Chapter 16
An Introduction to Design-Based Research with an Example From Statistics Education

Arthur Bakker and Dolly van Eerde

Abstract This chapter arose from the need to introduce researchers, including Master and PhD students, to design-based research (DBR). In Sect. 16.1 we address key features of DBR and differences from other research approaches. We also describe the meaning of validity and reliability in DBR and discuss how they can be improved. Section 16.2 illustrates DBR with an example from statistics education.

Keywords Design based research • Statistics education

16.1 Theory of Design-Based Research

16.1.1 Purpose of the Chapter

The purpose of this chapter is to introduce researchers, including Master and PhD students, to design-based research. In our research methods courses for this audience and in our supervision of PhD students, we noticed that students considered key publications in this field unsuitable as introductions. These publications have mostly been written to inform or convince established researchers who already have considerable experience with educational research. We therefore see the need to write for an audience that does not have that level of experience, but may want to know about design-based research. We do assume a basic knowledge of the main research approaches (e.g., survey, experiment, case study) and methods (e.g., interview, questionnaire, observation).

Compared to other research approaches, educational design-based research (DBR) is relatively new (Anderson and Shattuck 2012). This is probably the reason that it is not discussed in most books on qualitative research approaches. For example, Creswell (2007) distinguishes five qualitative approaches, but these do not include DBR (see also Denscombe 2007). Yet DBR is worth knowing about, espe-

A. Bakker (✉) • D. van Eerde
Freudenthal Institute for Science and Mathematics Education, Utrecht University,
Princetonplein 5, 3584 CC, Utrecht, The Netherlands
e-mail: a.bakker4@uu.nl; h.a.a.vaneerde@uu.nl

A. Bikner-Ahsbahs et al. (eds.), *Approaches to Qualitative Research in Mathematics Education*, Advances in Mathematics Education,
DOI 10.1007/978-94-017-9181-6_16

cially for students who will become teachers or researchers in education: Design-based research is claimed to have the potential to bridge the gap between educational practice and theory, because it aims both at developing theories about domain-specific learning and the means that are designed to support that learning. DBR thus produces both useful products (e.g., educational materials) and accompanying scientific insights into how these products can be used in education (McKenney and Reeves 2012; Van den Akker et al. 2006). It is also said to be suitable for addressing complex educational problems that should be dealt with in a holistic way (Plomp and Nieveen 2007).

In line with the other chapters in this book, Sect. 16.1 provides a general theory of the research approach under discussion and Sect. 16.2 gives an example from statistics education on how the approach can be used.

16.1.2 Characterizing Design-Based Research

In this section we outline some characteristics of DBR, compare it with other research approaches, go over terminology and history, and finally summarize DBR's key characteristics.

16.1.2.1 Integration of Design and Research

Educational design-based research (DBR) can be characterized as research in which the design of educational materials (e.g., computer tools, learning activities, or a professional development program) is a crucial part of the research. That is, the design of learning environments is interwoven with the testing or developing of theory. The theoretical yield distinguishes DBR from studies that aim solely at designing educational materials through iterative cycles of testing and improving prototypes.

A key characteristic of DBR is that educational ideas for student or teacher learning are formulated in the design, but can be adjusted during the empirical testing of these ideas, for example if a design idea does not quite work as anticipated. In most other interventionist research approaches design and testing are cleanly separated. See further the comparison with a randomized controlled trial in Sect. 16.1.2.5.

16.1.2.2 Predictive and Advisory Nature of DBR

To further characterize DBR it is helpful to classify research aims in general (cf. Plomp and Nieveen 2007):

- To describe (e.g., What conceptions of sampling do seventh-grade students have?)
- To compare (e.g., Does instructional strategy A lead to better test scores than instructional strategy B?)

- To evaluate (e.g., How well do students develop an understanding of distribution in an instructional sequence?)
- To explain or to predict (e.g., Why do so few students choose a bachelor in mathematics or science? What will students do when using a particular software package?)
- To advise (e.g., How can secondary school students be supported to learn about correlation and regression?)

Many research approaches such as surveys, correlational studies, and case studies, typically have descriptive aims. Experiments often have a comparative aim, even though they should in Cook's (2002) view "be designed to *explain* the consequences of interventions and not just to describe them" (p. 181, emphasis original). DBR typically has an explanatory and advisory aim, namely to give theoretical insights into how particular ways of teaching and learning can be promoted. The type of theory developed can also be of a predictive nature: Under conditions X using educational approach Y, students are likely to learn Z (Van den Akker et al. 2006).

Research projects usually have one overall aim, but several stages of the project can have other aims. For example, if the main aim of a research project is to advise how a particular topic (e.g., sampling) should be taught, the project most likely has parts in which phenomena are described or evaluated (e.g., students' prior knowledge, current teaching practices). It will also have a part in which an innovative learning environment has to be designed and evaluated before empirically grounded advice can be given. This implies that research projects are layered. Design-based research (DBR) has an overall predictive or advisory aim but often includes research stages with a descriptive, comparative, or evaluative aim.

16.1.2.3 The Role of Hypotheses and the Engineering Nature of DBR

In characterizing DBR as different from other research approaches, we also need to address the role of hypotheses in theory development. Put simply, a scientific theory can explain particular phenomena and predict what will happen under particular conditions. When developing or testing a theory, scientists typically use hypotheses—conjectures that follow from some emergent theory that still needs to be tested empirically. This means that hypotheses should be formulated in a form in which they can be verified or falsified. The testing of hypotheses is typically done in an experiment: Reality is manipulated according to a theory-driven plan. If hypotheses are confirmed, this is support for the theory under construction.

Just as in the natural sciences, it is not always possible to test hypotheses empirically within a short period of time. As a starting point design researchers, just like many scientists in other disciplines, use thought experiments—thinking through the consequences of particular ideas. When preparing an empirical teaching experiment, design researchers typically do a thought experiment on how teachers or students will respond to particular tools or tasks based on their practical and theoretical knowledge of the domain (Freudenthal 1991).

In empirical experiments, a hypothesis is formulated beforehand. A theoretical idea is operationalized by designing a particular setting in which only this particular feature is isolated and manipulated. To stay objective experimental researchers are often not present during the interventions. In typical cases, they collect only pre- and posttest scores. In design-based research, however, researchers continuously take their best bets (Lehrer and Schauble 2001), even if this means that some aspect of the learning environment during or after a lesson has to be changed. In many examples, researchers are involved in the teaching or work closely with teachers or trainers to optimize the learning environment (McClain and Cobb 2001; Smit and Van Eerde 2011; Hoyles et al. 2010). In the process of designing and improving educational materials (which we take as a prototypical case in this chapter), it does not make sense to wait until the end of the teaching experiment before changes can be made. This would be inefficient.

DBR is therefore sometimes characterized as a form of didactical engineering (Artigue, 1988): didactical engineering: Something has to be made with whatever theories and resources are available. The products of DBR are judged on innovativeness and usefulness, not just on the rigor of the research process that is more prominent in evaluating true experiments (Plomp 2007).

In many research approaches, changing and understanding a situation are separated. However, in design-based research these are intertwined in line with the following adage that is also common in sociocultural traditions: If you want to understand something you have to change it, and if you want to change something you have to understand it (Bakker 2004a, p. 37).

16.1.2.4 Open and Interventionist Nature of DBR

Another way to characterize DBR is to contrast it with other approaches on the following two dimensions: naturalistic vs. interventionist and open vs. closed. Naturalistic studies analyze how learning takes place without interference by a researcher. Examples of naturalistic research approaches are ethnography and surveys. As the term suggests, interventionist studies intervene in what naturally happens: Researchers deliberately manipulate a condition or teach according to particular theoretical ideas (e.g., inquiry-based or problem-based learning). Such studies are necessary if the type of learning that researchers want to investigate is not present in naturalistic settings. Examples of interventionist approaches are experimental research, action research, and design-based research.

Research approaches can also be more open or closed. The term *open* here refers to little control of the situation or data whereas *closed* refers to a high degree of control or a limited number of options (e.g., multiple choice questions). For example, surveys by means of questionnaires with closed questions or responses on a Likert scale are more closed than surveys by means of semi-structured interviews. Likewise, an experiment comparing two conditions is more closed than a DBR project in which the educational materials or ways of teaching are emergent and adjustable. Different research approaches can thus be positioned in a two-by-two table as in Table 16.1. DBR thus shares an interventionist nature with experiments and action research. We therefore continue by comparing DBR with experiments (16.1.2.5) and with action research (16.1.2.6).

Table 16.1 Naturalistic vs. interventionist and open vs. closed research approaches

	Naturalistic	Interventionist
Closed	Survey: questionnaires with closed questions	Experiment (randomized controlled trial)
Open	Survey: interviews with open questions	Action research
	Ethnography	Design-based research

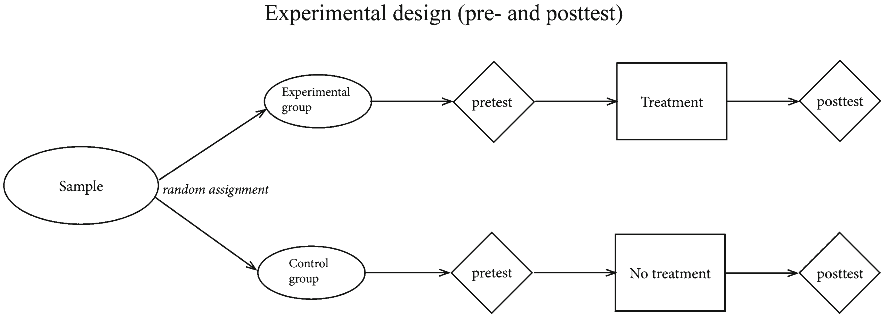


Fig. 16.1 A pre-posttest experimental design (randomized controlled trial)

16.1.2.5 Comparison of DBR with Randomized Controlled Trials (RCT)

A randomized controlled trial (RCT) is sometimes referred to as "true" experiment. Assume we want to know whether a new teaching strategy for a particular topic in a particular grade is better than the traditionally used one. To investigate this question one could randomly assign students to the experimental (new teaching strategy) or control condition (traditional strategy), measure performances on pre- and posttests, and use statistical methods to test the null hypothesis that there is no significant difference between the two conditions. The researchers' hope is that this hypothesis can be rejected so that the new type of intervention (informed by a particular theory) proves to be better. The underlying rationale is: If we know "what works" we can implement this method and have better learning results (see Fig. 16.1).

This so-called experimental approach of randomized controlled trials (Creswell 2005) is sometimes considered the highest standard of research (Slavin 2002). It has a clear logic and is a convincing way to make causal and general claims about what works. It is based on a research approach that has proven extremely helpful in the natural sciences.

However, its limitations for education are discussed extensively in the literature (Engeström 2011; Olsen 2004). Here we mention two related arguments. First, if we know what works, we still do not know why and when it works. Even if the new strategy is implemented, it might not work as expected because teachers use it in less than optimal ways.

An example can clarify this. When doing research in an American school, we heard teachers complain about their managers' decision that every teacher had to start every lesson with a warm-up activity (e.g., a puzzle). Apparently it had been proven by means of an RCT that student scores were significantly higher in the experimental condition in which lessons started with a warm-up activity. The negative effect in teaching practice, however, was that teachers ran out of good ideas for warm-up activities, and that these often had nothing to do with the topic of the lesson. Effectively, teachers therefore lost five minutes of every lesson. Better insight into how and why warm-up activities work under particular conditions could have improved the situation, but the comparative nature of RCT had not provided this information because only the variable of starting the lesson with or without warm-up activity had been manipulated.

