object and even a reasoning tool, for instance for arguing which battery brand was better.

In the thirteenth lesson, several students referred to 'bump' as the whole shape, whereas before they had only referred to the area surrounding the peak of the hill shape, which for them represented a group of values being close together. The development of the bump as an object was probably stimulated by questions about hypothetical situations in which students needed the bump as an object. When we asked about the shape of a graph from a much larger sample, one student argued that it would grow wider if the sample were bigger because there would be more extreme values. Other students reasoned that the bump would stay the same because there would also be more 'average' values. This suggests that those students had an image of the distribution independent of the sample size. Students also answered the question of what would happen if students of a higher grade were measured. One student said that the bump would be shifted to the right. In that sense, the bump had become an object encapsulating the data set as a whole. Several students were able to relate aspects of that shape to distribution aspects such as average, majority, groups of values, and several acknowledged the stability of the shape across sample size. They even hypothesized on the shape of a large sample, which means they modeled hypo-

thetical situations with a notion of distribution, which was indeed the end goal of the HLT.
The analysis presented here can be taken as a paradigmatic example of explaining symbolizing as embedded in a process of diagrammatic reasoning in which opportunities for hypostatic abstraction occur. We anticipate that modeling in mathematics and science education can be analyzed as diagrammatic reasoning as well.

10.1.3 Answer to the integrated research question

This section provides an answer to the integrated research question in grade 8 in a more general way than Section 9.8. The eighth-grade teaching experiment was organized around the recurring activity of growing samples to test the conjecture that students could develop a notion of distribution by reasoning about growing samples. The integrated research question was how students with little statistical background can develop a notion of distribution by diagrammatic reasoning about growing samples. As before, we cast the reconstructed HLT in terms of anticipatory conjectures that can be tested and revised in practice. We focus on the lessons in which activities were carried out related to growing samples.

1. Diagrammatization according to group characteristics of larger samples. Because variability is the most fundamental concept in statistics, we argue it is important to choose contexts in which students are likely to acknowledge variability: without such understanding there is no need for taking a sample, computing a mean, measuring spread, or looking at the shape of a distribution. Industrial contexts such as the life span of batteries appear to be unsuitable starting points. The experience in grade 8 demonstrates that much can be revealed about students' intuitions of statistical notions when they design their own methods of sampling. Students presumably choose sample sizes that are too small or want to test the whole population. To challenge those views and to promote attention for aggregate features of data sets, teachers can ask what a larger sample with a specific aggregate feature might look like. This promotes diagrammatization according to group characteristics of data sets but also mental experimentation (a what-if attitude) and reflection on aggregate features and sample size. We asked, for instance, what a larger sample of a good and a bad battery brand would look like but, as mentioned earlier, we would not choose such an industrial context again.

2. Extend the samples to populations and create the need for drawing continuous shapes. In short cycles of growing samples of a familiar context, for instance weight or a less sensitive context, students may be asked to predict diagrams of samples with a specific size and compare those with real samples of that size. In this way, reflection can be stimulated about the diagrams and conceptual aspects of the samples in terms of center, spread, and shape. For larger samples students can use dot plots or continuous sketches and predict the shape of the population distribution. Hill, bell curve, pyramid, and semicircle are among the many possibilities. It depends on the context whether students will acknowledge the skewness of unimodal distributions. In the weight context, students probably do not expect a skewed shape, but skewness can be made a topic of discussion by discussing left and right limits in relation to the mean (as in Section 9.4).

When students reflect on comparing predicted diagrams and real data sets, it is crucial that distribution aspects become clear topics of discussion, so that objects can be formed that can be refined during the remainder of the instructional sequence. In particular, we think of the following distribution aspects: average, low and high values, outliers, range, spread, and shape. Linguistically, transitions should be stimulated from predicates such as 'most', 'lying out', and 'spread out' to nouns such as 'majority', 'outlier', and 'spread'. The formed hypostatic abstractions can be predicated again: for example, "the majority lies between 23 and 35, there is an outlier at 107, the spread is large." Drawing on their context knowledge, students can use these statistical objects as reasoning tools about shapes. However, care should be taken that students do not just mimic their teachers in using particular nouns.

The challenge is to strike the balance between providing space for exploration, participation, and reinvention on the one hand, and guidance towards culturally accepted and precisely defined notions on the other. It can be demanding for teachers to see the potential of students' ideas and supporting students in the next step. At some point, it will be necessary to discuss conventional notions such as range, median, mode, and outliers to avoid confusion between range and spread; between mean, midrange, and median; and between extreme values and outliers. This is preferably done after students have experienced the need for such distinctions themselves.

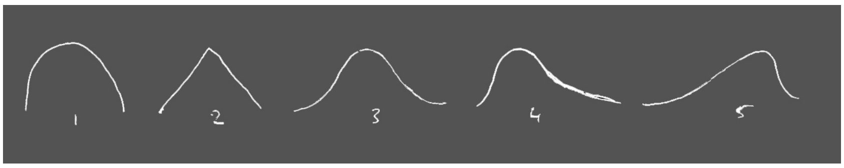


Figure 10.3: Shapes discussed in grade 8

3. Discussion of shapes. To address the distribution aspect of skewness, the different shapes students propose themselves probably need to be expanded with skewed shapes. In grade 8, students discussed the shapes of Figure 10.3 and used statistical notions such as mean and outlier to explain why certain shapes could not represent a weight distribution. During the diagrammatic reasoning about this, the statistical notions of mean and 'outlier', were used as reasoning tools. Students also implicitly reasoned about frequency and density in this phase. To make skewness a topic of discussion as well, we introduced two skewed shapes. Students finally learned to describe shapes and distributions as being uniform, normal, skewed to the left or to the right. This means that shape had become an object that they could reason about with appropriate names.

The next step, which we overlooked in the eighth-grade teaching experiment, is to support students in recognizing shapes in dot plots or other plots in which variation around the smooth curve makes it difficult to perceive the signal in the noise. It is likely that students do not see the same shapes as experts, because they need to develop conceptual structures to perceive those or model the situation with a well-known distribution such as the normal distribution. They also need to experience when it is legitimate to reduce the complexity of a data set with a model (the signal) that includes variation (noise) around it. This implies that also idealized shapes should be made topics of discussion as well as deviations from such shapes, for instance by asking what different data sets (with the same distribution) have in common.

It seems wise to use both case-value plots (such as dot plots) and aggregate plots (such as continuous shapes). Consider this example. Most students in the final interviews could indicate where the mode and mean were in a sketch of a right-skewed distribution, but mistook the median for the midrange. By mentally going back to the two-equal-groups option in Minitool 2, several students could correct themselves and indicate where about the median would be in the continuous sketch. In this way, the measures of center were treated as characteristics of a distribution and not just as outcomes of computations performed on individual data values.

Having sketched the most important steps in the reconstructed HLT, we now discuss students' notions of distribution. One of the end goals of the HLT was that distribution would become an object-like entity. In Chapter 5, we argue that a distribution is more like a composite unit than an object with a procedural and structural side and wonder what the procedural side of a distribution would be. In describing what a distribution was, students often used the term 'spread out': distribution is how the dots are spread out over the graph. From the way students talked about distribution, in particular during the final interviews, we inferred that they imagined the process of distributing dots over the variable as if growing a sample in a dot plot. We conjecture that such a process view of distribution could well be the procedural side of the concept, but we realize that our focus on growing samples has also fostered this view of distribution.

The analyses in Chapters 8 and 9 show that the reification process of distribution is in fact a complex process that involves many steps of hypostatic abstractions (cf.

Sfard, 2000a). Distribution is a multifaceted notion, the understanding of which requires understanding key aspects such as center, spread, density, and skewness. There even seems to be a reflexive relationship between the development of such characteristics of a distribution and the notion of distribution as an object or a shape: by reasoning about the occurrence of low, average, and high values, students expect a particular shape, and by reasoning about shape, students develop the meaning of distribution aspects such as mean, spread, density, and skewness.
We cannot answer the question of whether distribution had indeed become an object for the majority of the students without specifying what we mean by distribution and by object. The research has in fact yielded different levels of understanding distribution of which we mention a few.

