The potential of a classroom network to support teacher feedback

A study in statistics education

Jos Tolboom

Tolboom, J.L.J.

The potential of a classroom network to support teacher feedback; A study in statistics education

ISBN: 978-90-367-5442-2

NUR: 846

Print: Grafimedia, University of Groningen

Beeld omslag: Steven Sloof

rijksuniversiteit groningen

The potential of a classroom network to support teacher feedback

A study in statistics education

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vrijdag 15 juni 2012
om 14.30 uur

door

Johannes Leonardus Jacobus Tolboom
geboren op 3 september 1966
te Emmen

Promotores:	Prof. dr. H.W. Broer
	Prof. dr. W.A.J.M. Kuiper
Beoordelingscommissie:	Prof. dr. J.J.H. van den Akker
	Prof. dr. J.A. van Maanen
	Prof. dr. E.C. Wit

Table of contents

Chapter 1 Introduction

In this chapter we give an outline of this study and the rationale for it. In short, we describe all of the basic elements (feedback, information and communication technology, statistics education) that together form the skeleton of this research. We describe an initial study. We then formulate the main research question and sketch very briefly the research methodology applied in this study. The chapter ends with an overview of this research.

1.1 Rationale for this research

The inducement of this study is a 1998 curriculum reform in senior secondary education in The Netherlands (Staatsblad, 1997). This reform encompassed, on the one hand, new learning goals and contents, reallocation of the aimed study load, and on the other hand suggested pedagogical changes. The reform was called "The second stage" (Dutch: De tweede fase). The use of s*econd* aims at the second and last part of the curriculum of secondary education: the secondary grades 10 and 11 of *senior secondary education* (Dutch: havo) and at the secondary grades 10-12 of *pre university education* (Dutch: vwo). Besides changes in the written curriculum, a new pedagogical approach was suggested. This approach was called "The study house", often sketched as follows: students working self-regulated in a room larger than a traditional classroom, deliberating in small groups or working alone on problems, with the teachers walking between them, available for consult. This represents a change in the role of the teacher from "a spider in the web" to "a guide by the side".

In order to create time for teachers' coaching activities, school managers often decided to reduce time for the traditional face-to-face teaching practices. Students experienced an increase in workload, resulting amongst others in a students' strike on December 6th 1999, when about 20,000 students showed up in The Hague to demonstrate that The Second Stage was not academically feasible. Teachers suffered from a similar increase in workload.

Van Streun (2001) analysed the combination of official goals and the actual preconditions a school organization then faced, and concluded that the design of the second stage is based on at least seven systematic errors, from which we quote three:

1. The actual available study load, in which the new learning goals had to be mastered, is at most 70% of the study load that was used to determine the national exam standards.
2. Study programs are overloaded and fragmented, leading to a significantly heavier task for the teachers, having fewer lessons actually available.
3. Curricular overload and fragmentation go, with other factors, at the expense of the intended pedagogical innovation.

Apparently, according to this analysis, teachers and students perceived an existing curricular overload.

Above, we sketched more or less practical conditions that decreased teaching time in Dutch upper secondary education. But there are more timeless arguments for this study to be conducted. *Mathematics*, the overarching learning domain of this research, is perceived by students to be difficult (Berch & Mazzocco, 2007; Geary, 2010; Hembree,

1990; Küchemann, 1981; Rosnick & Clement, 1980). This is supported by the fact that there are specific terms to describe problems with the learning of mathematics: *dyscalculia* (McCloskey, Caramazza, & Basili, 1985), *innumeracy* (Paulos, 1988) and *mathematics learning disability* (or: *deficit*) (MLD) (Ginsburg, 1997). Especially poor performing students in mathematics consider statistics, the specific learning domain of this study, to be difficult (Hong & Karstensson, 2002). One interpretation of 'difficult' can be 'difficult to study without any help'. As we will illustrate in the next section (1.2) and in chapter 2 in more detail, *feedback* on the learner's work can be a strong help for those struggling with the difficulty of mathematics.

Summarising, in short, recent developments have urged the perpetual problem in mathematics education of the perception, by teachers and their students, of a lack of time to meet the learning goals. In the next section we will address the question: what could feedback contribute to the solution of this problem? The working hypothesis of this study is that an intensification of feedback can make the mathematics lessons more effective and that the application of information and communication technology (ICT) can support this feedback.