A second argument why RCT has its limitations is that a new strategy has to be designed before it can be tested, just like a Boeing airplane cannot be compared with an Airbus without a long tradition of engineering and producing such airplanes. In many cases, considerable research is needed to design innovative approaches. Design-based research emerged as a way to address this need of developing new strategies that could solve long-standing or complex problems in education.

Two discussion points in the comparison of DBR and RCT are the issues of generalization and causality. The use of random samples in RCT allows generalization to populations, but in most educational research random samples cannot be used. In response to this point, researchers have argued that theory development is not just about populations, but rather about propensities and processes (Frick 1998). Hence rather than generalizing from a random sample to a population (statistical generalization), many (mainly qualitative) research approaches aim for generalization to a theory, model or concept (theoretical or analytic generalization) by presenting findings as particular cases of a more general model or concept (Yin 2009).

Where the use of RCTs can indicate the intervention or treatment being the cause of better learning, DBR cannot claim causality with the same convincing rigor. This is not unique to DBR: All qualitative research approaches face this challenge of drawing causal claims. In this regard it is helpful to distinguish two views on causality: a regularity, variance-oriented understanding of causality versus a realist, process-oriented understanding of causality (Maxwell 2004). People adopting the first view think that causality can only be proven on the basis of regularities in larger data sets. People adopting the second view make it plausible on the basis of circumstantial evidence of observed processes that what happened is most likely caused by the intervention (e.g., Nathan and Kim 2009). The first view is underlying the logic of RCT: If we randomly assign subjects to an experimental and control condition, treat only the experimental group and find a significant difference between the two groups, then it can only be attributed to the difference in condition (the treatment). However, if we were to adopt the same regularity view on causality we would never be able to identify the cause of singular events, for example why a driver hit a tree. From the second, process-oriented view, if a drunk driver hits a tree we can judge the circumstances and judge it plausible that his drunkenness was an important explanation because we know that alcohol can cause less control, slower reaction time et cetera. Similarly, explanations for what happens in classrooms should be possible according to a process-oriented position based on what happens in response to particular interventions. For example, particular student utterances are very unlikely if not deliberately fostered by a teacher (Nathan and Kim 2009). Table 16.2 summarizes the main points of the comparison of RCT and DBR.

16.1.2.6 Comparison of DBR with Action Research

Like action research, DBR typically is interventionist and open, involves a reflective and often cyclic process, and aims to bridge theory and practice (Opie 2004). In both approaches the teacher can be also researcher. In action research, the researcher is not an observer (Anderson and Shattuck 2012), whereas in DBR s/he can be observer. Furthermore, in DBR design is a crucial part of the research, whereas in action research the focus is on action and change, which can but need not involve the design of a new learning environment. DBR also more explicitly aims for instructional theories than does action research. These points are summarized in Table 16.3.

Table 16.2 Comparison of experimental versus design-based research

Experiment (RCT)	Design-based research (DBR)
Testing theory	Developing and testing theory simultaneously
Comparison of existing teaching methods by means of experimental and control groups	Design of an innovative learning environment long
Proof of what works	Insight into how and why something works
Research interest is isolated by manipulating variables separately	Holistic approach long white word
Statistical generalization	Analytic or theoretical generalization, transferability to other situations
Causal claims based on a regularity view on causality are possible	Causality should be handled with great care and be based on a realist, process-oriented view on causality

Table 16.3 Commonalities and differences between DBR and action research

	DBR	Action research
Commonalities	Open, interventionist, researcher can be participant, reflective cyclic process	
Differences	Researcher can be observer	Researcher can only be participant
	Design is necessary	Design is possible
	Focus on instructional theory	Focus on action and improvement of a situation

16.1.2.7 Names and History of DBR

In its relatively brief history, DBR has been presented under different names. *Design-based research* is the name used by the Design-Based Research Collective (see special issues in Educational Researcher, 2003; Educational Psychologist 2004; Journal of the Learning Sciences 2004). Other terms for similar approaches are:

- Developmental or development research (Freudenthal 1988; Gravemeijer 1994; Lijnse 1995; Romberg 1973; Van den Akker 1999)
- Design experiments or design experimentation (Brown 1992; Cobb et al. 2003a; Collins 1992)
- Educational design research (Van den Akker et al. 2006)

The reasons for these different terms are mainly historical and rhetorical. In the 1970s Romberg (1973) used the term *development research* for research accompanying the development of curriculum. Discussions on the relation between research and design in mathematics education, especially on didactics, mainly took place in Western Europe in the 1980s and the 1990s, particularly in the Netherlands (e.g., Freudenthal 1988; Goffree 1979), France (e.g., Artigue 1988, cf. Artigue Chap. 17) and Germany (e.g., Wittmann 1992). The term *developmental research* is a translation of the Dutch *ontwikkelingsonderzoek*, which Freudenthal introduced in the 1970s to justify the development of curricular materials as belonging to a university institute (what is now called the Freudenthal Institute) because it was informed by and leading to research on students' learning processes (Freudenthal 1978; Gravemeijer and Koster 1988; De Jong and Wijers 1993). The core idea was that development of learning environments and the development of theory were intertwined. As Goffree (1979, p. 347) put it: "Developmental research in education as presented here, shows the characteristics of both developmental and fundamental research, which means aiming at new knowledge that can be put into service in continued development." At another Dutch university (Twente University), the term *ontwerpgericht* (design-oriented) research was more common, but there the focus was more on the curriculum than on theory development (Van den Akker 1999). One disadvantage of the terms 'development' and 'developmental' is their connotations to developmental psychology and research on children's development of concepts. This might be one reason that this term is hardly used anymore.

In the United States, the terms *design experiment* and *design research* were more common (Brown 1992; Cobb et al. 2003a; Collins 1992; Edelson 2002). One advantage of these terms is that design is more specific than development. One possible disadvantage of the term *design experiment* can be explained by reference to a critical paper by Paas (2005) titled *Design experiment: Neither a design nor an experiment*. The confusion that his pun refers to is two-fold. First, in many educational research communities the term *design* is reserved for research design (e.g., comparing an experimental with a control group), whereas the term in design research refers to the design of learning environments (Sandoval and Bell 2004). Second, for many researchers, also outside the learning sciences, the term *experiment* is reserved for "true" experiments or RCTs. In design experiments, hypotheses certainly play an important role, but they are not fixed and tested once. Instead they may be emergent, multiple, and temporary. In line with the Design-Based Research Collective, we use the term *design-based research* because this suggests that it is predominantly research (hence leading to a knowledge claim) that is based on a design process.

16.1.2.8 Theory Development in Design-Based Research

We have already stated that theory typically has a more central role in DBR than in action research. To address the role of theory in DBR, it is helpful to summarize diSessa and Cobb's (2004) categorization of different types of theories involved in educational research. They distinguish:

- Grand theories (e.g., Piaget's phases of intellectual development; Skinner's behaviorism)
- Orienting frameworks (e.g., constructivism, semiotics, sociocultural theories)
- Frameworks for action (e.g., designing for learning, Realistic Mathematics Education)
- Domain-specific theories (e.g., how to teach density or sampling)
- Hypothetical Learning Trajectories (Simon 1995) or didactical scenarios (Lijnse 1995; Lijnse and Klaassen 2004) formulated for specific teaching experiments (explained in Sect. 16.1.3).

As can be seen from this categorization, there is a hierarchy in the generality of theories. Because theories developed in DBR are typically tied to specific learning environments and learning goals, they are humble and hard to generalize. Similarly, it is very rare that a theoretical contribution to aerodynamics will be made in the design of an airplane; yet innovations in airplane design occur regularly. The use of grand theoretical frameworks and frameworks for action is recommended, but researchers should be careful to manage the gap between the different types of theory on the one hand and design on the other (diSessa and Cobb 2004). If handled with care, DBR can then provide the basis for refining or developing theoretical concepts such as meta-representational competence, sociomathematical norms (diSessa and Cobb), or whole-class scaffolding (Smit et al. 2013).

16.1.2.9 Summary of Key Characteristics of Design-Based Research

So far we have characterized DBR in terms of its predictive and advisory aim, particular way of handling hypotheses, its engineering nature and differences from other research methods. Here we summarize five key characteristics of DBR as identified by Cobb et al. (2003a):

1. The first characteristic is that its purpose is *to develop theories about learning and the means that are designed to support that learning*. In the example provided in Sect. 16.2 of in this chapter, Bakker (2004a) developed an instruction theory for early statistics education and instructional means (e.g. computer tools

and accompanying learning activities) that support the learning of a multifaceted notion of statistical distribution.

2. The second characteristic of DBR is its *interventionist* nature. One difference with RCTs is that interventions in the DBR tradition often have better ecological validity—meaning that learning already takes place in learning ecologies as they occur in schools and thus methods measure better what researchers want to measure, that is learning in natural situations. Findings from experiments do not have to be translated as much from controlled laboratory situations to the less controlled ecology of schools or courses. In technical terms, theoretical products of DBR "have the potential for rapid pay-off because they are filtered in advance for instrumental effect" (Cobb et al. 2003a, p. 11).
3. The third characteristic is that DBR has *prospective and reflective components* that need not be separated by a teaching experiment. In implementing hypothesized learning (the prospective part) the researchers confront conjectures with actual learning that they observe (reflective part). Reflection can be done after each lesson, even if the teaching experiment is longer than one lesson. Such reflective analysis can lead to changes to the original plan for the next lesson. Kanselaar (1993) argued that any good educational research has prospective and reflective components. As explained before, however, what distinguishes DBR from other experimental approaches is that in DBR these components are not separated into the formulation of hypotheses before and after a teaching experiment.
4. The fourth characteristic is the *cyclic* nature of DBR: Invention and revision form an iterative process. Multiple conjectures on learning are sometimes refuted and alternative conjectures can be generated and tested. The cycles typically consist of the following phases: preparation and design phase, teaching experiment, and retrospective analysis. These phases are worked out in more detail later in this chapter. The results of such a retrospective analysis mostly feed a new design phase. Other types of educational research ideally also build upon prior experiments and researchers iteratively improve materials and theoretical ideas in between experiments but in DBR changes can take place during a teaching experiment or series of teaching experiments.
5. The fifth characteristic of DBR is that the *theory* under development *has to do real work*. As Lewin (1951, p. 169) wrote: "There is nothing so practical as a good theory." Theory generated from DBR is typically humble in the sense that it is developed for a specific domain, for instance statistics education. Yet it must be general enough to be applicable in different contexts such as classrooms in other schools in other countries. In such cases we can speak of transferability.

16.1.3 Hypothetical Learning Trajectory (HLT)

DBR typically consists of cycles of three phases each: preparation and design, teaching experiment, and retrospective analysis. One might argue that the term 'retrospective analysis' is pleonastic: All analysis is in retrospect, after a teaching experiment. However, we use it here to distinguish it from analysis on the fly, which takes place during a teaching experiment, often between lessons.

A design and research instrument that proves useful during all phases of DBR is the *hypothetical learning trajectory* (HLT), which we regard as an elaboration of Freudenthal's thought experiment. Simon (1995) defined the HLT as follows:

> The hypothetical learning trajectory is made up of three components: the learning goal that defines the direction, the learning activities, and the hypothetical learning process—a prediction of how the students' thinking and understanding will evolve in the context of the learning activities. (p. 136)

Simon used the HLT for one or two lessons. Series of HLTs can be used for longer sequences of instruction (also see the literature on didactical scenarios in Lijnse 1995). The HLT is a useful research instrument to manage the gap between an instruction theory and a concrete teaching experiment. It is informed by general domain-specific and conjectured instruction theories (Gravemeijer 1994), and it informs researchers and teachers how to carry out a particular teaching experiment. After the teaching experiment, it guides the retrospective analysis, and the interplay between the HLT and empirical results forms the basis for theory development. This means that an HLT, after it has been mapped out, has different functions depending on the phase of the DBR and continually develops through the different phases. It can even change during a teaching experiment.