1 We assume that students, before instruction, know what is typical and what is not about specific contexts, and that typical values occur more often than exceptional values. This forms the basis for students' reasoning about distribution in natural contexts. However, students mostly lack the language in which they can express such intuitions.
2 The activities in the beginning of the teaching experiments support students in describing various distribution aspects in informal and context-bound ways, such as the reliability of a battery brand in relation to diagrams (predication). Predicates such as 'most' and 'lying out' can become hypostatized as 'majority' and 'outlier', though the meaning of such terms is still under development.
3 Students then learn and use statistical terms for such aspects as range, spread, median, distribution shapes, and skewness in relation to various diagrams.
4 After five lessons, most students in grade 8 were able to express that there were few low and high values and many around average, which could be seen from the relative height in continuous shapes.
5 However, it appears to be much harder to recognize ideal shapes (as signals) within the noise of a 'real' data set as represented in a dot plot (lesson 8). Students reasoned *about* shapes (e.g. in lesson 6) but not *with* shapes as reasoning tools to solve other statistical problems. Moreover, they found it hard to predict the shape of data sets of hypothetical situations such as train delays.
6 Furthermore, there seems to be a difference in viewing a distribution as a shape that emerges from growing a sample, the presumed procedural side of the concept, and distribution as a statistical object with characteristics such as range, spread, mean, median, and mode. In the latter case, shape is an object that can be mentally manipulated and that can be used to predict and model new situations. Students conceiving a distribution in the former way need not understand it in the latter way.

Though students in the teaching experiments did not show understanding of all those levels, we assume that their diagrammatic reasoning experience forms a fruitful intuitive basis for the more technical applications of the normal distribution they will encounter in grades 10 to 12.

10.2 Other elements of an instruction theory

In the previous section, we presented answers to the research questions. These answers were related to the HLTs we used. In our view, HLTs are the most important ingredients of an instruction theory, but there are also more general issues belonging to a developing instruction theory. In the present section we discuss such issues for an instruction theory for early statistics education. We start with the key concepts except distribution because it has already been discussed extensively, discuss the most important diagrams, and finally use the notion of progressive diagrammatic reasoning to integrate students' development of key concepts, diagrams, and language.

10.2.1 Key concepts

This section provides a summary of the most important findings of the present research concerning the key concepts, independently from a specific HLT.

Data

Moore (1990) characterizes data as numbers in context. If there is no close connection between data and context, two things can go wrong. First, if students only see a data set as a batch of numbers, they might be inclined to conceive statistics as 'number crunching'. A possible consequence is that they calculate the mean whenever a question sounds statistical (in contrast see the motto of this chapter). The second thing that can go wrong is that students neglect the data and reason from the context only (as happened in the exploratory interviews). This implies that the norm should be established that the available data should be used when answering a statistical question. The teacher plays a crucial role in establishing such norms and practices. Students should come to understand why they need data and why these data should be created in a proper way to come to an appropriate conclusion. In line with the results of the Nashville team, we stress the importance of talking through the process of data creation as necessary preparation to seeing data as numbers in context. In fact, talking through this process is also a way to address the measurement and sampling issues: what variable exactly is measured and how? However, such guided discussions alone may not suffice; in our view, students should also experience a whole investigative cycle from asking a question, collecting their own data,[58] analyzing

58. A survey in grade 8 showed that students favored the activity of collecting data on car colors because they liked the context and liked doing something.

data to communicating the results and perhaps refining the question and the data collection.

Center

Students can learn to calculate means and medians without too much effort, but this does not imply that they perceive means or medians as measures of center. There is considerable research showing that students generally do not see those values as group descriptors (for an overview see Konold & Higgins, 2003). We have offered remedies for learning the mean: in Chapter 6 we show how students can be supported in coordinating their computational knowledge of the mean with their intuitions of average. The estimation and compensation strategy proved useful for making this connection.

However, the median turned out to be conceptually even harder than we had expected. Like Cobb, McClain, and Gravemeijer (2003), we observed that the median can have two distinct meanings. The first, finding the median of a set of data points, is relatively easy; but the second, the median as a representative value or a characteristic of a distribution, is difficult to develop. One of the difficulties we identified was the following. For the students in the present study, the mean accounted for all the data points, but the median did not. One of the reasons is probably that the median is independent from the values of the differences. Therefore, many students viewed the mean as more precise. It seems especially counterintuitive to students to take a median, which is just order-dependent, in a representation with a rational scale (such as a dot plot).

In search of alternatives for learning a notion of center and measures of center, we have come to the hypothesis that student intuition of an average group or a modal clump can be used to develop a notion of center. Konold and colleagues (2002) argue that such 'modal clumps' can function as measures of center (and spread), and that these mostly encompass mean, median, and mode (at least in unimodal distributions). We assume that it is fruitful to let students indicate the ranges of where they see clumps in the data (Figure 10.4) as a precursor to viewing that part of the data as the center and later using formal measures to locate its position. In particular, the median could be a useful measure of where the clump in a skewed distribution is.

Another way to support students in developing a notion of center, which is not in the spirit of exploratory data analysis, is by focusing on true values and errors in repeated measurements. This is how the mean and normal distribution have been developed historically, and a few researchers have taken this path. Petrosino, Lehrer, and Schauble (2003), for instance, let students measure the height of a flag pole, the length of a pencil, and the height of model rockets in their flight. The activities helped in fostering a notion of a true value (median and mode) and variation of errors. In line with this idea, Konold & Pollatsek (2002) reason that the ideas of signal and noise are useful conceptual underpinnings for seeing an average value as a measure of center or a true value. The true value approach seems to ask for scientific contexts and the clump approach for contexts in which there is no true value. Both contexts appear to have their advantages and disadvantages, but care should be taken in combining them: as is clear from the history of statistics (e.g. Porter, 1986), the transition from variation in a measurement error context to natural variation (such as with height) was quite a leap conceptually.

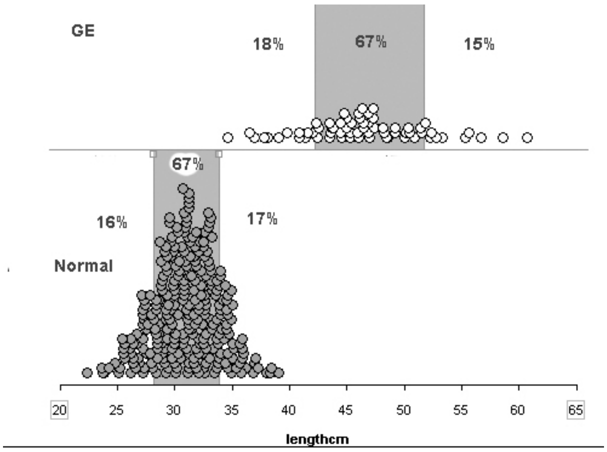


Figure 10.4: Clumps indicated with dividers and percentages (in Tinkerplots); lengths of normal fish and genetically engineered fish (hypothetical data set). The 'clump' of the genetically engineered fish is more to the right (higher values) but also has a wider range.

Spread

In recent years, researchers have come to call for more attention to the notions of spread and variation across all grade levels. The stress on the formal measures of center has probably kept spread in the background and, as a consequence, little is known about students' notions of spread and variation (Meletiou, 2002). One of the challenges of teaching spread is that the only easy way of measuring spread is by the range of a data set (historically the first measure of variation). This, however, is a rather crude measure of spread, just as the midrange is a crude measure of center. The next candidate is the interquartile range, but the present research shows that quartiles take much time to develop (except perhaps for a few *vwo*-students). One of the less formal ways to reason about spread is with the four-equal-groups option in Minitool 2 as a precursor to quartiles and box plots. If a computer tool makes such divisions, students need not be bothered with the precise definition of quartiles. A disadvantage is that students, along with their teachers, may think that exactly 25% of the data values will be in each part, whereas this percentage is rarely exactly 25%. It turned out that the transition to viewing spread as dispersed from a mean or a median was a big leap for these students. For them there seemed to be no reason in the contexts we used to measure differences from the mean, as is necessary for standard deviation. In the context of repeated measurements, this way of conceiving spread probably makes more sense to students, because they can think of possible causes of errors. For their fourth graders, Petrosino, Lehrer, and Schauble (2003) used a so-called 'spread number', which was defined as the median of the absolute differences with the median of the data set.[59]

Sampling

As in the case of spread, sampling receives little attention in most middle school curricula. This may be caused by the lack of numerical calculations that can be trained and tested. Moreover, statistics is taught as part of the mathematics curriculum, which generally focuses on well-defined notions and calculations rather than on statistical reasoning. Another reason could be that sampling is a difficult notion for students to develop. Yet it is important to address the sampling issue, for instance to make sure that students understand how the data were created. The activity of growing a sample is an intuitively reasonable way of addressing sampling. Resampling can later be used to address the variation between samples of the same size.