1.2 Why feedback?

All learning includes dialoguing. This can be a dialogue between a teacher and a student (like the archetypical dialogue between Socrates and Menon's slave (Plato, 2002)), a dialogue between learners (peer learning), or an internal dialogue (dialogue with one's self). For improving the efficacy of the dialogue with respect to learning, feedback on students' answers is very effective (Hattie & Timperley, 2007).

Teaching activities should serve the end goal of a student being able to set up effective and fruitful internal dialogues. With this competency, a learner is able to learn more independently from an instructor (parent, tutor, teacher, etc.).

When looking for powerful instruments for instruction, a meta-analysis conducted by Marzano, Pickering and Pollock (2001) offers nine categories with statistically significant effects on student performance. 'Providing feedback'—a key component of formative assessment— is one of these nine. Hattie (in Marzano et al. 2001, p. 96) formulates it even stronger: "*The most powerful single modification that enhances achievement is feedback."* From a synthesis of twelve meta analyses, Hattie and Timperley (2007) report feedback as having a top five position with respect to learning gain (mean effect size of 0.79). In the model they propose to underpin feedback they give a central place to three feedback questions: "Where am I going?", "How am I going?" and "Where to next?", referring to Ramaprasad's (1983) three key-processes in learning and teaching: establishing where the learners are in their learning, establishing where they are going, and establishing what needs to be done to get them there. We summarise this as the *guiding function* of feedback.

Hattie's belief that feedback is a major accelerator of learning has had extensive uptake in educational research over the last forty years (Anderson, Kulhavy, & Andre, 1971; Butler & Winne, 1995; Hattie & Timperley, 2007; Kulhavy, 1977; Mory, 2004; Mory & Jonassen, 1996; Shute, 2008).

There are several aspects of feedback that should be thought through. We will address these aspects in chapter 2 in more detail. For now, we mention the following eight aspects:

1. There is *directive* versus *facilitative* feedback.
2. The *specificity* of the feedback is to be chosen.
3. The *complexity* of the feedback is to be taken into account.
4. The *addressing* of the feedback matters: to the student or to the task?
5. The timing of the feedback is relevant: immediately after the task or with a delay?
6. The *presentation* of the feedback: written or orally?
7. The *personalisation* of the feedback: to what extent the personal circumstances of the student are taken into account?
8. *Delivery* of feedback: by a human or computer?

Feedback, if implemented correctly, may be important for student learning. However, giving feedback is a very time-consuming activity for teachers. Educational technology has developed to a point where it may provide a solution for a substantial part of this time problem (Collis, de Boer, & Slotman, 2001). For example, a test implemented in a web-based learning environment, completed by the students, can generate basic feedback to students automatically in real time. The question then rises: *how* can we implement feedback using educational technology so students are able to profit from the teacher feedback?

In this study we explore the potential of an intervention that aims to support the process of the supply of feedback to students. This feedback is to be given on students' exercises in statistics, while utilising ICT. In the next section (1.3) we will specify the ICT setting we used in this study. With the support of the feedback process, our aim is to support the setup of an interactive classroom discourse, based on students' work. Black and Wiliam (2009, p. 8) indicate this as "*engineering effective classroom discussions*", the second step in the process widely known as *formative assessment* (Bloom, Hastings, & Madaus, 1971; Sadler, 1989). This classroom discourse will offer the teacher chances to provide feedback. But as we will see in chapter 5, it works the other way round as well: the supply of feedback as a starting point for a classroom discussion.

In research reported so far, no principle distinction with respect to the domains of study – of which mathematics is one – has been made when it comes to giving feedback. This could be caused by the fact that it is not a domain itself that asks for feedback, but the characteristics of a task that make a specific implementation of the feedback structure effective. As a working hypothesis, we assume domain-neutrality of the concept of feedback, convinced by the few research reports specifically focussing on feedback in mathematics (McIntosh, 1997). In chapter 8 we will look back in order to verify or falsify this assumption.

1.3 The feedback potential of the graphing calculator in a classroom network

So far, we discussed the value of feedback with respect to learning in general and the urge to apply it to Dutch upper secondary mathematics education specifically. In this section we will highlight the use of ICT to supply feedback, as is advocated as one of the five recommendations for how educational technology should transform American education (Atkins, et al., 2010). When we came to realise ourselves that the lack of contact time in mathematics education could possibly be compensated by an intensification and improvement of teacher feedback, we asked ourselves: how could we realise this?