16.1.3.1 HLT in the Design Phase

The development of an HLT starts with an analysis of how the mathematical topic of the design study is elaborated in the curriculum and the mathematical textbooks, an analysis of the difficulties students encounter with this topic, and a reflection on what they should learn about it. These analyses result in the formulation of provisional mathematical learning goals that form the orientation point for the design and redesign of activities in several rounds. While designing mathematical activities the learning goals may become better defined. During these design processes the researcher also starts formulating hypotheses about students' potential learning and about how the teacher would support students' learning processes. The confrontation of a general rationale with concrete tasks often leads to a more specific HLT, which means that the HLT gradually develops during the design phase (Drijvers 2003).

An elaborated HLT thus includes mathematical learning goals, students' starting points with information on relevant pre-knowledge, mathematical problems and assumptions about students' potential learning processes and about how the teacher could support these processes.

16.1.3.2 HLT in Teaching Experiment

During the teaching experiment, the HLT functions as a guideline for the teacher and researcher for what to focus on in teaching, interviewing, and observing. It may happen that the teacher or researcher feels the need to adjust the HLT or instructional activity for the next lesson. As Freudenthal wrote (1991, p. 159), the cyclic alternation of research and development can be more efficient the shorter the cycle is. Minor changes in the HLT are usually made because of incidents in the classroom such as student strategies that were not foreseen, activities that were too difficult, and so on. Such adjustments are generally not accepted in comparative experimental research, but in DBR, changes in the HLT are made to create optimal conditions and are regarded as elements of the data corpus. This means that these changes have to be reported well and the information is stronger when changes are supported by theoretical considerations. The HLT can thus also change during the teaching experiment phase.

16.1.3.3 HLT in the Retrospective Analysis

During the retrospective analysis, the HLT functions as a guideline determining what the researcher should focus on in the analysis. Because predictions are made about students' learning, the researcher can contrast those conjectures with the observations made during the teaching experiment. Such an analysis of the interplay between the evolving HLT and empirical observations forms the basis for developing an instruction theory. After the retrospective analysis, the HLT can be reformulated, often more drastically than during the teaching experiment, and the new HLT can guide a subsequent design phase.

An HLT can be seen as a concretization of an evolving domain-specific instruction theory. Conversely, the instruction theory is informed by evolving HLTs. For example, if patterns of an HLT stabilize after a few cycles, these generalized patterns in learning or instruction and the insights of how these patterns are supported by instructional means can become part of the emerging instruction theory.

Overall, the idea behind developing an HLT is not to design the perfect instructional sequence, which in our view does not exist, but to provide empirically grounded results that others can adjust to their local circumstances. The HLT remains hypothetical because each situation, each teacher, and each class is different. Yet patterns can be found in students' learning that are similar across different teaching experiments. Those patterns and the insights of how particular educational activities support students in particular kinds of reasoning can be the basis for a more general instructional theory of how a particular domain can be taught. Bakker (2004a), for example, noted that when estimating the number of elephants in a picture, students typically used one of four strategies, and these four strategies reoccurred in all of the five classrooms in which he used the same task. Having observed such a pattern in strategy use, the design researcher can assume the pattern to be an element of the instruction theory.

For some readers, the term 'trajectory' might have a linear connotation. Although we aim for a certain direction, like the course of a ship, Bakker's (2004a) HLTs were non-linear in the sense that he did not make a linear sequence of activities in advance that he strictly adhered to (cf. Fosnot and Dolk 2001). Moreover, two subtrajectories came together later on in the sequence. In the following sections we give a more detailed description of the three phases of a DBR cycle and discuss relevant methodological issues. Further details about hypothetical learning trajectories can be found in a special issue of *Mathematical Thinking and Learning* (Mathematical Thinking and Learning 2004, volume 6, issue 2) devoted to HLTs.

The term HLT stems from research in which the teacher was a researcher or a member of the research team (Simon 1995). However, if the teacher is not so familiar with the research team's intentions it may be necessary to pay extra attention to what the teacher can or should do to realize the potential of the learning activities. In such cases, the terms *hypothetical teaching and learning trajectory* (HTLT) or *teaching and learning strategy* (Dierdorp et al. 2011) may be more appropriate.

16.1.4 Phases in DBR

16.1.4.1 Phase 1: Preparation and Design

It is evident that the relevant present knowledge about a topic should be studied first. Gravemeijer (1994) characterizes the design researcher as a tinkerer or, in French, a *bricoleur*, who uses all the material that is at hand, including theoretical insights and practical experience with teaching and designing.

In the first design phase, it is recommended to collect and invent a set of tasks that could be useful and discuss these with colleagues who are experienced in designing for mathematics education. An important criterion for selecting a task is its potential role in the HLT towards the mathematical end goal. Could it possibly lead to types of reasoning that students could build upon towards that end goal? Would it be challenging? Would it be a meaningful context for students?

There are several design heuristics, principles, and guidelines. In Sect. 16.2 we explain heuristics from the theory of Realistic Mathematics Education.

16.1.4.2 Phase 2: Teaching Experiment

The notion of a teaching experiment arose in the 1970s. Its primary purpose was to experience students' learning and reasoning first-hand, and it thus served the purpose of eliminating the separation between the practice of research and the practice of teaching (Steffe and Thompson 2000). Over time, teaching experiments proved useful for a broader purpose, namely as part of DBR. During a teaching experiment, researchers and teachers use activities and types of instruction that according to the HLT seem most appropriate at that moment. Observations in one lesson and theoretical arguments from multiple sources can influence what is done in the next lesson. Observations may include student or teacher deviations from the HLT.

Hence, this type of research is different from experimental research designs in which a limited number of variables are manipulated and effects on other variables are measured. The situation investigated here, the learning of students in a new context with new tools and new end goals, is too complicated for such a set-up.

Besides that, a different type of knowledge is looked for, as pointed out earlier in this chapter: We do not want to assess innovative material or a theory, but we need prototypical educational materials that could be tested and revised by teachers and researchers, and a domain-specific instruction theory that can be used by others to formulate their own HLTs suiting local contingencies.

During a teaching experiment, data collection typically includes student work, tests before and after instruction, field notes, audio recordings of whole-class discussions, and video recordings of every lesson and of the final interviews with students and teachers. We further find 'mini-interviews' with students, lasting from about twenty seconds to four minutes, very useful provided that they are carried out systematically (Bakker 2004a).

16.1.4.3 Retrospective Analysis

We describe two types of analysis useful in DBR, a task oriented analysis and a more overall, longitudinal, cyclic approach. The first is to compare data on students' actual learning during the different tasks with the HLT. To this end we find the data analysis matrix (Table 16.4) described in Dierdorp et al. (2011) useful. The left part of the matrix summarizes the HLT and the right part is filled with excerpts from relevant transcripts, clarifying notes from the researcher as well as a quantitative impression of how well the match was between the assumed leaning as formulated in the HLT and the observed learning. With such analysis it is possible to give an overview, as in Table 16.5, which can help to identify problematic sections in the educational materials. Insights into why particular learning takes place or does not

Table 16.4 Data analysis matrix for comparing HLT and actual learning trajectory (ALT)

Hypothetical learning trajectory			Actual learning trajectory		
Task number	Formulation of the task	Conjecture of how students would respond	Transcript excerpt	Clarification	Match between HLT and ALT: Quantitative impression of how well the conjecture and actual learning matched (e.g., –, 0, +)

Table 16.5 ALT result compared with HLT conjectures for the tasks involving a particular type of reasoning

+				x	x			x	x	x	x	x		x	x	x	x	x	x
±	x		x										x						
–		x				x	x												
Task:	5d	5f	6a	6c	7	8	9c	9e	10b	11c	15	17	23b	23c	24a	24c	25d	34a	42

Note: an x means how well the conjecture accompanying that task matched the observed learning (– refers to confirmation for up to 1/3 of the students, and + to at least 2/3 of the students) take place help to improve the HLTs in subsequent cycles of DBR. This iterative process allows the researcher to improve the predictive power of HLTs across subsequent teaching experiments.

An elaborated HLT would include assumptions about students' potential learning and about how the teacher would support students' learning processes. In this task-oriented analysis above no information is included about the role of the teacher. If there are crucial differences between students' assumed and observed learning processes or if the teaching has been observed to diverge radically from what the researcher had intended, the role of the teacher should be included into the analysis in search of explanations for these discrepancies.

A comparison of HLTs and observed learning is very useful in the redesign process, and allows answers to research questions that ask how particular learning goals could be reached. However, in our experience additional analyses are often needed to gain more theoretical insights into the learning process. An example of such additional analysis is a method inspired by the *constant comparative method* (Glaser and Strauss 1967; Strauss and Corbin 1998) and Cobb and Whitenack's (1996) method of longitudinal analyses. Bakker (2004a) used this type of analysis in his study in the following way. First, all transcripts were read and the videotapes were watched chronologically episode-by-episode. With the HLT and research questions as guidelines, conjectures about students' learning and views were generated and documented, and then tested against the other episodes and other data material (student work, field notes, tests). This testing meant looking for confirmation and counter-examples. The process of conjecture generating and testing was repeated. Seemingly crucial episodes were discussed with colleagues to test whether they agreed with our interpretation or perhaps could think of alternative interpretations. This process is called *peer examination*.

For the analysis of transcripts or videos it is worth considering computer software such as Atlas.ti (Van Nes and Doorman 2010) for coding the transcripts and other data sources. As in all qualitative research, data triangulation (Denscombe 2007) is commonly used in design-based research.

16.1.5 Validity and Reliability

Researchers want to analyze data in a reliable way and draw conclusions that are valid. Therefore, validity and reliability are important concerns. In brief, validity concerns whether we really measure what we intend to measure. Reliability is about independence of the researcher. A brief example may clarify the distinction. Assume a researcher wants to measure students' mathematical ability. He gives everyone 7 out of 10. Is this a valid way of measuring? Is this a reliable way?

It is a very reliable way because the instruction "give all students a 7" can be reliably carried out, independently of the researcher. However, it is not valid, because there is most likely variation between students' mathematical ability, which is not taken into account with this way of measuring.

We should emphasize that validity and reliability are complex concepts with multiple meanings in different types of research. In qualitative research the meanings of validity and reliability are slightly different than in quantitative research. Moreover, there are so many types of validity and reliability that we cannot address them all. In this chapter we have focused on those types that seemed most relevant to us in the context of DBR. The issues discussed in this section are inspired by guidelines of Maso and Smaling (1998) and Miles and Huberman (1994), who distinguish between internal and external validity and reliability.

16.1.5.1 Internal Validity

Internal validity refers to the quality of the data and the soundness of the reasoning that has led to the conclusions. In qualitative research, this soundness is also labeled as *credibility* (Guba 1981). In DBR, several techniques can be used to improve the internal validity of a study.

- During the retrospective analysis conjectures generated and tested for specific episodes are tested for other episodes or by data triangulation with other data material, such as field notes, tests, and other student work. During this testing stage there is a search for counterexamples to the conjectures.
- The succession of different teaching experiments makes it possible to test the conjectures developed in earlier experiments in later experiments.

Theoretical claims are substantiated where possible with transcripts to provide a rich and meaningful context. Reports about DBR tend to be long due to the *thick descriptions* (Geertz 1973) required. For example, the paper by Cobb et al. (2003b) is 78 pages long!