The present research confirms various results of other studies (e.g. Watson & Moritz, 2000). For example, students in our study indeed did not specify the selection of items and too easily assumed that taking 'a few' gives a fair image of what the question was about. Yet, a little instruction can change this: as it turned out in grade 7 and 8, it was relatively easy to promote the insight that larger samples generally provide a better image but are also more time-consuming and costly than small samples. On a concrete level, students were sensitive to bias: they understood that counting car colors before a (Dutch) post office would yield too high a percentage of red cars. However, promoting the insight that there is variation between samples of the same size proved difficult in grade 8.

10.2.2 Diagrams

In this section we discuss the results concerning various diagram types: value-bar graph, dot plot, histogram, and box plot.

Value-bar graph

Value-bar graphs are not commonly used in statistics and they are only useful for relatively small data sets. Yet there are several reasons to use value-bar graphs in a middle school statistics curriculum. First, students easily interpret value-bar graphs,

59. Assume that 9.1, 9.3, 9.7, 9.8 and 10.6 are the measurements of the height of a flag pole. The median is 9.7 meter, and the absolute differences are 0.6, 0.4, 0.1, and 0.9. The median of these four absolute differences is 0.5 meter, the spread number.

probably because they are already acquainted with bar graphs (with categorical data). Second, the present research shows that value-bar graphs can help students estimate the mean visually and understand the connection between the computation of the mean and its qualitative aspects such as intermediacy, balance, and representativeness. We assume that value-bar graphs provide a better model than the balance model because students at this age probably do not know the physical laws of balance (cf. Hardiman et al., 1984; Pine & Messer, 2000) even though they might have some experience with see-saws. Third, letting students compare different representations of distributions turned out useful; one of the successful comparisons was that of value-bar graphs and mound-shaped graphs (Chapters 6 and 8). Fourth, if students have not already developed a notion of variable, experimenting with the value tool in Minitool 1 and reasoning about the endpoints of value bars may help them develop such a sense of a variable (Gravemeijer, 2000b). Fifth, when students are asked to make a graph of a data set, many of them make value-bar graphs, mostly with vertical bars. We conclude that value-bar graphs are useful in a statistics unit at the middle school level.

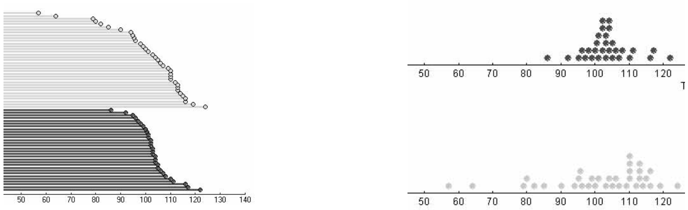


Figure 10.5: Value-bar graph in Minitool 1 and a dot plot in Minitool 2, both of the battery data set used in classes 1B and 1C

Dot plot

Like value-bar graphs, dot plots are easy to read for students but not often seen in statistics textbooks. An advantage of dot plots is that they can represent much larger data sets than value-bar graphs. Though dot plots are sometimes used by professional statisticians (Scheaffer, 2000; Wilkinson, 1999) and provided by commercial educational statistics software packages (Fathom, Tinkerplots), we also heard criticism on dot plots: a dot plot has no vertical axis, which means that the height of a dot has no formal meaning, unless dots are stacked as in Figure 10.6. In an introductory course, in which there is no way of discriminating between frequency distributions and probability density functions, this may also surface as an advantage: students have to construct a meaning to the height of the graph and they can come to see it as an informal measure of density. As a consequence, the shape of a distribution is oftentimes more visible from a dot plot than from a histogram.

We conclude that the dot plot is a useful graphical representation in an introductory statistics course as it allows students to see the individual data points.

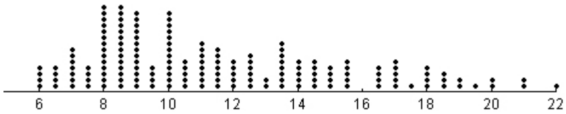


Figure 10.6: Dot plot in which dots are neatly stacked due to rounded values and small dots; thus, height represents frequency

Histogram

The histogram is one of the most commonly used diagrams in statistics. However, it is common knowledge that many students mistake histograms for bar graphs (Baker et al., 2002). For instance, many students misinterpret the height of a bar as the height of people instead of a frequency. This mistake also occurred in the present research in grade 7, despite preparation with dot plots and the fixed-interval-width option in Minitool 2.

In the rationale of the Minitools, there was a learning trajectory from value-bar graphs, to dot plots in which students can organize data points with their own groups and with fixed intervals. Once fixed intervals are chosen, students using the revised version of Minitool 2 can select the histogram option and hide the data points. The jeans activity (7.12) was intended to motivate the choice for fixed intervals, but in grade 7 it did not lend itself so well. In both grades, students preferred very different plots for solving such problems as the jeans problem. In grade 8, some students preferred making their own groups even after we had shown them the options of fixed intervals. We concluded that for them the interval option was not transparent, let alone the histogram option. There were students who preferred to keep the dots in their plots whereas others found the histogram and box plots options without the dots "less busy" or "better organized". This leads us to the conclusion that providing the full range of possibilities gives all students the chance to use a representation they can use as a meaningful reasoning tool (cf. Treffers, Streefland, & De Moor, 1994).

We advise against introducing histograms in early middle school grades for the following reasons. Students in grade 7 needed much time to develop a sense of center and spread, and this took away time allotted for learning the ins and outs of histograms and box plots. Of course, it is not difficult for students to read off particular values from histograms or box plots, but reading off values (elements in a complex sign) is semiotically seen as quite different from interpreting a sign as a diagram (representing relations between data values) or as a symbol standing for one object (for instance of a normal distribution or a frequency distribution). As we argue in Chapter 2, we do not see much use in teaching students notions (mean, median) and graphs (histogram, box plot) that they can only interpret on a superficial level (albeit in an exact way) and not use as meaningful reasoning tools for analyzing data.

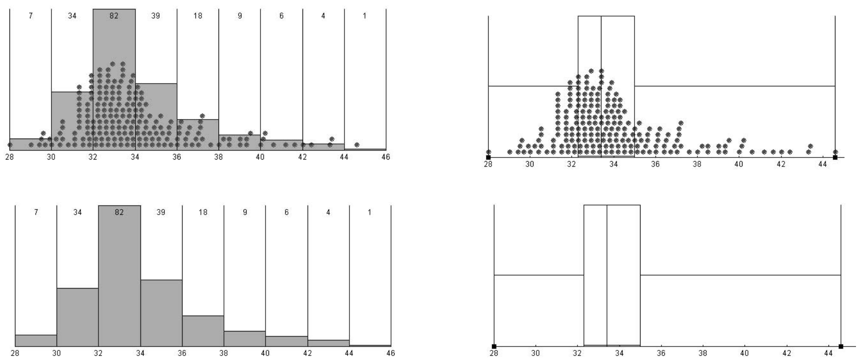


Figure 10.7: Histogram and box plot options in Minitool 2 (without and with hiding data)

Box plot
Box plots are useful representations for summarizing and comparing distributions. For students, however, this type of plot incorporates several statistical notions such as median and quartiles, which are conceptually demanding (Friel et al., 2001). It is therefore no surprise that students have troubles with box plots, even at the high school level (Konold et al., 1997). The pitfall of introducing box plots at the middle school level is that precious time is spent on how to make box plots and interpret superficial features, for instance what the median is. In our view, this time is better spent on doing statistical data analysis with simpler tools that are more likely to become meaningful reasoning tools. We therefore propose not to use box plots at the middle school level. For the high school level, routes as proposed in Chapter 2 can be used (Figure 10.7).