Computers and feedback have been a successful combination ever since the computer was introduced in education (Kulik & Kulik, 1991; Pressey, 1926, 1950). Since the curriculum reform in 1998, Dutch students all have a rather sophisticated, very compact computer at hand: the graphing calculator (Quesada & Maxwell, 1994). So when we were able to utilise this computer for feedback, we wouldn't be confronted with 'reservations of a computer classroom' and other availability related problems. Supply of feedback is technically most easily realised when using a network. The most flexible network in a classroom situation is a wireless network, provided, of course, that everything works as expected.

These considerations made us choose a wireless classroom network (CN) of graphing calculators (Hegedus & Kaput, 2001) in order to support the feedback from teacher to students.

1.4 The supply of feedback in a classroom network

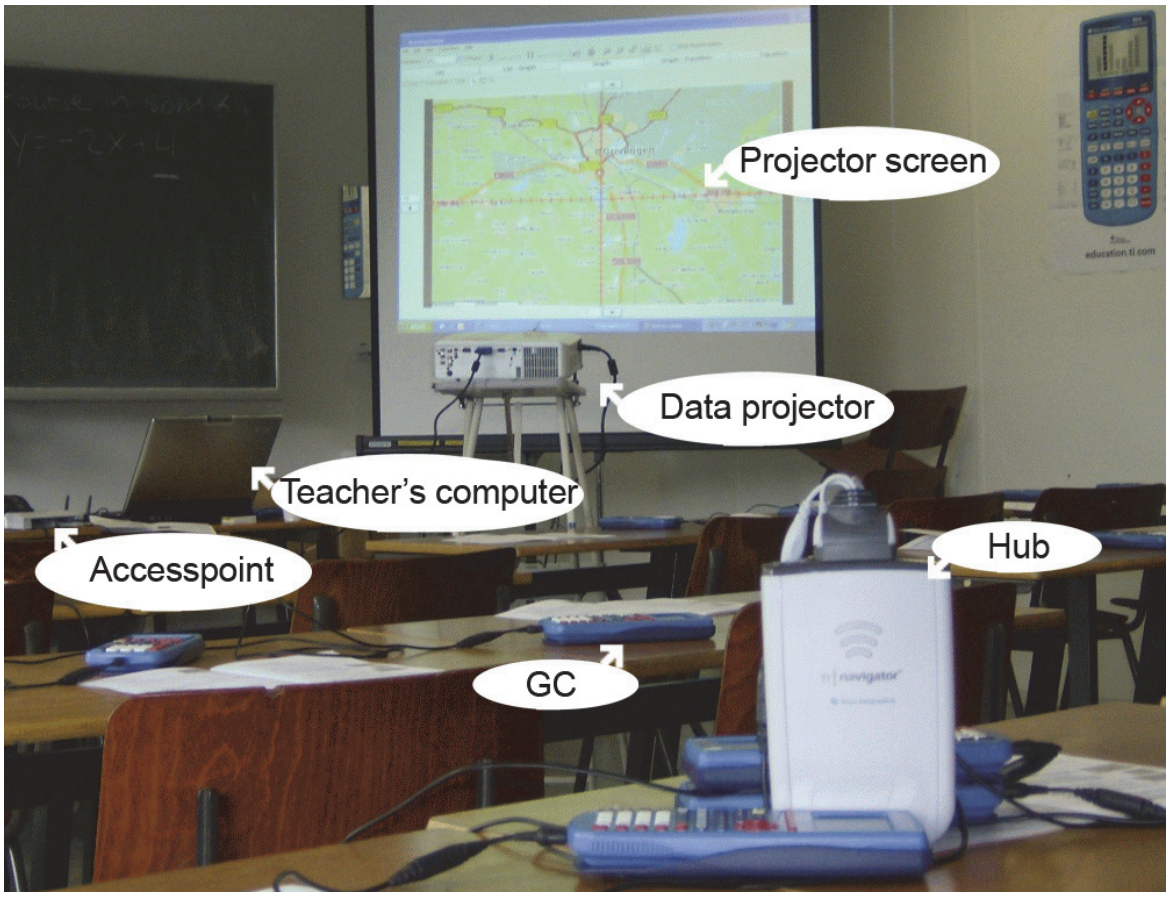


Figure 1.1 An overview of the components of a classroom network

In order to get a concrete idea of classroom activity, we first describe in terms of ICT-facilities what we see in the classroom when using a classroom network (CN). This is illustrated in figure 1.1. The functional interaction takes place between the teacher's computer and the students' graphing calculator (GC). Technically, the GCs are in groups of four, connected to a hub. This hub communicates with an access point that in turn communicates with the teacher's computer. On the teacher's computer there is software that is able to analyse the students' work and represent the analysis in a slide show. Using

the data projector and the projector screen, the teacher can make this analysis available for the students.