16.1.5.2 External Validity

External validity is mostly interpreted as the generalizability of the results. The question is how we can generalize the results from these specific contexts to be useful for other contexts. An important way to do so is by framing issues as instances of something more general (Cobb et al. 2003a; Gravemeijer and Cobb 2006). The challenge is to present the results (instruction theory, HLT, educational activities) in such a way that others can adjust them to their local contingencies.

In addition to generalizability as a criterion for external validity we mention *transferability* (Maso and Smaling 1998). If lessons learned in one experiment are successfully applied in other experiments, this is a sign of successful generalization. At the end of Sect. 16.2 we give an example of how a new type of learning activity was successfully enacted in a new research project in another country.

16.1.5.3 Internal Reliability

Internal reliability refers to the degree of how independently of the researcher the data are collected and analyzed. It can be improved with several methods. Data collection by objective devices such as audio- and video registrations contribute to the internal reliability. During his retrospective analysis Bakker (2004a) ensured reliability by discussing the critical episodes, including those discussed in Sect. 16.2, with colleagues for peer examination. For measuring interrater reliability, the agreement among independent researchers, it is advised to calculate not only the percentage of agreement but also use Cohen's kappa or another measure that takes into account the probability of agreement by chance (e.g., Krippendorff's alpha). It is not necessary for a second coder to code all episodes, but ensure that a random sample should be of sufficient size: The larger the number of possible codes, the larger the sample required (Bakkenes et al. 2010; Cicchetti 1976). Note that the term internal reliability can also refer to the consistency of responses on a questionnaire or test, often measured with help of Cronbach's alpha.

16.1.5.4 External Reliability

External reliability usually denotes replicability, meaning that the conclusions of the study should depend on the subjects and conditions, and not on the researcher. In qualitative research, replicability is mostly interpreted as virtual replicability. The research must be documented in such a way that it is clear how the research has been carried out and how conclusions have been drawn from the data. A criterion for virtual replicability is 'trackability' (Gravemeijer and Cobb 2006), 'traceability' (Maso and Smaling 1998), or transparency (Akkerman et al. 2008). This means that the reader must be able to track or trace the learning process of the researchers and to reconstruct their study: failures and successes, procedures followed, the conceptual framework used, and the reasons for certain choices must all be reported. In Freudenthal's words:

> Developmental research means: experiencing the cyclic process of development and research so consciously, and reporting on it so candidly that it justifies itself, and that this experience can be transmitted to others to become like their own experience. (1991, p. 161)

We illustrate the general characterization and description of DBR of Sect. 16.1 by an example of a design study on statistics education in Sect. 16.2.

16.2 Example of Design-Based Research

In this second section we illustrate the theory of design-based research (DBR) as outlined in Sect. 16.1 with an example from Bakker's (2004a, b) PhD thesis on DBR in statistics education. We briefly describe the aim and theoretical background of this DBR project and then focus on one design idea, that of growing samples, to illustrate how it is related to different layers of theory and how it was analyzed. Finally we discuss the issue of generalizability. In the appendix we provide a structure of a DBR project with examples from this Sect. 16.2.

16.2.1 Relevance and Aim

The background problem addressed in Bakker's (2004a) research on statistics education was that many stakeholders were dissatisfied with what and how students learned about statistics. For example, in many curricula there was a focus on computing arithmetic means and making bar charts (Friel et al. 2001). Moreover, there was very little knowledge about how to use innovative educational statistics software (cf. Biehler et al. 2013, for an historical overview).

To solve these practical problems, Bakker's (2004a) aim was to contribute to an empirically grounded instruction theory for early statistics education with new computer tools for the age group from 11 to 14. Such a theory should specify patterns in students' learning as well as the means supporting that learning in the domain of statistics education. Like Cobb et al. (2003b), Bakker (2004a) focused his research on the concept of distribution as a key concept in statistics. One problem is that students tend to see isolated data points instead of a data set as a whole (Bakker and Gravemeijer 2004; Konold and Higgins 2003). Yet statistics is about features of data sets, in particular distributions of samples. The selected learning goal was therefore that distribution had to become an object-like entity with which students could see data sets as an entity with characteristics.

16.2.2 Research Question

Bakker's initial research question was: How can students with little statistical background develop a notion of distribution? In trying to answer this question in grade 7, however, Bakker came to include a focus on other statistical key concepts such as data, center, and sampling because these are so intricately connected to that of distribution (Bakker and Derry 2011). The concept of distribution also proved hard for seventh-grade students. The initial research question was therefore reformulated for grade 8 as follows: How can coherent reasoning about distribution be promoted in relation to data, variability, and sampling in a way that is meaningful for students with little statistical background?

Our point here is that research questions can change during a research project. Indeed, the better and sharper your research question is in the beginning of the project, the better and more focused your data collection will be. However, our experience is that most DBR researchers, due to progressive insight, end up with slightly different research questions than they started with.

As pointed out in Sect. 16.1, DBR typically draws on several types of theories. Given the importance of graphical representations in statistics education, it made sense for Bakker to draw on semiotics as an orienting framework. He came to focus on semiotics, in particular Peirce's ideas on diagrammatic reasoning. The domain-specific theory of Realistic Mathematics Education proved a useful framework for action in the design process even though it had hardly been applied in statistics education.

16.2.3 Orienting Framework: Diagrammatic Reasoning

The learning goal was that distribution would become an object-like entity. Theories on reification of concepts (Sfard and Linchevski 1992) and the relation between process and concept (cf. Tall et al. 2000, on *procept*) were drawn upon. One theoretical question unanswered in the literature was what the process nature of a distribution could be. It is impossible to make sense of graphs without having appropriate conceptual structures, and it is impossible to communicate about concepts without any representations. Thus, to develop an instruction theory it is necessary to investigate the relation between the development of the meaning of graphs and concepts. After studying several theories in this area, Bakker deployed Peirce's semiotic theory on diagrammatic reasoning (Bakker 2007; Bakker and Hoffmann 2005). For Peirce, a diagram is a sign that is meant to represent relations. Diagrammatic reasoning involves three steps:

1. The first step is to *construct* a diagram (or diagrams) by means of a representational system such as Euclidean geometry, but we can also think of diagrams in computer software or of an informal student sketch of statistical distribution. Such a construction of diagrams is supported by the need to represent the relations that students consider significant in a problem. This first step may be called *diagrammatization*.
2. The second step of diagrammatic reasoning is to *experiment* with the diagram (or diagrams). Any experimenting with a diagram is executed within a not necessarily perfect representational system and is a rule or habit-driven activity. Contemporary researchers would stress that this activity is situated within a practice. What makes experimenting with diagrams important is the rationality immanent in them (Hoffmann 2002). The rules define the possible transformations and actions, but also the constraints of operations on diagrams. Statistical diagrams such as dot plots are also bound by certain rules: a dot has to be put above its value on the x axis and this remains true even if for instance the scale is changed. Peirce stresses the importance of doing something when thinking or reasoning with diagrams:

   > Thinking in general terms is not enough. It is necessary that something should be DONE. In geometry, subsidiary lines are drawn. In algebra, permissible transformations are made. Thereupon the faculty of observation is called into play. (CP 4.233—CP refers to Peirce's collected papers, volume 4, section 233)

In the software used in this research, students can do something with the data points such as organizing them into equal intervals or four equal groups.

3. The third step is to observe the results of experimenting. We refer to this as the *reflection* step. As Peirce wrote, the mathematician observing a diagram "puts before him an icon by the observation of which he detects relations between the parts of the diagram other than those which were used in its construction" (Peirce 1976 III, p. 749). In this way he can "discover unnoticed and hidden relations among the parts" (Peirce CP 3.363; see also CP 1.383). The power of diagrammatic reasoning is that "we are continually bumping up against hard fact. We expected one thing, or passively took it for granted, and had the image of it in our minds, but experience forces that idea into the background, and compels us to think quite differently" (Peirce CP 1.324).

Diagrammatic reasoning, in particular the reflection step, is what can introduce the 'new'. New implications within a given representational system can be found, but possibly the need is felt to construct a new diagram that better serves its purpose.

16.2.4 Domain-Specific Framework for Action: Realistic Mathematics Education (RME)

As pointed out by diSessa and Cobb (2004), grand theories and orienting frameworks do not tell the design researcher how to design learning environments. For this purpose, frameworks for action can be useful. Here we discuss Realistic Mathematics Education (RME).

Our research took place in the tradition of RME as developed over the last 40 years at the Freudenthal Institute (Freudenthal 1991; Gravemeijer 1994; Treffers 1987; van den Heuvel-Panhuizen 1996). RME is a theory of mathematics education that offers a pedagogical and didactical philosophy on mathematical learning and teaching as well as on designing educational materials for mathematics education. RME emerged from research and development in mathematics education in the Netherlands in the 1970s and it has since been used and extended, also in other countries.

The central principle of RME is that mathematics should always be meaningful to students. For Freudenthal, mathematics was an extension of common sense, a system of concepts and techniques that human beings had developed in response to phenomena they encountered. For this reason, he advised a so-called *historical phenomenology* of concepts to be taught, a study of how concepts had been developed in relation to particular phenomena. The insights from such a study can be input for the design process (Bakker and Gravemeijer 2006).

The term 'realistic' stresses that problem situations should be 'experientially real' for students (Cobb et al. 1992). This does not necessarily mean that the problem situations are always encountered in daily life. Students can experience an abstract mathematical problem as real when the mathematics of that problem is meaningful to them. Freudenthal's (1991) ideal was that mathematical learning should be an enhancement of common sense. Students should be allowed and encouraged to invent their own strategies and ideas, and they should learn mathematics on their own authority. At the same time, this process should lead to particular end goals. This process is called *guided reinvention*—one of the design heuristics of RME. This heuristic points to the question that underlies much of the RME-based research, namely that of how to support this process of engaging students in meaningful mathematical and statistical problem solving, and using students' contributions to reach certain end goals.

The theory of RME is especially tailored to mathematics education, because it includes specific tenets on and design heuristics for mathematics education. For a description of these tenets we refer to Treffers (1987) and for the design heuristics to Gravemeijer (1994) or Bakker and Gravemeijer (2006).

16.2.5 Methods

The absence of the type of learning aimed for is a common reason to carry out design research. For Bakker's study in statistics education, descriptive, comparative, or evaluative research did not make sense because the type of learning aimed for could not be readily observed in classrooms. Considerable design and research effort first had to be taken to foster specific innovative types of learning. Bakker therefore had to design HLTs with accompanying educational materials that supported the desired type of learning about distribution. Design-based research offers a systematic approach to doing that while simultaneously developing domain-specific theories about how to support such learning for example here on the domain of statistics. In general, DBR researchers first need to create the conditions in which they can develop and test an instruction theory, but to create those conditions they also need research.

Teaching experiment. Bakker designed educational materials with accompanying HLTs in several cycles. Here we focus on the last cycle, involving a teaching experiment in grade 8. Half of the lessons were carried out in a computer lab and as part of them students used two minitools (Cobb et al. 1997), simple Java applets with which they analyzed data sets on, for instance, battery life span, car colours, and salaries (Fig. 16.3). The researcher was responsible for the educational materials and the teacher was responsible for the teaching, though we discussed in advance on a weekly basis both the materials and appropriate teaching style. Three preservice teachers served as assistants and helped with videotaping and interviewing students and with analyzing the data.

In the example that we elaborate we focus on the fourth of a series of ten lessons, each 50 min long. In this specific lesson, students reasoned about larger and larger samples and about the shape of distributions.

Subjects. The teaching experiment was carried out in an eighth-grade class with 30 students in a state school in the center of a Dutch city. The students in this study were being prepared for pre-university (*vwo*) or higher professional education (*havo*). The students in the class reported on here were not used to whole-class discussions, but rather to be "taken by the hand" as the teacher called it; they were characterized by the three research assistants as "passive but willing to cooperate." These students had no prior instruction in statistics; they were acquainted with bar and line graphs, but not with dot plots, histograms, or box plots. Students already knew the mean from calculating their report grades, but mode and median were not introduced until the second half of the educational sequence after variability, data, sampling, and shape had been topics of discussion.