10.2.3 Progressive diagrammatic reasoning

The notion of diagrammatic reasoning turned out to be useful in capturing the integrated development of key concepts, diagrams, and student language. We have added the adjective 'progressive' to the notion of diagrammatic reasoning in analogy to 'progressive schematizing', which is commonly used in the RME theory (Gravemeijer, 1994; Treffers, 1987). What can we learn from the results for an instruction theory for statistics education? In short, it is pivotal to stimulate the steps of diagrammatic reasoning: diagrammatization, experimentation, and reflection. In our view, these are three core steps in learning statistical reasoning. These steps need not occur in a particular order and, in stressing the importance of those three steps, we do not intend to be exhaustive (cf. Wild & Pfannkuch, 1999). In the remainder of this section we discuss these steps in more detail.

Diagrammatization

For the step of making diagrams, we distinguish cases in which students make their own diagrams and cases in which students are offered diagrams, for instance in computer software.

Throughout the research we have noted that it proves productive to let students make their *own diagrams* (cf. the third RME tenet of stimulating students' own constructions). Students' own diagrams are likely to be meaningful to them and the variety of diagrams can be used to discuss different aspects of the data sets or diagrams (see also DiSessa, 2002). However, apart from offering opportunities for reinvention, there also needs to be careful guidance. Otherwise, students might draw many bar graphs, but not many other plots. If students make their own diagrams, they are likely to use diagrams they know, because habits are formed during previous experimenting with them (the interpretant is a response to a sign, and this response is often habitual[60]). Mike's diagrammatization of his bump, for instance, involved grouping data, symbolizing them with dots, and drawing a line through those dots, which are all habits formed in previous experiences. Emily, for instance, symbolized the mean with the value tool with which students in previous lessons had estimated means.

If diagrams are *offered*, it is important that these diagrams are well prepared and relatively easy to interpret. Otherwise, they are unlikely to become tools in students' reasoning and problem solving. The value-bar graphs and dot plots offered in the two Minitools proved relatively easy for students to interpret as representations of data values and relations between those values (such as "few low and high values" or "many around average"). Much of the research literature on visualization and representation turns out relevant to this aspect of diagrammatic reasoning (e.g. Goldin & Janvier, 1998; Janvier, 1987; Kadunz, 1998; Presmeg, 1998).

Experimentation

Experimentation with diagrams can be done physically and mentally. Computer software can be useful for *physically* experimenting with diagrams. Students get the opportunity to dynamically interact with a data set and make different diagrams of that data set. The advantage of this possibility of exploring data sets over learning how to make specific plots in a procedural way is that students can experience that one data set can be represented in different ways and that different features of the data set can be deduced from different diagrams. However, such experience does not emerge automatically. For instance, though histogram and box plot options were prepared in grade 8 with computer tool options and instructional activities, only few students found those representations helpful. More lessons were required for discussing the merits of the different representations than were available.

Apart from physically experimenting with diagrams, students can also *mentally* ex-

60. Cf. the development of instrumentation schemes (Drijvers, 2003, p. 98).

periment with them. This ability is important for being able to choose a suitable representation of the data and anticipate what data look like in a particular type of diagram. In the present research we tried to promote mental experimentation in different ways. One was by letting students report their findings in the role of data analyst (as opposed to just solving a statistical problem). This probably helps them imagine what a data set would look like in a specific type of diagram. Another way of promoting mental experimentation was by asking questions that could promote a 'what-if' attitude. We asked, for instance, what a graph would look like if the sample were much bigger and what would happen to the bump if an older group of students were weighed.

Reflection

The best reflection that we have observed was during whole-class discussions and mini-interviews. It proved particularly difficult to organize reflective discussion in a computer lab.[61] We highlight two important ingredients of the reflection phase: stimulating predication and creating opportunities for hypostatic abstraction.

Predication is the process of describing features of diagrams or what they stand for. By using suitable data sets of appropriate problem situations, students can be guided in the type of predicates they use. In the present research, students predicated values or groups of values in terms of 'lying out', 'staying behind', 'most values', 'the average box', 'high values', and so on. Furthermore, they predicated dots as 'close together', 'bunched up', 'packed', or 'spread out'. They also talked about data sets as consisting of low, average, and high values. The importance of predication is that specific aspects of diagrams can become topics of common attention ('objects') and that distinctions can be made, for instance between range and spread.

One of the ways to create opportunities for *hypostatic abstraction* is to ask students what they precisely mean by terms, for instance during class discussions. It is advisable to use questions that we asked in the mini-interviews, such as the following. What do you mean by spread? What exactly do you mean by 'average'? What do you call 'low values'? What would you take as the 'average box' in this case? Where do you see the majority? In this way, students develop a vocabulary with which they can more clearly define the objects they talk about. Sometimes it is possible to introduce an official name for what students already describe: the median for the middlemost value, the midrange for the mean of the minimum and maximum values, and the mode for the value that occurs most often. In the present research there have been several instances of such reinventions. More common, however, is the necessity to deliberately introduce notions such as range to avoid confusion.

Another way of stimulating hypostatic abstraction is to create situations in which predicates can be used as objects. When the teacher in grade 7 asked what would

61. It would help if the teacher could switch off the screens with one button.

happen with the bump if older students were weighed, several students could operate with the bump as an object by shifting it to the right. Furthermore, it is useful to compare two diagrams of the same data set (such as Mike and Emily's graphs). We can explain this in the following way. If the diagrams, or aspects of them, look different whereas students know that they represent the same data set, there could be a cognitive conflict that needs to be explained (e.g. bump versus a straight part in the diagrams). Then there may be a need, from a student perspective, for a conceptual object that underlies the two diagrams (majority, bump).

This step of reflection links well with research on communication (Cobb et al., 2000; Steinbring, 2000), discourse (Sfard, 2000b), interaction (Dekker & Elshout-Mohr, 1998; Nelissen, 1987), and the proactive role of the teacher in class discussions (McClain, 2000).

10.3 Discussion

In this section we discuss the following topics: methodology, RME design heuristics, symbolizing, computer tools, and a comparison with the Nashville results.

10.3.1 Methodology

The methodology chosen for the present research as to contribute to an instruction theory for statistics education was design research. In this section we discuss a few of its merits and constraints.

By staying close to educational practice (the study was carried out in regular classrooms with their regular teachers), the research gains ecological validity. This implies, as Cobb, Confrey, et al. (2003) write, that the results are already filtered for instrumental effect. Moreover, design research may well help in bridging the gap between theory and practice. However, we sometimes felt the need to gain more certainty about particular issues. For instance, we became interested in the question of whether it makes a difference to students if value bars are presented vertically or horizontally when estimating means or finding medians. Comparative empirical research, within the setting created by the design research, could yield more insight into such questions, but this requires a larger team of researchers. Likewise, we could not reach the same rigor in analyzing students' notions of spread as for instance Watson and Moritz (2000) do in analyzing students' notions of sampling using clinical interviews. Such large-scale interview studies are rich sources for what can be expected when we start instruction to students of a particular age. Nevertheless, they do not prescribe how students' notions can be improved. For instance, we also encountered 'small samplers' in our study, but it turned out relatively easy to convey that larger samples are generally preferable to small samples, but that drawing larger samples takes more time and money. One merit of design research is that it can be used to analyze the development of students' notions when we try to support that development (cf. Frick, 1998). We propose combining design research with

small comparative studies to gain more certainty on crucial decisions within the design (cf. Brown, 1992).
Another issue that reveals both a merit and a constraint of design research is the generality of its results. We contrast it with a method common in educational research to arrive at general results: pose general but dichotomous questions and work those out in concrete situations (e.g. Milo, 2003). One such dichotomous question is whether we should provide students with a model or let them design one themselves (for a discussion see Van Dijk et al., 2003). A significantly better score in one condition may lead to a general result. In the present design research we worked the other way round. We developed activities through a cyclic process of design, trial, analysis, and revision and arrived at activities that worked out well in a specific content domain. We then tried to generalize the results by describing patterns in student learning and the means of supporting that learning. In this way, a subtle answer is given to the general question of whether to provide students with models or let them design diagrams themselves. We should do both, and the interplay is a subtle balance (e.g. in the growing samples activity in Section 9.4). This is where we need domain-specific instruction theories, also with respect to computer tools. As Hoyles and Noss (2003) observe: small changes in technological tools can have major influences on students' learning. Dichotomously set up comparative research does not appear to offer the small grain size to explain the effects of those seemingly small changes in design.
Last we discuss how generalizable or replicable the results of the present research are if we take a learning effect of the mini-interviews into account. The mini-interviews stimulated reflection. Due to the questions we (including the assistants) asked, students were invited to predicate features of diagrams and to explain what they meant by particular words. It is likely that the mini-interviews have not only accelerated the learning process but have also improved its quality. The mini-interview questions were formulated before the lessons and can therefore be taken as part of the HLT. If the HLT is to be used again, the designer and teacher should make sure that questions asked in the mini-interviews are also asked in the instructional materials and during class or small-group discussions.