Bakker (2004) states that the actual trend of self-regulated learning in Dutch mathematics education, as is also used in the 'Studyhouse approach', makes it harder for a teacher to set up a classroom discussion. As stated in section 1.3, the supply of feedback could be a good starting point for a classroom discussion and vice versa (Black & Wiliam, 2009). A CN offers the possibility for this supply and can thus be a propelling force for classroom interaction. The classroom discussion can then be based on the students' work that was collected and analysed with the CN. This approach seamlessly fits with the concept of self-regulated learning. First, the students receive the exercises through the CN. Then they work individually or cooperatively on these exercises. They receive immediate feedback from their graphing calculator when they complete an exercise. Then the teacher uses the CN to collect the assignments and immediately evaluates the students' performances. This evaluation can then be exported to a presentation to be shown, using a data projector, to the students. Then the process of teacher feedback starts.

This is a basic description of the workflow we aim to establish in the intervention to be designed, developed, piloted and evaluated in this study.

1.5 The graphing calculator

A 'graphing calculator' (Quesada & Maxwell, 1994), also known as 'graphic(al) calculator' (Berger, 1998) or 'graphics calculator' (Drijvers & Doorman, 1996), is a specific handheld device dedicated to the original task of a computer, as stated by Webster's: 'calculating' or even 'mathematics'. The graphing calculator (GC) was the first handheld computer to be structurally implemented in education. In stand-alone mode, its influence on the teaching process has been studied by many researchers. We are specifically interested in what research says about: GCs, classroom network, mathematics education and feedback, if possible in coherence with each other.

Doerr and Zangor (2000) describe how students used the GC to construct mathematical meaning out of particular tasks. They conclude that five patterns and modes of graphing calculator tool use emerged in this practice: computational tool, transformational tool, data collection and analysis tool, visualising tool, and checking tool. With respect to the learning environment we developed, besides these five modes, we add a sixth one: in this study we try to utilise the graphing calculator as *a network device for direct formative assessment*. This means that the graphing calculator will have to function as a *communication tool* too. Further, Doerr and Zangor (2000) conclude amongst others that the use of the graphing calculator as a personal device can inhibit communication in a small group setting, while its use as a shared device supported mathematical learning in the whole class setting. A classroom network exactly transforms the GC from a strictly private tool into a more shared device. This transformation can be compared to the one that turns a stand-alone computer into an internet workstation that uses a web browser to participate in online communities.

Four years of observing mathematics classes using a wireless classroom network made Davis (2003) conclude that this way of working effected both the content of the lessons as well as the whole learning environment. She explicitly mentions that formative assessment got a more natural place in the teaching landscape.

Roschelle and Pea (2002), Roschelle (2003), and Roschelle, Tatar, Vahey, Kaput, and Hegedus (2003) demonstrate that there are three main problems to be solved before

wirelessly connected handhelds can have a large-scale break through. Incompatibility of soft- and hardware, although of vital practical importance, will not be fully problematised in this study. The other two are:

1. Research needs to arrive at a more precise understanding of the attributes of wireless networking that meet acclaimed pedagogical requirements and desires.
2. Further research should investigate the possibility of rich pedagogical practice arising out of simple wireless and mobile technologies.

By designing content that tries to utilise the specific technology in an optimal way (with respect to feedback) and by observing very closely the actions of the teacher and the students during feedback situations, we try to tell a story 'of rich pedagogical practice' (point 2). By our analysis we try to 'arrive at a more precise understanding of the attributes of wireless networking that meet the acclaimed pedagogical desire of feedback' (point 1).

As always when working with ICT there is some terminological confusion. Beside 'classroom network', a couple of other terms are used to indicate systems with quite similar functionality. Fies and Marshal (2006) review the research literature on 'classroom response systems'. When reviewing students opinions on 'clicker classrooms', Trees and Jackson (2007) noted that 'voting systems' can transform classroom dynamics. This transformation can actually be interpreted in the way Hegedus and Kaput (2002) describe with their remark that, after implementation of a classroom network in mathematics education, the relationship between mathematical and classroom social structure was radically strengthened, which was confirmed by Stroup et al. (2005).