Data collection. The collected data on which the results presented in this chapter are based include student work, field notes, and the audio and video recordings of class activities that the three assistants and researcher made in the classroom. An essential part of the data corpus was the set of mini-interviews we held during the lessons; they varied from about twenty seconds to four minutes, and were meant to find out what concepts and graphs meant for students, or how the minitools were used. These mini-interviews influenced students' learning because they often stimulated reflection. However, we think that the validity of the research was not put in danger by this, since the aim was to find out how students learned to reason with shape or distribution, not whether teaching the sequence in other eighth-grade classes would lead to the same results in the same number of lessons. Furthermore, the interview questions were planned in advance as part of the HLT, and discussed with the assistants.

Retrospective analysis. In this example we do not illustrate how HLTs can be compared with observed learning (see Dierdorp et al. 2011). Here we highlight one type of analysis that in Bakker's case yielded more theoretical insights: a method resembling Glaser and Strauss's constant comparative method (Glaser and Strauss 1967). For the analysis, Bakker watched the videotapes, read the transcripts, and formulated conjectures on students' learning on the basis of transcript episodes. Numbering the conjectures served as useful codes to work with during the analysis. Examples of such codes and conjectures were:

C1. Students divide imaginary data sets into three groups of low, 'average', and high values.

C2. Students either characterize spread as range or look very locally at spread

C3. Students are inclined to think of small samples when first asked about how one could test something (batteries, weight).

C5. What-if questions work well for letting students think of aggregate features of a graph or a situation. What would a weight graph of older students look like? What would the graph look like if a larger sample was taken? What would a larger sample of a good battery brand look like?

C7. Students' notions of spread, distribution, and density are not yet distinguished. When explaining how data are spread out, they often describe the distribution or the density in some area.

C9. Even when students see a large sample of a particular distribution, they often do not see the shape we see in it.

The generated conjectures were tested against other episodes and the rest of the collected data (student work, field observations, and tests) in the next round of anal-

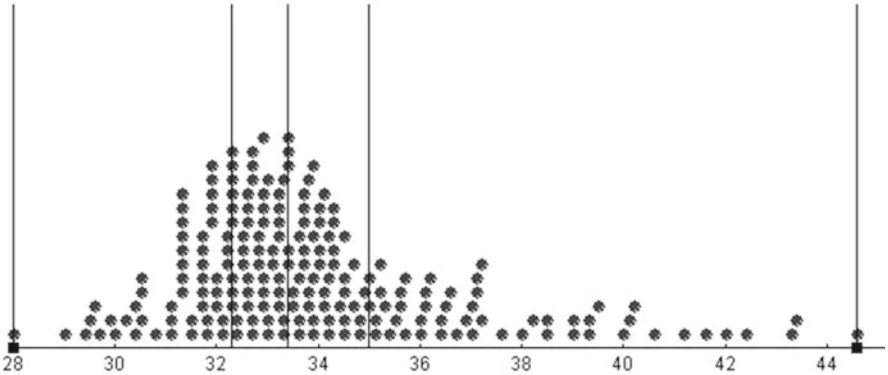


Fig. 16.2 Jeans data with four equal groups option in Minitool 2

ysis by data triangulation. Conjectures that were confirmed remained in the list; conjectures that were refuted were removed from the list. Then the whole generating and testing process was repeated. The aforementioned examples were all confirmed throughout this analysis.

To get a sense of the interrater reliability of the analysis, about one quarter of the episodes including those discussed in this chapter and the conjectures belonging to these episodes were judged by the three assistants who attended the teaching experiment. The amount of agreement among judges was very high: all four judges agreed about 33 out of 35 codes. A code was only accepted if all judges agreed after discussion. We give an example of a code that was finally rejected and one that was accepted. This example stems from the seventh lesson in which two students used the four equal groups option in Minitool 2 for a revised version of the jeans activity. Their task was to advise a jeans factory about frequencies of jeans sizes to be produced (Fig. 16.2).

Sofie Because then you can best see the spread, how it is distributed.
Int. How it is distributed. And how do you see that here [in this graph]?
What do you look at then? (…)
Sofie Well, you can see that, for example, if you put a [vertical] line here,
here a line, and here a line. Then you see here [two lines at the right]
that there is a very large spread in that part, so to speak.

In the first line, Sofie seems to use the terms spread and distributed as almost synonymous. This line was therefore coded with C7, which states that "students' notions of spread, distribution, and density are not yet distinguished. When explaining how data are spread out, they often describe the distribution or the density in some area." In the second line, Sofie appears to look at spread very locally, hence it was coded with C2, which states that "students either characterize spread as range or look very locally at spread."

We also give an example of a code assignment that was dismissed in relation to the same diagram.

Int. What does this tell you? Four equal groups?
Melle Well, I think that most jeans are between 32 and 34 [inches].

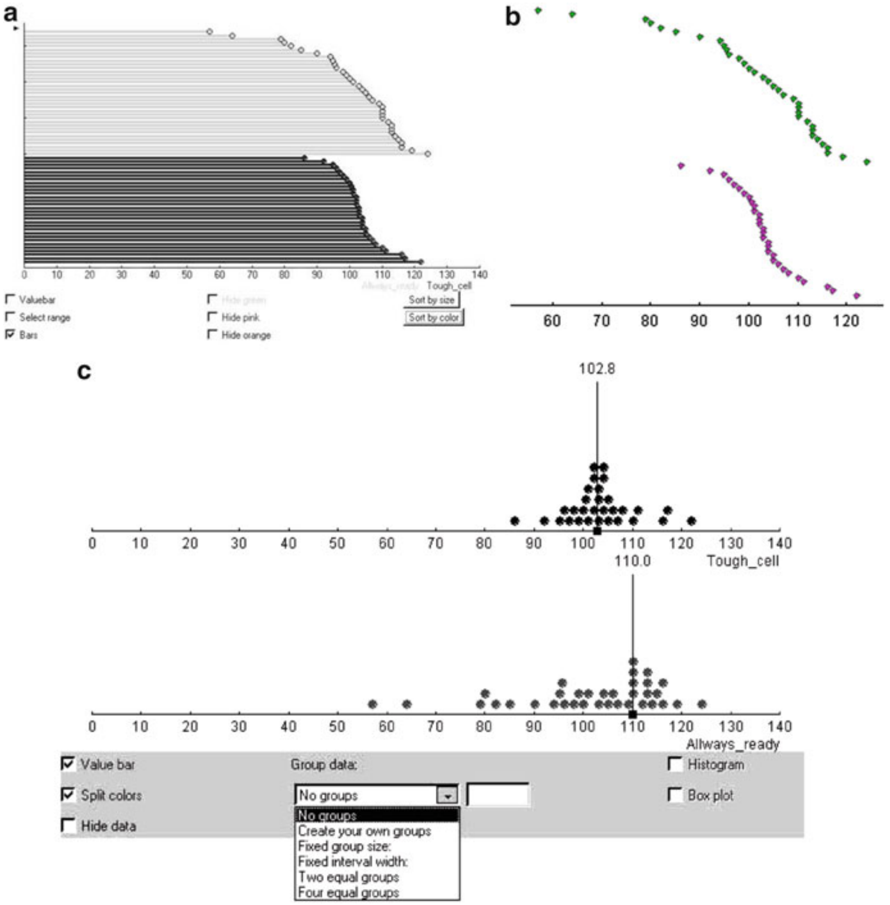
We had originally assigned the code C1 to the this episode (students talk about data sets as consisting of three groups of low, 'average', and high values), because "most jeans are between 32 and 34" implies that below 32 and above 34 the frequencies are relatively low. In the episode, however, this student did not talk about three groups of low, average, and high values or anything equivalent. We therefore removed the code from this episode.

16.2.6 HLT and Retrospective Analysis

To illustrate relationships between theory, method, and results, this section presents the analysis of students' reasoning during one educational activity which was carried out in the fourth lesson. Its goal was to stimulate students to reason about larger and larger samples. We summarize the HLT of that lesson: the learning goal, the activity of growing a sample and the assumptions about students' potential learning processes and about how the teacher could support these processes. We then present the retrospective analysis of three successive phases in growing a sample.

The overall *goal* of the growing samples activity as formulated in the hypothetical learning trajectory for this fourth lesson was to stimulate students' diagrammatic reasononing about shape in relation to sampling and distribution aspects in the context of weight. This implied that students should first make diagrams, then experiment with them and reflect on them. The idea was to start with ideas invented by the students and guide them toward more conventional notions and representations. This process of guiding students toward these culturally accepted concepts and graphs while building on their own inventions is called guided reinvention. We had noted in previous teaching experiments that students were inclined to choose very small samples initially. It proved necessary to stimulate reflection on the disadvantages of such small samples and have them predict what larger samples would look like. Such insights from the analyses of previous teaching experiments helped to better formulate the HLT of a new teaching experiment. More particularly, Bakker assumed that starting with students' initial ideas about small samples and asking for predictions about larger samples would make students aware of various features of distributions.

The *activity* of growing a sample consisted of three phases of making sketches of a hypothetical situation and comparing those sketches with graphs displaying real data sets. In the first phase students had to make a graph of their own choice of a predicted weight data set with sample size 10. The results were discussed by the teacher to challenge this small sample size, and in the subsequent phases students had to predict larger data sets, one class and three classes in the second phase, and all students in the province in the third phase. Thus, three such phases took place as



□ Valuebar
□ Select range
☑ Bars
□ Hide green
□ Hide pink
□ Hide orange

☑ Value bar
☑ Split colors
□ Hide data
□ Histogram
□ Box plot

Fig. 16.3 (**a**) Minitool 1 showing a value-bar graph of battery life spans in hours of two brands. (**b**) Minitool 1, but with bars hidden. (**c**) Minitool 2 showing a dot plot of the same data sets

described and analyzed below. Aiming for guided reinvention, the teacher and researcher tried to strike a balance between engaging students in statistical reasoning and allowing their own terminology on the one hand, and guiding them in using conventional and more precise notions and graphical representations on the other. Figure 16.3b is the result of focusing only on the endpoints of the value bars in Fig. 16.3a. Figure 16.3c is the result of these endpoints falling down vertically on the x-axis. In this way, students can learn to understand the relationship between value-bar graphs and dot plots, and what distribution features in different representations look like (Bakker and Hoffmann 2005).

16.2.6.1 Analysis of the First Phase of Growing a Sample

The text of the student activity sheet for the fourth lesson contained a number of tasks that we cite in the following subsections. The sheet started as follows:

> Last week you made graphs of predicted data for a balloon pilot. During this lesson you will get to see real weight data of students from another school. We are going to investigate the influence of the sample size on the shape of the graph.
>
> Task a. Predict a graph of ten data values, for example with the dots of minitool 2.

The sample size of ten was chosen because the students had found that size reasonable after the first lesson in the context of testing the life span of batteries. Figure 16.4 shows examples for three different types of diagrams the students made to show their predictions: there were three value-bar graphs (such as in minitool 1—e.g., Ruud's diagram), eight with only the endpoints (such as with the option of minitool 1 to "hide bars"—e.g., Chris's diagram) and the remaining nineteen plots were dot plots (such as in minitool 2—e.g., Sandra's diagram). For the remainder of this section, the figures and written explanations of these three students are demonstrated, because their work gives an impression of the variety of the whole class. Those three students were chosen because their diagrams represent all types of diagrams made in this class, also for other phases of growing a sample.