10.3.2 Heuristics of Realistic Mathematics Education

In this section we reflect on the RME heuristics of historical and didactical phenomenology, and guided reinvention (2.1). What has the historical phenomenology contributed to the design? Many of the historical hypotheses functioned as elements of the evolving HLT. Of the 28 hypotheses we formulated in Chapter 4, we found empirical support for 11. One, H10, needed refinement: the Trial of the Pyx turned out unsuitable because the historical context became an extra obstacle to overcome (Van Amerom, 2002, makes a similar observation). We have not been able to test the remaining 14 hypotheses, which were mainly about median and distribution.

More specifically, the most apparent success was the idea to start the HLT with estimation to support reasoning about average and sampling (H1, H6). The elephant and balloon activities were directly inspired by the historical study. Next, we would probably not have been sensitive to students' use of the midrange as a possible precursor to the mean if we had not known that the midrange was one of the possible precursors of the mean. Instead of avoiding the midrange, as happens in all textbooks we know, we took advantage of its turning up in discussions and used skewed distributions to challenge its use. Thus students became more sensitive to how data values were distributed.

More generally, the historical study turned out to be helpful for looking through the eyes of the students. Instead of taking the precisely defined historical end products as our framework, our own statistical concepts became more 'fluid' again. As designers, we sometimes need to go through a process opposite to reification. This 'deconstruction' process is a first essential step of a didactical phenomenology.[62] The next step is to reconstruct a revised and improved version of the historical learning process that young learners might go through while using the knowledge they have and their ancestors did not have (Section 4.1).

Using the lignification metaphor by Frege, the motto of Chapter 9, we re-address the issue of guided reinvention. On the one hand, we need to provide students enough opportunities to think freely and creatively. This implies that it is beneficial to let students invent their own diagrams and ideas. Activities with a playful element such as the growing samples activity can succeed in engaging students in statistical reasoning (Sections 9.3 and 9.4). Students can work at their own level: in making a diagram they can think of the statistical notions and meanings as they have at hand (cf. final test in 1B). For the software that is used, this means that it should be easy enough for students to make sense of what the software can do. This all comprises the "soft and sappy" reinvention part of guided reinvention.

On the other hand, students need to be guided carefully: their reasoning needs to be formalized, and their notions have to become more abstract and general. Apart from the instructional activities, the instructional software has to be designed carefully. They must provide enough structure and direction to afford level raising (emergent modeling, abstraction, generalization, formalization). And probably the most important guidance comes from the teacher. This is the 'guided' part of guided reinvention. Our experience is: the more autonomy we want to allow students, the more we have to invest in planning (cf. Bakker & Gravemeijer, 2003).

62. By 'deconstruction' we do not mean task analysis, which aims at decomposing tasks into smaller and easier steps, but 'liquefying' the historically reified concepts.

10.3.3 Symbolizing

In this section, we discuss the advantages of using semiotics for analyzing students' symbolizing process, and in particular of using Peirce's semiotics as opposed to chains of significations. Some readers may ask what using semiotics as an instrument of analysis adds to using common sense. By framing episodes of student learning as examples of more general issues such as diagrammatic reasoning or hypostatic abstraction, the results can be generalized and the insights can be applied in new situations. For example, insights into students' diagrammatic reasoning can be linked with insights into students' ways of modeling in mathematics and science education. In this way, the present study gains theoretical validity and contributes to theory development. Apart from this issue of generality, our experience is that the semiotic framework used provided insights that we had probably not gained without it and provided a vocabulary to express insights more precisely than without those notions.

In Chapter 8 we argue why Peirce's triadic sign notion better suits the purpose of analyzing such non-linear processes as learning statistical reasoning than the dyadic sign of Saussure or Lacan. What makes Peirce's notion of sign non-linear is that the interpretant can be the reaction to multiple signs and that it can be the production of multiple new signs. Several researchers have used the notion of chain of signification for analyzing processes of symbolizing. In Chapter 8, we conclude that this notion is too limited for analyzing more complex learning processes, for example, if they compare signs. Instead, we use the notions of diagrammatic reasoning, predication, and hypostatic abstraction to analyze student learning. The adjective 'progressive' was added to diagrammatic reasoning to stress that it should lead to, in this case, reasoning about distribution aspects with increasingly sophisticated notions and signs. Note that it is possible to describe chain-like signification processes with the Peircean notion, because the interpretant can be a new, more complex sign, the interpretant of which can again be a new, more sophisticated sign, and so on. We therefore regard Peirce's theory as superseding (in the sense of Steffe & Thompson, 2000) the theory of chains of signification.

Another advantage is that the notion of symbolizing can be embedded in a superseding framework of diagrammatic reasoning. Symbolizing is literally making a symbol, and this includes both making the sign, for instance by diagrammatization, and forming the abstract object it represents (e.g. by hypostatic abstraction).

One feature of Peirce's semiotics that could be seen as a limitation is its lack of grounding in a psychological theory. Hence, researchers have combined semiotic and psychological perspectives. For instance, Seeger (2001, 2002) has compared the theories of Peirce and Vygotsky, and Van Oers (2000) has proposed a psychosemiotic theory which draws on both semiotic and psychological theories. Moreover, several authors of Anderson, Sáenz-Ludlow, Zellweger, and Cifarelli (2003) have contributed to a social semiotics as their framework. One further research challenge we see is to develop a semiotics for IT applications. Semiotics has been developed for static signs, but signs in IT tools can be dynamic and interactive: if an interpreter dynamically interacts with an IT sign, how should the interpretant be defined? The interpreter's response is tied to the variety of rules of the IT tool that learners need not be aware of. One way to investigate this issue is to unravel the relationship between semiotics and the instrumental approach, which theorizes on how artefacts such as IT tools become instruments for solving mathematical problems in student hands (Artigue, 2002; Drijvers, 2003).

10.3.4 Computer tools

What can we learn from the experiences with the educational statistics software? In this section, we tentatively discuss a few affordances and constraints of educational software such as the Minitools and formulate hypotheses about different types of educational statistics software that have been developed for middle school students.

One attractive feature of the Minitools is that it hardly takes any time to learn the technical aspects of the tools, unlike with computer algebra systems (Drijvers, 2003). However, there are few opportunities for exploring different representations, which we view as a constraint. Throughout the research we have looked for ways to let students make their own graphs and compare different representations of the same data set. When students made their own visual representations, these were generally very similar to the Minitools representations. This implies that software can heavily influence the way in which students make diagrams of situations. It is striking that many students made vertical bar graphs though Minitool 1 only offers horizontal bars. This can be interpreted in two ways: the first is that the applet in its present form is too restrictive; the second is that students were not bound to the representation they had used. We assume the students in our study could have benefitted from using a more expressive tool (in the sense of Doerr & Zangor, 2000).

How can computer tools support diagrammatic reasoning? Computer software such as the Minitools appears most useful for experimentation with diagrams. For example, Minitool 1 with its value tool supported a visual compensation strategy in the present study (but not in the Nashville research). The experimentation experience with particular types of diagrams forms the basis for reflection. In supporting diagrammatic reasoning, computer tools should in our view also offer user-friendly options for diagrammatization.

The three Minitools form a series of small applications (or applets) that are designed for a particular HLT. We characterize such series as route-type software (Bakker, 2002). There are also larger applications such as Fathom, Tabletop, Tinkerplots, and VU-Stat that are not tied to specific HLTs, but that offer a landscape of possibilities for analyzing data. We call this landscape-type software in analogy to Fosnot and Dolk's (2001) notion of the landscape of learning. On the basis of our experience with the Minitools and Tinkerplots we have formulated the following hypotheses, which can be generalized to route-type and landscape-type software.