Simpson and Oliver (2007) compared theory and practice around 'voting systems' and concluded, amongst others, that the field in practice is still in the preserve of the enthusiast and research has just started. This does not hold Mazur (2009) from suggesting that lecturing in higher education, without the interaction between lecturer and students, made possible by technologies as voting systems, may be disappearing.

We conclude, with Simpson and Oliver (2007), that there still is not a complete research base with respect to classroom networks. There are nevertheless promising possibilities for classroom feedback, inducing enhanced student engagement and a strengthened classroom discourse.

In the next section we will describe the learning domain in which we will deploy the intervention for this study – mathematics, or more specifically statistics – with respect to the characteristics that could possibly be relevant for the implementation of the feedback.

1.6 Statistics education

In The Netherlands statistics was introduced into the secondary school curriculum in 1968 (Hemelrijk, 1968), building upon the work of, amongst others, Bunt (1956). In the United States statistics was introduced in 1989 (Romberg, et al., 1989). We therefore consider statistics to be relatively new in the curriculum of secondary mathematics education. The fact that the introduction of statistics into the Dutch secondary school mathematics curriculum was a real innovation can, amongst others, be deduced from the minutes of a 1967 meeting from the Dutch Association of Mathematics Teachers. In these minutes it is mentioned that a considerable number of the mathematics teachers did not consider themselves skilful enough to teach statistics (Kleijne, 2008). In the U.S.A., despite

introductory problems, statistics acquired a firm position in the curriculum (Shaughnessy, 2010).

Shaughnessy (2010) reasons that statistics, from the quantitative techniques represented by the secondary school curriculum, may be the most common element of mathematics used after completion of secondary school. In tertiary education, students use statistics during multiple quantitative courses. Professionals use statistics in their working and even in their private lives. Whether Shaughnessy exaggerated a little while stating that statistics is more often used than, for instance, elementary algebra, or not, we agree with him that statistics is for many students very useful. This usefulness is perhaps the most important argument for the firm place statistics nowadays has in the curriculum. The usefulness, together with the applied character of statistics, attributed to the decision to include statistics in the curriculum for mathematics A, the specific variant of secondary mathematics for those students who aim to study in higher economics education or social sciences.

In this study we restricted ourselves to the sub-domain of *descriptive* statistics. We have no reason to believe that feedback in descriptive statistics is more important than in other sub-domains of statistics education. The main reason is that descriptive statistics is usually the introduction to statistics education, making it interesting because it is possible to give a good impression of statistics. The use of descriptive statistics is that widespread that it is hard to open up a newspaper without being exposed to it. Apart from the fact that descriptive statistics is used in a wide variety of sciences, it is important to have knowledge of its concepts and methods for those who want to be a member of today's information-based society (Gal & Garfield, 1997). But utility for and future study of secondary school students are not the only reasons. Pereira-Mendoza and Swift (1981) added aesthetics to the curricular goals of statistics, which we adopt as a third reason for choosing it as a learning domain to conduct our research. In our view, the beauty should be sought, for the target group in this study (grade 10 senior secondary education), primarily in the possibility to analyse real life phenomena, caught in authentic contexts (Wijers, Jonker, & Kemme, 2004), with the use of mathematical techniques. This is a utilitarian view on beauty through which, hopefully, the intrinsic mathematical beauty will shine.

Besides these reasons, we had a couple of organisational reasons for descriptive statistics as a domain for our study. The first one is that this topic was conveniently programmed at the schools where our intervention took place. A second pragmatic reason (for the third round of pilots in 2010) is that in 2007 the domain of descriptive statistics was no longer assessed by a central exit examination, but by a school examination. This offers schools more freedom in organising their education, thus making participation in this research easier.

1.7 An initial study

As we concluded in section 1.3, there are promising possibilities for the improvement of feedback through the use of a classroom network. This could provide a solution to the problem we sketched earlier this chapter: the shortcomings of contact time in Dutch upper secondary mathematics education.

In this section we describe the results of a first initial intervention study (Tolboom, 2005) that tries to identify what we observe when we utilise a wireless classroom network in mathematics education. We hypothesised that there were opportunities for the improvement of interaction and we were aware of some reported research mentioning an

improvement of student engagement (Hegedus & Kaput, 2002). However, we wanted to get a more precise pedagogical picture of what this engagement looked like and, if possible, what the didactical structure of this engagement was likely to be.