To stimulate the reflection on the graphs, the teacher showed three samples of ten data points on the blackboard and students had to compare their own graphs (Fig. 16.4) with the graphs of the real data sets (Fig. 16.5).

> Task b. You get to see three different samples of size 10. Are they different from your own prediction? Describe the differences.

The reason for showing three small samples was to show the variation among these samples. There were no clear indications, though, that students conceived this variation as a sign that the sample size was too small for drawing conclusions, but they generally agreed that larger samples were more reliable. The point relevant to the analysis is that students started using predicates to describe aggregate features of the graphs. The written answers of the three students were the following:

Ruud	Mine looks very much like what is on the blackboard.
Chris	The middle-most [diagram on the blackboard] best resembles mine because the weights are close together and that is also the case in my graph. It lies between 35 and 75 [kg].
Sandra	The other [real data] are more weights together and mine are further apart.

Ruud's answer is not very specific, like most of the written answers in the first phase of growing samples. Chris used the predicate "close together" and added numbers to indicate the range, probably as an indication of spread. Sandra used such terms as "together" and "further apart," which address spread. The students in the class used common predicates such as "together," "spread out" and "further apart" to describe features of the data set or the graph. For the analysis it is important to

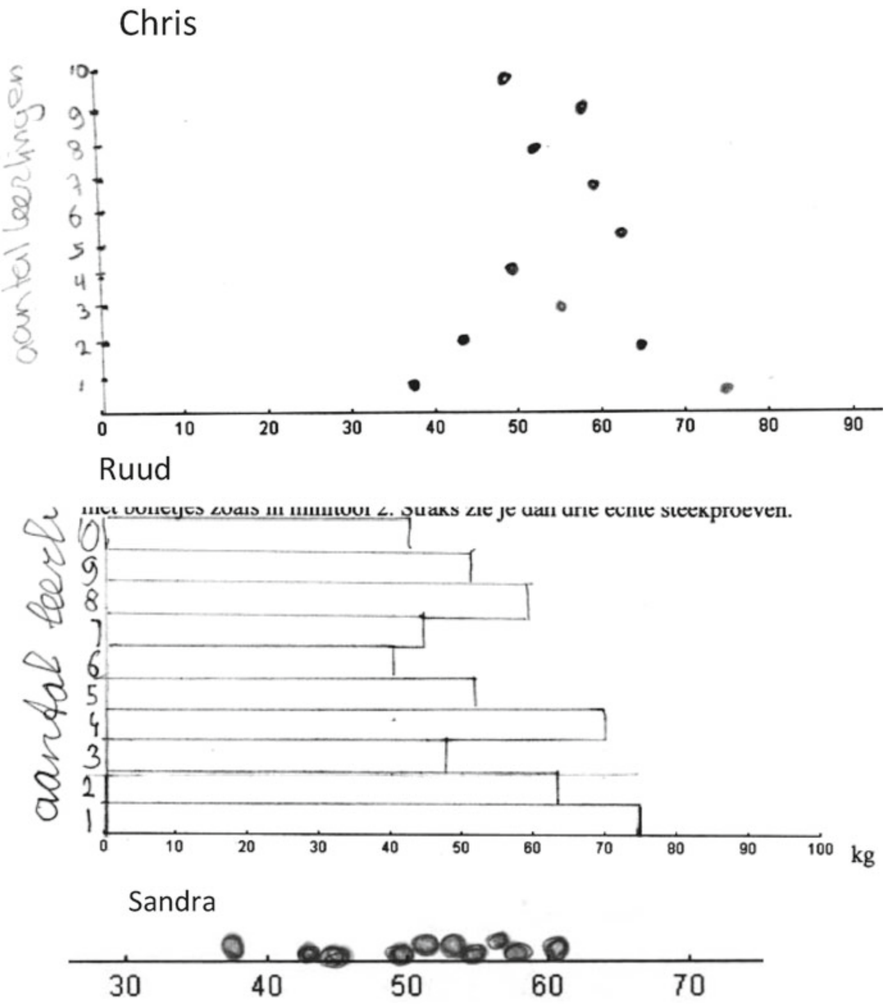
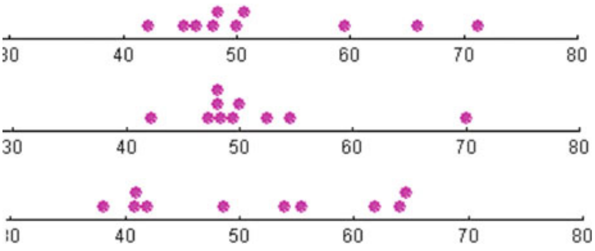


Fig. 16.4 Student predictions (Ruud, Chris, and Sandra) for ten data points (weight in kg) (Bakker 2004a, p. 219)

Fig. 16.5 Three real data sets in minitool 2 (Bakker 2004a, p. 219)



note that the students used predicates (together, apart) and no nouns (spread, average) in this first phase of growing samples. Spread can only become an object-like concept, something that can be talked about and reasoned with, if it is a noun. In the semiotic theory of Peirce, such transitions from the predicate "the dots are spread out" to "the spread is large" are important steps in the formation of concepts (see Bakker and Derry 2011, for our view on concept formation).

16.2.6.2 Analysis of the Second Phase of Growing a Sample

The students generally understood that larger samples would be more reliable. With the feedback students had received after discussing the samples of ten data points in dot plots, students had to predict the weight graph of a whole class of 27 students and of three classes with 67 students (27 and 67 were the sample sizes of the real data sets of eighth graders of another school).

> Task c. We will now have a look how the graph changes with larger samples. Predict a sample of 27 students (one class) and of 67 students (three classes).
>
> Task d. You now get to see real samples of those sizes. Describe the differences. You can use words such as majority, outliers, spread, average.

During this second phase, all of the students made dot plots, probably because the teacher had shown dot plots on the blackboard, and because dot plots are less laborious to draw than value bars (only one student started with a value-bar graph for the sample of 27, but switched to a dot plot for the sample of 67). The hint on statistical terms was added to make sure that students' answers would not be too superficial as (often happened before) and to stimulate them to use such notions in their reasoning. It was also important for the research to know what these terms meant to them. When the teacher showed the two graphs with real data, once again there was a short class discussion in which the teacher capitalized on the question of why most student predictions now looked pretty much like what was on the blackboard, whereas with the earlier predictions there was much more variation. No student had a reasonable explanation, which indicates that this was an advanced question. The figures of the same three students are presented in Figs. 16.6 and 16.7 and their written explanations were:

Ruud My spread is different.

Chris Mine resembles the sample, but I have more people around a certain weight and I do not really have outliers, because I have 10 about the 70 and 80 and the real sample has only 6 around the 70 and 80.

Sandra With the 27 there are outliers and there is spread; with the 67 there are more together and more around the average.

Here, Ruud addressed the issue of spread ("my spread is different"). Chris was more explicit about a particular area in her graph, the category of high values. She also correctly used the term "sample," which was newly introduced in the second lesson. Sandra used the term "outliers" at this stage, by which students meant "extreme values," which did not necessarily mean exceptional or suspect values.

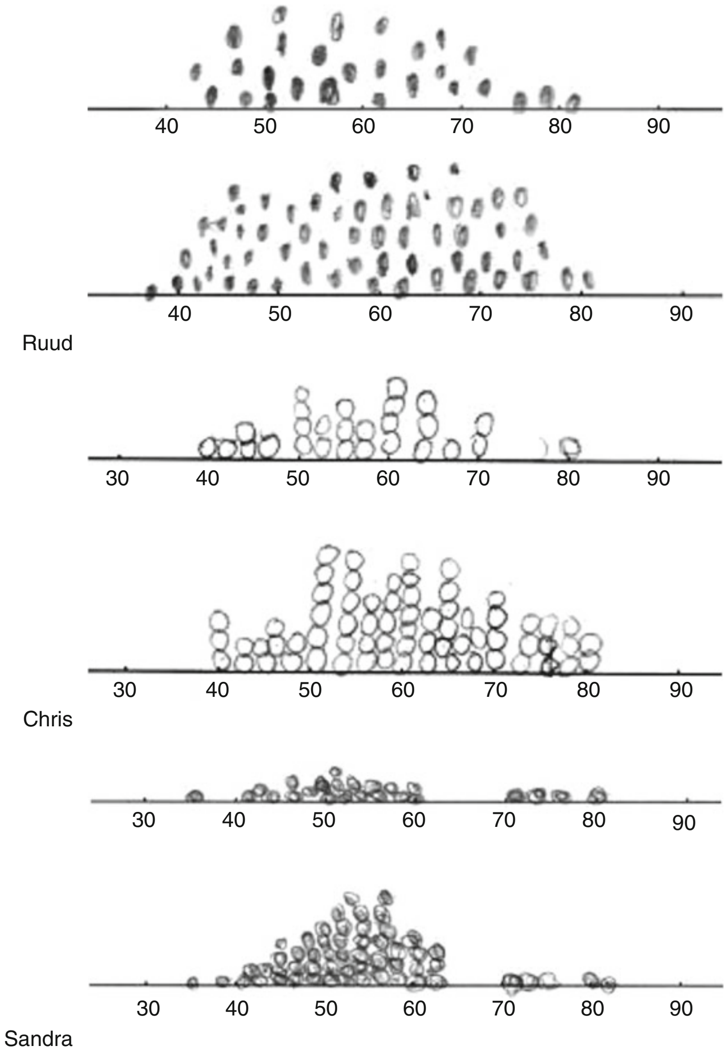


Fig. 16.6 Predicted graphs for one class (n = 27, top plot) and three classes (n = 67, bottom plot) by Ruud, Chris, and Sandra (Bakker 2004a, p. 222)

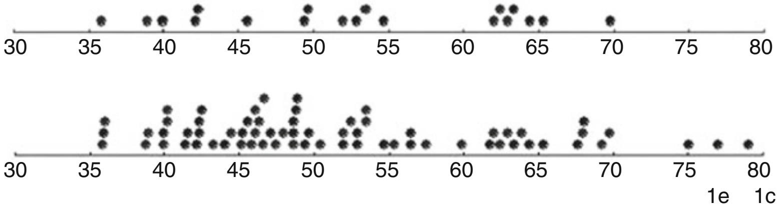


Fig. 16.7 Real data sets of size 27 and 67 of students from another school (Bakker 2004a, p. 222)

She also seemed to locate the average somewhere and to understand that many students are about average. These examples illustrate that students used statistical notions for describing properties of the data and diagrams.

In contrast to the first phase of growing a sample, students used nouns instead of just predicates for comparing the diagrams. Like others Ruud used the noun "spread" ("my spread is different") whereas students earlier used only predicates such as "spread out" or "further apart" (e.g., Sandra). Of course, this does not always imply that if students use these nouns that they are thinking of the right concept. Statistically, however, it makes a difference whether we say, "the dots are spread out" or "the spread is large." In the latter case, spread is an object-like entity that can have particular aggregate characteristics that can be measured, for instance by the range, the interquartile range, or the standard deviation. Other notions such as outliers, sample, and average, are now used as nouns, that is as conceptual objects that can be talked about and reasoned with.

16.2.6.3 Analysis of the Third Phase of Growing a Sample

The aim of the hypothetical learning trajectory was that students would come to draw continuous shapes and reason about them using statistical terms. During teaching experiments in the seventh-grade experiments (Bakker and Gravemeijer 2004), reasoning with continuous shapes turned out to be difficult to accomplish, even if it was asked for. It often seemed impossible to nudge students toward drawing the general, continuous shape of data sets represented in dot plots. At best, students drew spiky lines just above the dots. This underlines that students have to construct something new (a notion of signal, shape, or distribution) with which they can look differently at the data or the variable phenomenon.