1 Small applications such as the Minitools are useful for teaching specific issues, such as visually estimating means or preparing the introductions to histograms or box plots if there is no or little time to learn the software. However, there is a risk of a narrow path offering little exploration space for genuine data analysis.
2 It is easier for teachers to guide students in using simple tools and to discuss their reasoning with the tools because of the limited variety. Conversely, if students use a larger application in which they can make many different plots such as in Tinkerplots, it is more demanding for teachers to guide their students.
3 When using larger applications, students will spend more time finding a good representation of the data and on doing genuine data analysis, but perhaps learn less about specific topics that smaller applications can draw the attention to.
4 When using larger applications such as Tinkerplots, students need much time to learn specific features of the software and will not understand many of the plots they produce. Much reflection time is likely to be spent on the meaning of unconventional plots. Hence, students' meta-representational skills (DiSessa, 2002) might improve, but their knowledge about conventional plots could deteriorate when using large applications.

It might be sensible to start with simple tools if students are young and inexperienced in analyzing data, and use larger applications for students who already have some statistical understanding and skills in reasoning with various plots. We assume that the students in our study could well have coped with a larger application such as Tinkerplots. Moreover, Tinkerplots offers the option to gray out options so that smaller environments such as value-bar graphs or dot plots with limited grouping options can be presented to students.

The question arises of what criteria there are for selecting a computer tool. We suggest two. First, is the tool likely to become meaningful to students? Second, can teachers guide students in learning to reason with this tool? Whichever tool is chosen, the instructional activities, assessment, and the way of teaching should be in tune with the tool, and vice versa (Chapter 9; cf. Kanselaar et al., 1999).

10.3.5 Comparison with the Nashville research

In this section, we compare the present research with the Nashville research. After mentioning the practical differences we compare the most important results.

The Nashville research was carried out with one group of students over a two-year period (37 lessons in grade 7 with 29 students, and 41 lessons in grade 8 with 11 students). Because we could not get more than twelve lessons, we carried out several teaching experiments: four in grade 7 and one in grade 8 (between ten and fifteen lessons per experiment). The level of the Dutch *havo-vwo* students was probably a little higher than that of the Nashville students: about 35-45% of the Dutch students attend the *havo-vwo* tracks of education, whereas the Nashville students joined a 'magnet' school which was attended by the top 50% (Cortina and McClain, personal communication, February 2, 2004). The Nashville research was carried out by a large group (Section 2.3), whereas the present research was not. The two teachers in the Nashville research were team members who only gave one lesson per day, whereas the teachers in our study also taught several other classes per day as part of their regular job.

In retrospect, we characterize our teaching experiments as less linear in different ways. We often asked students to make and compare their own graphs and to compare Minitool 1 and 2 representations. As a consequence, we needed another semiotic framework that would allow for network-like analyses of students' learning, especially when comparing representations. In Chapter 8, we argue why the Peircean semiotics best suited our purpose. The Nashville team had much more time to establish the norms and practices they wanted to establish in their class. Thus they were in a better position to be confident that the meanings of words were 'taken-as-shared' before moving to a more advanced issue or tool. This may explain why the Nashville team could describe the collective learning as a chain of signification. Because our situation more captured regular school practice and we had less time per teaching experiment, there was more variety between students within one class. The time span was too short to focus on group processes such as emerging norms and practices. Instead we were interested in students' conceptual development of center, spread, sampling, and distribution.

We now compare the main points of departures and results, which are indicated as P# and R# as in Chapter 2.

P2. We did not restrict our contexts to ones that were *socially important* such as AIDS and CO_2 emission, because that would cut down the number of possibilities drastically. Activities that students participated well in, such as the elephant estimation, the car color activity, and growing samples activity, were often not situated in socially important contexts.

P6. We did not only ask students to *compare distributions*, but also to describe and structure single distributions. One reason was to avoid vertical comparisons of slices of distributions. Another reason was we considered it dull for students to compare two distributions in each activity.

P7, R15. As we argue in Chapters 4 and 5, we did not want to avoid the *mean* and Chapters 6 and 7 show it was not necessary to do so. Moreover, the mean is probably the most used statistic.

P11 and R8. We have not stressed *multiplicative reasoning* because the Dutch students were already reasonably fluent with percentages. Moreover, we tried to avoid a situation in which they would vertically compare slices of two distributions or would just compare percentages left and right of one value. For instance, we changed the question of the speed trap activity to avoid students from focusing on the speed limit as a cutting point.

R1. Talking through the *data creation process* indeed turned out very important to bring the context to life and address the sampling issue. Without a sense of the sampling issue, students often do not understand what the data stand for.

R2, R17. We have therefore paid more and more attention to *sampling.* Though sampling is a complex notion, it also has aspects that can be developed rather easily. For instance, students quickly came to understand that larger samples are mostly more reliable and several sampling and distribution aspects could be coherently addressed by discussing growing samples (Chapter 9).

R5. As in the Nashville research, the initial *case-oriented* views (reasoning about features of data points) were extended with *aggregate* views of data (reasoning about distribution aspects).

R6. Both in the Nashville research and the present research, students came to reason with *shape* as bumps and hills as symbols, not just visual images.

R9. In the present research, students learned to estimate means visually with the value tool, but in the Nashville research this was not the case. We see two possible explanations. First, the Dutch students already had a better understanding of the mean, and second, the activities of the first lessons in grade 7 enabled them to reinvent the compensation strategy.

R12. We found empirical support for the importance of the role as *data analyst.* For instance, when the battery problem was cast into the context of the *Consumentenbond*, the Dutch equivalent of *Consumer Reports*, students gave more and better arguments than in a factory context in which they were inclined to sell an advertising pitch.

R16. As with the Nashville team, we encountered the difficulty of designing good activities to support students' development of the *median as a measure of center*. In line with the Nashville team we assume that students should first have a sense of the center of a distribution before they can measure that center (e.g. a clump in the center of a data set) with a median, and use medians to compare distributions.

In short, the present study supports many of the results found in the Nashville research, but there are also a few differences. It is noteworthy that very similar patterns in students' learning occurred in activities such as the battery life span problems, although the two populations differed in level and educational context. We furthermore designed a few new activities that support students' developments of aggregate views, for example the data invention tasks, comparison of representations, and growing samples activities.

10.4 Towards a new statistics curriculum

In this section, we make suggestions for a new statistics curriculum of 30 to 40 lessons in grade 7 and 8 over a two-year period.

The goal of such a curriculum is that students become statistically literate, in particular they learn to analyze data and communicate about statistical information. To achieve this, students should extend their case-oriented views with aggregate views of data. For describing and predicting aggregate features of data sets, particular statistical key concepts turn out to be indispensable: variability, sampling, data, center, spread, and distribution. The most fundamental key concepts in statistics are variability and uncertainty. Without a sense of the variability of a certain phenomenon (e.g. life span of batteries), there is no reason for students to think of a sample or a distribution either. It is therefore important to choose a problem situation in which students expect variability or acknowledge the uncertainty involved in the context. One suitable problem situation is estimation of large numbers.

The curriculum focuses on the key concepts of data, center, spread, sampling, and distribution (shape) in relation to case-value plots such as value-bar graphs and dot plots. These concepts and diagram types have to become tools in diagrammatic reasoning; for instance, when comparing distributions or when describing stable distribution aspects (for example whilst growing a sample). Though coherent reasoning about these key concepts is favorable, they cannot always be addressed at the same time. We suggest starting with center and spread as the most important distribution aspects (as in Chapter 7), and then address sampling and shape issues. In terms of diagrammatic reasoning, we advocate the following:

1 students make diagrams of data sets and hypothetical situations (diagrammatization);
2 they experiment with diagrams (mental experimentation can be stimulated with what-if questions and physical experimentation is preferably done with a computer tool that allows dynamic interaction with diagrams);
3 they reflect on the diagrams (the teacher and the instructional design are important here).

We would use a computer tool that allows diagrammatization and user-friendly ways of performing genuine data analysis. The Minitools have limited options to diagrammatize (values need to be entered one by one and there is no drawing tool), and when organizing a data set within one Minitool, students can only use one type of plot, such as a dot plot. Tinkerplots (Konold & Miller, 2004), a recently developed construction tool for statistical data analysis, more closely fits our criteria (10.3.4).

Somewhere in the curriculum, students should experience a complete investigative cycle of data analysis from a question, design, sampling, data analysis to communication of the results and formulation of a new or more precise question (cf. NCTM, 2000; Wild & Pfannkuch, 1999). Every now and then, graphs or messages from the media can be used to foster a critical attitude (De Lange et al., 1993).

From the historical phenomenology, it is clear that we should distinguish three fields in which students can learn different aspects of statistical notions: error theory with repeated measurements, natural phenomena with symmetrical distributions, and social contexts with irregular data. Accordingly, we envision different routes of developing notions of center and spread that probably need to be combined.

The first route roughly follows the historical development of error theory in science. It uses students' notions of a true value when conducting repeated measurements of one item. Konold and Pollatsek (2002) argue that students can thus develop a sense of signal and noise, which can be seen as a conceptual underpinning of understanding average values.[63] This is the route that Petrosino, Lehrer, and Schauble (2003) took in a fourth-grade classroom. We expect students to learn to reason about measurements as being around a true value, that errors on both sides are equally likely, and that large errors are less likely than small errors (H13). The mean, which Dutch seventh graders already know as an algorithm, can be used to estimate the true value. The median can be introduced as the middle-most value (in particular if students think that negative and positive errors are equally likely).

Spread then appears as dispersion from the true value (or the measure of center that estimates that true value). We then let students compare situations with small and large errors to make spread a topic of discussion. The range can be made equal to avoid a conflict between dispersion and range. We anticipate that students come to see the majority of the measurements as a clump of data close together, and that there will always be some low and high values. To guide the reinvention of quartiles, we could ask, "between which two values is about half of the errors?" On the basis of the historical development of quartiles we assume that looking at halves is natural to students (H14). Computer tool options such as four equal groups can become ways of quantifying spread and center if suitable contexts are chosen.

Such a repeated measurement approach could be used for combining science and mathematics lessons (Erickson, 2002). Within the Dutch education system, however, this is difficult to accomplish because Dutch students do not receive instruction in physics until grade 8 and chemistry until grade 9. One drawback of this first route of error theory is that is appears to be at odds with the EDA approach.

The second route of developing notions of center and spread starts from students' knowledge of what is typical and what is not in natural contexts with roughly symmetrical and smooth distributions. As we have seen throughout the teaching experiments, students know what typical weights and heights are for their age. Using such

63. More generally, a distribution, a trend, or a model is a way of capturing a signal, the variation around which is considered noise.

familiar contexts, we can help them express a categorization into three groups of low, typical, and high values, both in a natural and a diagrammatic language. In terms of diagrammatic reasoning it is important that students learn to predicate features of data sets that are represented in diagrams. For example, where is the majority of the data values? These majorities, for instance represented as ‘clumps’ in dot plots, can function as an initial notion of center. When students are stimulated to more precisely describe where they see these clumps, as a range, these clumps can come to function as an informal measure of spread as well. If two distributions with the same range have the clumps in the same place, but one clump has a larger range than the other, we expect that students will see that the former has a larger spread than the latter (in Figure 10.4 the clumps have different ranges).

As shown in Chapter 7, it is important to provide students with tools that can help them become more precise in their reasoning. For instance, it turns out useful to introduce the notion of range as distinct from spread to avoid confusion in the discourse and enhance student abilities to express what they see in diagrams. If opportunities appear to introduce conventional definitions (e.g., median, mode), we can take advantage of them.

Using such data invention tasks as we used in the battery context, we can stimulate students to express aggregate features of center and spread in diagrams: what would a reliable battery brand with a short life span look like in a diagram? In such activities, reliability is taken as a contextual basis for a notion of spread, and average life span as a basis for center. With the compensation strategy of visually estimating the mean, students can learn that the mean accounts for all data values and how it is influenced by extreme values. Gradually the mean can come to function as measuring the position of a clump (the center of the distribution), for instance when comparing different data sets. This can be done with an activity similar to the speed trap. A disadvantage of the speed trap context is that the speed limit can be a distracting value that evokes reasoning below and above that limit, whereas we try to foster reasoning with centers of distributions as a whole. When comparing different degrees of being spread out, students can use computer tool options such as making their own groups, equal group size, and four equal groups. It should be emphasized that it makes sense to have a convention of using a standard way, for instance four equal groups (quartiles). As we argue in Chapter 7, students should be stimulated to describe how data are spread out. Doing so, they often describe how data are distributed.

From early lessons onwards, attention should be paid to sampling. In the beginning, also in grade 7, this could be done by letting students collect data of something simple (such as reaction time or car color). Another way of addressing sampling is talking through the process of data creation. A more explicit way of addressing sampling is asking students to design a method of testing something. Next, an activity similar to growing samples can be done to stimulate diagrammatic reasoning about distribution aspects (as in Sections 9.4 and 9.5). The advantage of growing a sample is that it starts from a sample size that students initially find reasonable and ends with the shape of a population distribution. Students expect such features as the mean to be stable after some time, and some may even consider the shape to stabilize. The situation of comparing predictions and real data sets supports students to reflect in aggregate terms, because comparing individual cases is clearly not very helpful.

When reasoning about shapes they propose themselves, in addition to skewed shapes we think need to be discussed, students can use the statistical notions they know and show the understanding they have of shape. As in Chapter 9, we expect students to acknowledge that there are only a few low and high values in most unimodal distributions. This is represented as low horizontal parts in a continuous shape, and the large group of average values is represented as a high flat part in the shape; and of course, there are values in between that cause the slopes of the shape.

In higher grades, we can ask students to describe new distribution shapes, such as normal, bimodal, uniform, and skewed, all represented in dot plots. The data sets should be large enough for students to recognize the shapes and come to see them as signals in noise. Next, students learn to operate with these shapes as objects, for instance mentally shifting them along the axis or predicting shapes of new situations. At some point, students can handle social science contexts in which irregular data are not so easy to model with distributions.

We anticipate that a curriculum, as described here, leads to statistical understanding that forms the basis for a more formal introduction of the normal distribution in higher grades. More generally, we expect that the approach promoted here contributes to statistical literacy (Gal, 2002).

10.5 Recommendations for teaching, design, and research

10.5.1 Recommendations for teaching and instructional design

Statistics chapters in most mathematical textbooks introduce students to the important statistical notions and graphs in what we call a ‘topic-topic-topic approach’. The rationale of such an approach is probably that once students have mastered those statistical notions and graphs, they will have learned statistics or at least be prepared to carry out statistical analyses. In Chapter 2, we criticize this approach of introducing the theory before the application: research in statistics education shows that this mostly leads to disappointing results. The mean, for instance, is more than an algorithm performed on data values. As shown in several chapters, the mean has many qualitative and quantitative aspects. To know the algorithm is certainly not to know when to use the mean or to know how to use it as a group descriptor or a representative value. Similarly, it is not difficult for middle school students to read values from a histogram, but it is demanding to interpret a histogram as a diagram or a symbol of a frequency distribution. Therefore, researchers in statistics education have been in search of ways to engage students in genuine data analysis while offering opportunities to develop statistical notions and graphs as meaningful tools. The present research is within this tradition.

Though the present design research focused on instructional design and students' learning, we noticed that the influence of the teacher could hardly be overestimated. In previous chapters, we mentioned a few issues for the instruction theory in which the role of the teacher is pivotal.

First, *talking through the process of data creation* helps to bring contexts to life, so that students will know what the data values stand for. This process takes time, which implies that one context per lesson is to be preferred over many contexts per lesson (cf. Van den Boer, 2003). One of the core questions is: how do you think the data were collected? As shown in Chapter 9, it can be very informative to let students design an investigation.

Second, the teacher as well as the designer can stimulate students' *roles as data analysts* to ensure that students do not just solve the problem at issue, but also think about ways to communicate the results clearly to others such as decision makers and discuss the statistics at issue. This includes reflection on the conveying power of diagrams (cf. DiSessa, 2002).

Third, good *practices and norms* do not emerge automatically. We have noticed that students do not automatically understand that they should base their arguments on the available data. Students are not always used to listening to each other and asking questions if they do not understand what the teacher or their classmates are saying (cf. Yackel & Cobb, 1996). As a matter integral to the learning process, it is necessary to come to taken-as-shared topics of discussion. Additionally, a what-if attitude ("what would a diagram look like if...?") is only developed if a teacher establishes a practice of asking what-if questions.