In this last phase of growing the sample, the task was to make a graph showing data of all students in the city, not necessarily with dots. The intention of asking this was to stimulate students to use continuous shapes and dynamically relate samples to populations, without making this distinction between sample and population explicit yet. The conjecture was that this transition from a discrete plurality of data values to a continuous entity of a distribution is important to foster a notion of distribution as an object-like entity with which students could model data and describe aggregate properties of data sets. The task proceeded as follows:

> Task e. Make a weight graph of a sample of all eighth graders in the city. You need not draw dots. It is the shape of the graph that is important.
>
> Task f. Describe the shape of your graph and explain why you have drawn that shape.

The figures of the same three students are presented in Fig. 16.8 and their written explanations were:

Ruud Because the average [values are] roughly between 50 and 60 kg.
Chris I think it is a pyramid shape. I have drawn my graph like that because I found it easy to make and easy to read.
Sandra Because most are around the average and there are outliers at 30 and 80 [kg].

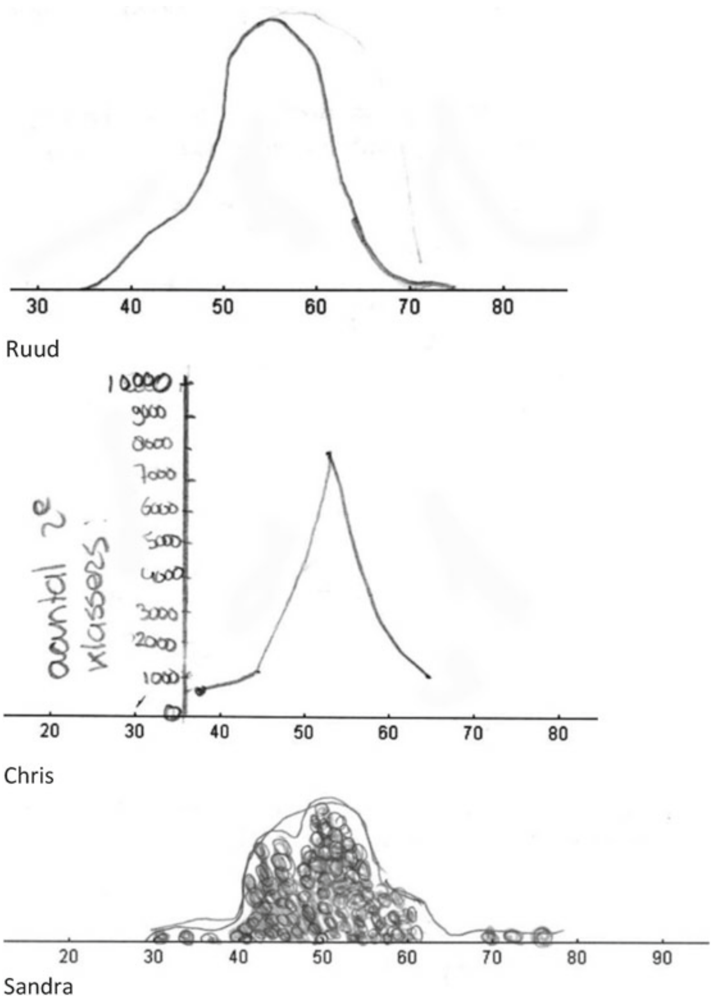


Fig. 16.8 Predicted graphs for all students in the city by Ruud, Chris, and Sandra (Bakker 2004a, p. 224)

Ruud's answer focused on the average group. During an interview after the fourth lesson, Ruud like three other students literally called his graph a "bell shape," though he had probably not encountered that term in a school situation before. This is probably a case of *reinvention*. Chris's graph was probably inspired by line graphs that the students made during mathematics lessons. She introduced the vertical axis with frequency, though such graphs had not been used before in the statistics course. Sandra may have started with the dots and then drawn the continuous shape.

In this third phase of growing a sample, 23 students drew a bump shape. The words they used for the shapes were pyramid (three students), semicircle (one), and bell shape (four). Many students drew continuous shapes but these were all symmetrical. Since weight distributions are not symmetrical and because skewness is an important concept, a subsequent lesson addressed asymmetrical shapes in relation to the weight data (see Bakker 2004b).

16.2.7 Reflection on the Example

The research question we addressed in the example is: How can coherent reasoning about distribution be promoted in relation to data, variability, and sampling in a way that is meaningful for students with little statistical background? We now discuss those key elements for the educational activity and speculate about what can be learned from the analysis presented here.

The activity of growing a sample involved short phases of constructing diagrams of new hypothetical situations, and comparing these with other diagrams of a real sample of the same size. The activity has a broader empirical basis than just the teaching experiment reported in this chapter, because it emerged from a previous teaching experiment (Bakker and Gravemeijer 2004) as a way to address shape as a pattern in variability.

To theoretically generalize the results, Bakker analyzed students' reasoning as an instance of diagrammatic reasoning, which typically involves constructing diagrams, experimenting with them, and reflecting on the results of the previous two steps. In this growing samples activity, the quick alternation between prediction and reflection during diagrammatic reasoning appears to create ample opportunities for concept formation, for instance of spread.

In the first phase involving the prediction of a small data set, students noted that the data were more spread out, but in subsequent phases, students wrote or said that the spread was large. From the terms used in this fourth lesson, we conclude that many statistical concepts such as center (average, majority), spread (range and range of subsets of data), and shape had become topics of discussion (object-like entities) during the growing samples activity. Some of these words were used in a rather unconventional way, which implies that students needed more guidance at this point. Shape became a topic of discussion as students predicted that the shape of the graph would be a semicircle, a pyramid, or a bell shape, and this was exactly what the HLT targeted. Given the students' minimal background in statistics and the fact that this was only the fourth lesson of the sequence, the results were promising. Note, however, that such activities cannot simply be repeated in other contexts; they need to be adjusted to local circumstances if they are to be applied in other situations.

The instructional activity of growing samples later became a connecting thread in Ben-Zvi's research in Israel, where it also worked to help students develop statistical concepts in relation to each other (Ben-Zvi et al. 2012). This implies that this instructional idea was transferable to other contexts. The transferability of instructional ideas from the USA to the Netherlands to Israel, even to higher levels of education, illustrates that generalization in DBR can take place across contexts, cultures and age group.

16.2.8 Final Remarks

The example presented in Sect. 16.2 was intended to substantiate the issues discussed in Sect. 16.1, and we hope that readers will have a sense of what DBR could look like and feel invited to read more about it. It should be noted that there are many variants of DBR. Some are more focused on theory, some more on empirically grounded products. Some start with predetermined learning outcomes, others have more open-ended goals (cf. Engeström 2011). DBR may be a challenging research approach but it is in our experience also a very rewarding one given the products and insights that can be gained.

Acknowledgments The research was funded by the Netherlands Organization for Scientific Research under grant number 575-36-003B. The writing of this chapter was made possible with a grant from the Educational and Learning Sciences Utrecht awarded to Arthur Bakker. Section 2.6 is based on Bakker (2004b). We thank our Master students in our Research Methodology courses for their feedback on earlier versions of this manuscript. Angelika Bikner-Ahsbahs's and reviewers' careful reading has also helped us tremendously. We also acknowledge PhD students Adri Dierdorp, Al Jupri, and Victor Antwi, and our colleague Frans van Galen for their helpful comments, and Nathalie Kuijpers and Norma Presmeg for correcting this manuscript.

Appendix: Structure of a DBR Project with Illustrations

In line with Oost and Markenhof (2010), we formulate the following general criteria for any research project:

1. The research should be **anchored** in the literature.
2. The research aim should be **relevant**, both in theoretical and practical terms.
3. The formulation of aim and questions should be **precise**, i.e. using concepts and definitions in the correct way.
4. The method used should be **functional** in answering the research question(s).
5. The overall structure of the research project should be **consistent**, i.e. title, aim, theory, question, method and results should form a coherent chain of reasoning.

In this appendix we present a structure of general points of attention during DBR and specifications for our statistics education example, including references to relevant sections in the chapter. In this structure these criteria are bolded. This structure could function as the blueprint of a book or article on a DBR project.

	General points	Examples
Introduction:	1. Choose a topic	1. Statistics education at the middle school level
	2. Identify common problems	2. Statistics as a set of unrelated concepts and techniques
	3. Identify knowledge gap and relevance	3. How middle school students can be supported to develop a concept of distribution and related statistical concepts
	4. Choose mathematical learning goals	4. Understanding of distribution (2.1)

(continued)

	General points	Examples
Literature review forms the basis for formulating the research aim (the research has to be **anchored** and **relevant**)		
Research aim:	It has to be clear whether an aim is descriptive, explanatory, evaluative, advisory etc. (1.2.2)	Contribute to an empirically and theoretically grounded instruction theory for statistics education at the middle school level (advisory aim) (2.1)
Research aim has to be narrowed down to a research question and possibly subquestions with the help of different theories		
Literature review (theoretical background):	Orienting frameworks	Semiotics (2.3)
	Frameworks for action	Theories on learning with computer tools
	Domain-specific learning theories (1.2.8)	Realistic Mathematics Education (2.4)
With the help of theoretical constructs the research question(s) can be formulated (the formulation has to be **precise**)		
Research question:	Zoom in what knowledge is required to achieve the research aim	How can students with little statistical background develop a notion of distribution?
It should be underpinned why this research question requires DBR (the method should be **functional**)		
Research approach:	The lack of the type of learning aimed for is a common reason to carry out DBR: It has to be enacted so it can be studied	Dutch statistics education was atomistic: Textbooks addressed mean, median, mode, and different graphical representations one by one. Software was hardly used. Hence the type of learning aimed for had to be enacted.
Using a research method involves several research instruments and techniques		
Research instruments and techniques	Research instrument that connects different theories and concrete experiences in the form of testable hypotheses.	Series of hypothetical learning trajectories (HLTs)
	1. Identify students' prior knowledge	1. Prior interviews and pretest
	2. Professional development of teacher	2. Preparatory meetings with teacher
	3. Interview schemes and planning	3. Mini-interviews, observation scheme
	4. Intermediate feedback and reflection with teacher	4. Debrief sessions with teacher
	5. Determine learning yield (1.4.2)	5. Posttest
Design	Design guidelines	Guided reinvention; Historical and didactical phenomenology (2.4)
Data analysis	Hypotheses have to be tested by comparison of hypothetical and observed learning. Additional analyses may be necessary (1.4.3)	Comparison of hypothetical and observed learning
		Constant comparative method of generating conjectures and testing them on the remaining data sources (2.6)

(continued)

	General points	Examples
Results	Insights into patterns in learning and means of supporting such learning	Series of HLTs as progressive diagrammatic reasoning about growing samples (2.6)
Discussion	Theoretical and practical yield	Concrete example of an historical and didactical phenomenology in statistics education
		Application of semiotics in an educational domain
		Insights into computer use in the mathematics classroom
		Series of learning activities
		Improved computer tools
The aim, theory, question, method and results should be aligned (the research has to be **consistent**)		

References

Akkerman, S. F., Admiraal, W., Brekelmans, M., & Oost, H. (2008). Auditing quality of research in social sciences. *Quality & Quantity, 42*, 257–274.

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational Researcher, 41*, 16–25.

Artigue, M. (1988). Ingénierie didactique [Didactical engineering]. In M. Artigue, G. Brousseau, J. Brun, Y. Chevallard, F. Conne, & G. Vergnaud (Eds.), *Didactique des mathematiques* [Didactics of mathematics]. Paris: Delachaux et Niestlé.

Bakkenes, I., Vermunt, J. D., & Wubbels, T. (2010). Teachers learning in the context of educational innovation: Learning activities and learning outcomes of experienced teachers. *Learning and Instruction, 20*(6), 533–548.

Bakker, A. (2004a). *Design research in statistics education: On symbolizing and computer tools*. Utrecht: CD-Bèta Press.

Bakker, A. (2004b). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64–83. Online http://www.stat.auckland.ac.nz/~iase/serj/SERJ3(2)_Bakker.pdf

Bakker, A. (2007). Diagrammatic reasoning and hypostatic abstraction in statistics education. *Semiotica, 164*, 9–29.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*, 5–26.

Bakker, A., & Gravemeijer, K. P. E. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 147–168). Dordrecht: Kluwer.

Bakker, A., & Gravemeijer, K. P. E. (2006). An historical phenomenology of mean and median. *Educational Studies in Mathematics, 62*(2), 149–168.

Bakker, A., & Hoffmann, M. (2005). Diagrammatic reasoning as the basis for developing concepts: A semiotic analysis of students' learning about statistical distribution. *Educational Studies in Mathematics, 60*, 333–358.

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM The International Journal on Mathematics Education, 44*, 913–925.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clement, A. J. Bishop, C. Keitel, J. Kilpatrick, & A. Y. L. Leung (Eds.), *Third international handbook on mathematics education* (pp. 643–689). New York: Springer. doi:10.1007/978-1-4614-4684-2_21.

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2*, 141–178.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129*, 452–456.

Cobb, P., & Whitenack, J. W. (1996). A method for conducting longitudinal analyses of classroom videorecordings and transcripts. *Educational Studies in Mathematics, 30*(3), 213–228.

Cobb, P., Yackel, E., & Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. *Journal for Research in Mathematics Education, 23*, 2–33.1.

Cobb, P., Gravemeijer, K.P.E., Bowers, J., & McClain, K. (1997). *Statistical Minitools*. Designed for Vanderbilt University, TN, USA. Programmed and revised (2001) at the Freudenthal Institute, Utrecht University, the Netherlands.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003a). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

Cobb, P., McClain, K., & Gravemeijer, K. P. E. (2003b). Learning about statistical covariation. *Cognition and Instruction, 21*, 1–78.

Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15–22). New York: Springer.

Cook, T. (2002). Randomized experiments in education: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175–199.

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (2nd ed.). Upper Saddle River: Pearson Education.

Creswell, J. W. (2007). *Qualitative inquiry and research design. Choosing among five traditions* (2nd ed.). Thousand Oaks: Sage.

De Jong, R., & Wijers, M. (1993). *Ontwikkelingsonderzoek: Theorie en praktijk* [Developmental research: Theory and practice]. Utrecht: NVORWO.

Denscombe, M. (2007). *The good research guide* (3rd ed.). Maidenhead: Open University Press.

Dierdorp, A., Bakker, A., Eijkelhof, H. M. C., & Van Maanen, J. A. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning, 13*, 132–151.

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. *Educational Researcher, 32*(1), 77–103.

Drijvers, P. H. M. (2003). *Learning algebra in a computer algebra environment: Design research on the understanding of the concept of parameter*. Utrecht: CD-Beta Press.

Edelson, D. C. (2002). Design research: What we learn when we engage in design. *Journal of the Learning Sciences, 11*, 105–121.

Educational Researcher. (2003). *Special issue on design-based research collective, 32*(1–2).

Educational Psychologist. (2004). *Special issue design-based research methods for studying learning in context, 39*(4).

Engeström, Y. (2011). From design experiments to formative interventions. *Theory and Psychology, 21*(5), 598–628.

Fosnot, C. T., & Dolk, M. (2001). *Young mathematicians at work. Constructing number sense, addition, and subtraction*. Portsmouth: Heinemann.

Freudenthal, H. (1978). *Weeding and sowing: Preface to a science of mathematical education*. Dordrecht: Reidel.

Freudenthal, H. (1988). Ontwikkelingsonderzoek [Developmental research]. In K. Gravemeijer & K. Koster (Eds.), *Onderzoek, ontwikkeling en ontwikkelingsonderzoek* [Research, development and developmental research]. Universiteit Utrecht, the Netherlands: OW&OC.

Freudenthal, H. (1991). *Revisiting mathematics education: China lectures*. Dordrecht: Kluwer.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers, 30*(3), 527–535.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal of Research in Mathematics Education., 32*(2), 124–158.

Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In C. Geertz (Ed.), *The interpretation of cultures: Selected essays* (pp. 3–30). New York: Basic Books.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Goffree, F. (1979). *Leren onderwijzen met Wiskobas. Onderwijsontwikkelingsonderzoek 'Wiskunde en Didaktiek' op de pedagogische akademie* [Learning to teach Wiskobas. Educational development research]. Rijksuniversiteit Utrecht, The Netherlands.

Gravemeijer, K. P. E. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education, 25*(5), 443–471.

Gravemeijer, K. P. E., & Cobb, P. (2006). Design research from a learning design perspective. In J. Van den Akker, K. P. E. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 17–51). London: Routledge.

Gravemeijer, K. P. E., & Koster, K. (Eds.). (1988). *Onderzoek, ontwikkeling en ontwikkelingsonderzoek* [Research, development, and developmental research]. Utrecht: OW&OC.

Guba, E. G. (1981). Criteria for assessing trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal, 29*(2), 75–91.

Hoffmann, M. H. G. (2002). Peirce's "diagrammatic reasoning" as a solution of the learning paradox. In G. Debrock (Ed.), *Process pragmatism: Essays on a quiet philosophical revolution* (pp. 147–174). Amsterdam: Rodopi Press.

Hoyles, C., Noss, R., Kent, P., & Bakker, A. (2010). *Improving mathematics at work: The need for techno-mathematical literacies*. Abingdon: Routledge.

Journal of the Learning Sciences (2004). Special issue on design-based research, 13(1), guest-edited by S. Barab and K. Squire.

Kanselaar, G. (1993). Ontwikkelingsonderzoek bezien vanuit de rol van de advocaat van de duivel [Design research: Taking the position of the devil's advocate]. In R. de Jong & M. Wijers (Red.) (Eds.), *Ontwikkelingsonderzoek, theorie en praktijk*. Utrecht: NVORWO.

Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston: National Council of Teachers of Mathematics.

Mathematical Thinking and Learning (2004). *Special issue on learning trajectories in mathematics education*, guest-edited by D. H. Clements and J. Sarama, 6(2).

Lehrer, R., & Schauble, L. (2001). *Accounting for contingency in design experiments.* Paper presented at the annual meeting of the American Education Research Association, Seattle.

Lewin, K. (1951). Problems of research in social psychology. In D. Cartwright (Ed.), *Field theory in social science; selected theoretical papers*. New York: Harper & Row.

Lijnse, P. L. (1995). "Developmental Research" as a way to an empirically based "didactical structure" of science. *Science Education, 29*(2), 189–199.

Lijnse, P. L., & Klaassen, K. (2004). Didactical structures as an outcome of research on teaching-learning sequences? *International Journal of Science Education, 26*(5), 537–554.

Maso, I., & Smaling, A. (1998). *Kwalitatief onderzoek: praktijk en theorie* [Qualitative research: Practice and theory]. Amsterdam: Boom.

Maxwell, J. A. (2004). Causal explanation, qualitative research and scientific inquiry in education. *Educational Researcher, 33*(2), 3–11.

McClain, K., & Cobb, P. (2001). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45*, 103–129.

McKenney, S., & Reeves, T. (2012). *Conducting educational design research*. London: Routledge.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: A sourcebook of new methods*. Beverly Hills: Sage.

Nathan, M. J., & Kim, S. (2009). Regulation of teacher elicitations in the mathematics classroom. *Cognition and Instruction, 27*(2), 91–120.

Olsen, D. R. (2004). The triumph of hope over experience in the search for "what works": A response to Slavin. *Educational Researcher, 33*(1), 24–26.

Oost, H., & Markenhof, A. (2010). *Een onderzoek voorbereiden* [Preparing research]. Amersfoort: Thieme Meulenhoff.

Opie, C. (2004). *Doing educational research*. London: Sage.

Paas, F. (2005). Design experiments: Neither a design nor an experiment. In C. P. Constantinou, D. Demetriou, A. Evagorou, M. Evagorou, A. Kofteros, M. Michael, C. Nicolaou, D. Papademetriou, & N. Papadouris (Eds.), *Integrating multiple perspectives on effective learning environments. Proceedings of 11th biennial meeting of the European Association for Research on Learning and Instruction* (pp. 901–902). Nicosia: University of Cyprus.

Peirce, C. S. (1976). *The new elements of mathematics* (C. Eisele, Ed.). The Hague: Mouton.

Peirce, C. S. (CP). *Collected papers of Charles Sanders Peirce* 1931–1958. In C. Hartshorne & P. Weiss (Eds.), Cambridge, MA: Harvard University Press.

Plomp, T. (2007). Educational design research: An introduction. In N. Nieveen & T. Plomp (Eds.), *An introduction to educational design research* (pp. 9–35). Enschede: SLO.

Plomp, T., & Nieveen, N. (Eds.). (2007). *An introduction to educational design research*. Enschede: SLO.

Romberg, T. A. (1973). *Development research. Overview of how development-based research works in practice.* Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin-Madison, Madison.

Sandoval, W. A., & Bell, P. (2004). Design-dased research methods for studying learning in context: Introduction. *Educational Psychologist, 39*(4), 199–201.

Sfard, A., & Linchevski, L. (1992). The gains and the pitfalls of reification — The case of algebra. *Educational Studies in Mathematics, 26*(2–3), 191–228.

Simon, M. (1995). Reconstructing mathematics pedagogy from a constructivistic perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145.

Slavin, R. E. (2002). Evidence-based educational policies: Transforming educational practice and research. *Educational Researcher, 31*, 15–21.

Smit, J., & Van Eerde, H. A. A. (2011). A teacher's learning process in dual design research: Learning to scaffold language in a multilingual mathematics classroom. *ZDM The International Journal on Mathematics Education, 43*(6–7), 889–900.

Smit, J., van Eerde, H. A. A., & Bakker, A. (2013). A conceptualisation of whole-class scaffolding. *British Educational Research Journal, 39*(5), 817–834.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiments methodology: Underlying principles and essential elements. In R. Lesh & A. E. Kelly (Eds.), *Research design in mathematics and science education* (pp. 267–307). Hillsdale: Erlbaum.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research techniques and procedures for developing grounded theory* (2nd ed.). London: Sage.

Tall, D., Thomas, M., Davis, G., Gray, E., & Simpson, A. (2000). What is the object of the encapsulation of a process? *Journal of Mathematical Behavior, 18*, 223–241.

Treffers, A. (1987). *Three dimensions. A model of goal and theory description in mathematics instruction. The Wiskobas project*. Dordrecht: Kluwer.

Van den Akker, J. (1999). Principles and methods of development research. In J. van den Akker, R. M. Branch, K. Gustafson, N. Nieveen, & T. Plomp (Eds.), *Design approaches and tools in education and training* (pp. 1–14). Boston: Kluwer.

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006). *Educational design research*. London: Routledge.

Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht: CD-Bèta Press.

Van Nes, F., & Doorman, L. M. (2010). The interaction between multimedia data analysis and theory development in design research. *Mathematics Education Research Journal, 22*(1), 6–30.

Wittmann, E. C. (1992). Didaktik der Mathematik als Ingenieurwissenschaft. [Didactics of mathematics as an engineering science.]. *Zentralblatt für Didaktik der Mathematik, 3*, 119–121.

Yin, R. K. (2009). *Case study research: Design and methods*. Thousand Oaks: Sage.