Fourth, we observed that class discussions were difficult to organize in the computer lab, because students are easily distracted by what is visible on the screens. A practical suggestion is to have students switch off the monitors when discussing in the computer lab. For creating a common object of attention, we prefer *discussion in the regular classroom* using slides of screen shots or, even better, a computer projector.

From the semiotic analyses, it follows that the steps of diagrammatic reasoning are key elements in learning statistics. Teachers are therefore recommended to stimulate students to diagrammatize aggregate features, to experiment with diagrams in software and in the mind (by establishing a habit of asking 'what-if' questions), and to reflect. The reflection phase includes students precisely describing what they see in diagrams (e.g. clump, majority, bump, but also the shape of the bump) and where they see it (from which value to which value). This works into two directions: students learn to express their thoughts as part of their conceptual development, and teachers can assess students' statistical ideas. The insights acquired are a prerequisite for supporting students in their conceptual development.

Experts such as teachers and researchers may easily interpret students' explanations in a more advanced way than students actually think. For instance, when students first used the term 'bump', we thought they referred to the whole shape, but from the retrospective analyses we concluded that they initially referred to the 'majority' of the data or the top part of the whole shape only. Another tricky notion is 'most'. If students say 'most' they can refer to many different things such as the highest values, the most frequent values, or the largest group. At an early stage, we often unconsciously chose the most plausible interpretation, but a closer look sometimes proved us wrong.

The semiotic analyses show how intimately conceptual development is linked to the development of a suitable natural and diagrammatic language. From a semiotic point of view it is therefore important that the classroom interaction focuses on clear topics of common attention, which are visible in a diagram and described in the natural language, because this is the way objects can be formed (in Peirce's terms, what is talked or thought about is an object). Discussing such objects or topics of common attention provides opportunities for hypostatic abstraction and thus supports students' conceptual development.

Throughout the chapters, we have presented a few design heuristics for instructional design in statistics education. The main goal, in our view, is for students to learn to think in aggregate terms about a data set as a whole. To avoid a case-oriented view, it can sometimes help to "stay away from data," as we refer to in Chapter 6. Related to this, it can help to "ask about forests instead of trees," for instance by asking questions about hypothetical situations (e.g. weight diagram of a group of older students, the shape of a larger sample). Furthermore, comparing multiple representations can, but need not, support students in using conceptual tools. For instance, when asked about the common ground of two different representations of the same data set, students may use and develop statistical notions that help in describing the common feature (such as spread or bump). However, we have also observed that students easily resort to comparing superficial features of the representations (cf. Seeger, 1998, Van Someren et al., 1998). In that case, students do not interpret the signs as diagrams or symbols, but compare iconic or indexical elements of the signs.

What are the consequences of the semiotic analyses for instructional design? Just as in the recommendations for teaching, instructional designers are recommended to stimulate the steps of diagrammatic reasoning and create opportunities for predication and hypostatic abstraction. Our experience is that conveying the core ideas of a hypothetical learning trajectory to teachers is far from trivial, which means that instructional designers are recommended to think of ways of supporting teachers helping their students.

For criteria for selecting statistics software, we refer to Section 10.3.4.

10.5.2 Recommendations for future research

Statistics education research is a relatively young discipline, especially the domain of using computer tools dedicated to *learning* statistical data analysis as opposed to *performing* data analysis. In this section we present a few research challenges for statistics education in which using computer tools is integral to learning statistical data analysis.

As Noss and Hoyles (1996) write, “new technologies – *all* technologies – inevitably alter how knowledge is constructed and what it means to any individual” (p. 106). This raises the epistemological question of how using statistics tools changes students’ learning and the type of knowledge they acquire. This influence may be rather drastic: from statisticians we have heard that the software they use is so important for their way of working and thinking that they often characterize themselves according to the software they use (e.g., “I am a Minitab statistician”).

Computer tools have some evident advantages: students can dynamically interact with large data sets and different graphical representations in a way that is impossible by hand. Using such tools may shift the focus from calculating means and drawing histograms of small data sets to exploring large data sets with multiple representations and ready available means and medians. Furthermore, computer tools such as the Minitools and Tinkerplots offer ways of grouping data into four equal groups and equal intervals that are laborious by hand, but prepare insight into distributions as seen in box plots and histograms.

However, computer tools can easily be almost independent worlds with their own rules and peculiarities. In that case, the software itself needs to be learned before it can effectively mediate between the learner and what is to be learned. This would not be problematic if learning software were part of the curriculum but, in practice, people expect learning with computers to lead to similar knowledge in faster and better ways. It would be more realistic to expect learning and the acquired knowledge to change; how is a topic of further research.

We see at least two theoretical frameworks that can help to gain a deeper understanding of the influence of using IT tools on student learning. The first is the instrumental approach (Artigue, 2002; Drijvers, 2003), which theorizes on how artefacts such as IT tools become instruments for solving mathematical problems in students’ hands. As Drijvers has shown for learning algebra with computer algebra systems, the development of ‘instrumentation schemes’ always has a technical and a conceptual side, which are interwoven. It is certainly not the case that having an IT tool that does the laborious computations allows students to focus on the conceptual side of solving problems (Artigue, 2002). Technical peculiarities of the tool can hinder students, but sometimes can also be used to improve students’ conceptual understanding (see Drijvers, 2003, for examples). It is likely that similar and different observations can be found for the instrumentation process of using statistics tools for solving statistical problems. One difference is that the educational statistics software we report on is specially designed for middle school students and has fewer technical hurdles.

A second theoretical framework that we consider potentially fruitful is a semiotics of IT tools, which still has to be worked out. When semiotics came to birth, signs were static. Though Peirce's notion of interpretant offers the opportunity to stress the dynamic aspects of sign activity, the interpretant is only the interpreter's response, not the reaction of a computer tool to the interpreter's actions. A computer tool reacts according to hidden rules and is therefore not transparent to all users. In our view, an elaborated sign notion is needed that takes the dynamic interaction with computer tools into account.

So far, we have focused on the influence of using IT tools on learning, but it is evident that drastic changes in learning demands other ways of teaching, assessing, and designing.

Teaching. In the present research we have focused on students' learning and have kept the role of the teacher in the background. Partially we could do so because we worked with experienced teachers who were well acquainted with the RME philosophy. However, in a follow-up teaching experiment we would certainly focus more on the teacher's role. One reason is that the success of particular instructional activities depends on the teacher. Another reason is that the approach we have taken in this study is not easy for teachers. They have to lead class discussions on students' reasoning about a variety of diagrams and arguments. The more options a computer tool offers, the more difficult it seems to be for the teacher. Particularly if the teacher wants to let students benefit from their own ideas, she or he needs to see the potential of students' often imprecise formulations. Furthermore, this approach requires a good understanding of data analysis techniques and of students' learning of those techniques. This implies the need for research into the professionalization of teachers in statistical data analysis, especially if software is used. In particular, we see the need of investigating how teachers can grasp the core ideas of an HLT.

Assessment. Nor is students' learning in the approach of this study easy to assess. We have used a number of assessment tasks, some of which were more informative than others, but most of them were rather laborious to score (e.g. question 2 from the test in class 1B). Such questions are unlikely to be used in national tests. Therefore, there is a need of assessment items that do assess what students learned and that might be used in large-scale tests (cf. Konold & Khalil, 2003).

Design. The hypotheses mentioned in Section 10.3.4 lead to the question of what the relative merits are of route and landscape-type tools for learning data analysis. In this thesis we have formulated a few questions about the affordances and constraints of particular representations such as value-bar graphs and dot plots. Do students, within settings as created in the present research, more readily see spread in a dot plot than in a value-bar graph? Can students more easily estimate the mean in a value-bar graph than in a dot plot or histogram? Does it make a difference for learning about the median in a value-bar graph whether the bars are vertical or horizontal? The answers to such questions are necessary to inform future designs, of instructional materials including software. Nonetheless, we were unable to answer those questions within the methodology chosen. Nor do we think that purely quantitative comparative research can help to answer those questions, because those diagrams gain their meanings within communities of learners (Meira, 1998). A combination of design research that creates the right conditions, and comparative research that isolates specific aspects may yield empirically grounded answers to such questions. Such a set-up requires a larger team than we had.

Finally, we need to gain more insight into which types of statistics education lead to statistical literacy so that citizens of the future knowledge society will wisely use its most important resource: information.