The topic of this initial study was separate from the rest of this research: the use of the normal distribution in probability theory. We converted a chapter of a regular Dutch mathematics textbook to digital content that could be used in a classroom network. This chapter discussed the relationship between the three basic parameters of the normal distribution (mean, standard deviation, and the limit(s) of the value of the stochastic variable) and the probability. The interventional character was thus mainly a matter of media: we exchanged the textbook for digital materials, while keeping the content the same. We observed the lessons and discussed our observations during individual interviews with three students and the teacher. Meanwhile, our research question for this study was posed as "What is the most far reaching didactical change in the classroom discourse that was caused by the introduction of a classroom network?"

When operationalising the phrase *didactical change* we had to define a reference level. Therefore we took the expectations from the teacher and observers based on their experience of mathematics education without the support of a classroom network.

After analysis of the observational data we concluded that the technology was not developed enough for large-scale use. About 30% of the effective time in the classroom was lost through technical problems (both with handhelds and the network). But more interesting, the most far reaching pedagogical change in the classroom discourse was a remarkable intense discourse on mathematical objects and processes (Sfard, 1991). We discussed this discourse after the intervention in open semi-structured individual interviews with the teacher and three students (one with weak, one with average and one with good competence with respect to mathematics, according to the teacher). It turned out that teacher feedback on the students' work seemed the pivot for the intensified classroom discourse. The teacher and students indicated that the use of the classroom network was the facilitator of teacher feedback.

The results of this initial study motivated us to set up a study into the potentials of a classroom network to support the supply of feedback in statistics education.

1.8 Main research question and research methodology

With the results of the initial study (see section 1.7) in mind, our basic interest with respect to this study can best be described by the main research question:

> "*What are the potentials of a classroom network in supporting teachers in providing feedback in statistics education?*"

In the operationalisation of the feedback we are interested in the question:

> "*Can feedback, as supported by a classroom network, be a pivot to an interactive classroom discourse?*"

We are interested in the questions: *Does* the teacher utilise the support to set up an interactive classroom discourse? And if so, *how* does she or he utilise it? With the answers to these questions, we try to identify the conditions under which feedback in statistics education can be provided as well as enhanced with the support of ICT.

We choose to use the methodology of *educational design research* to conduct this study. In chapter 4, we will justify this choice more firmly. For now, we restrict ourselves with the following arguments.

The aim of this study is to find out *if*, and if so, *how,* the supply of feedback in mathematics education can be improved by using a classroom network. Several authors (Kelly, 2003; van den Akker, Gravemeijer, McKenney, & Nieveen, 2006; van den Akker & Kuiper, 2008) argue that *educational design research* (EDR) is appropriate for leading this kind of improvement processes. Our main research question starts with 'what', implicating a *how*, that is: exploring a mode, eventually trying to find out *why* (an intervention succeeds or fails with respect to its goal). EDR is supposed to be a logical paradigm for this type of research. Kelly (2007) states that design research is most appropriate for *open wicked* (Rittel & Webber, 1973) problems. Our problem could be *wicked*, because feedback, although a classical theme in learning science, is still not completely understood (Cohen, 1985; Shute, 2008) and not very well structurally implemented in classroom practice. Our problem is *open* since it is highly unlikely that there is just one way to just one answer. What is more, the technology we explored and its use were very new, to the extent we presumed we needed several iterations in order to create a teaching setting specific enough to yield data that could possibly lead to an answer to our research question.

Since the paradigm of educational design research is eclectic in its nature (Gravemeijer, 1994), we picked out those methodological elements that seem suitable for this research. First, we conducted a literature study with respect to feedback, statistics education and ICT. After this, we formulated the design principles that guided the design of a prototype. We conducted a content analysis, with respect to the design principles, of two predominantly used mathematics textbooks. With the design principles, and the conclusions from the content analysis, we developed a prototype including corresponding so-called hypothetical learning trajectories. The design of the prototype was reviewed by field experts and adjusted to their findings. We tested the prototype in classroom practice evaluated through student questionnaires and interviews with teachers and students. Analysis of these data offered input for a revision of the prototype. Again, the prototype was reviewed by experts and again adapted. Following this process the prototype was put into practice with the same instruments: observation, questionnaires and interviews. We will conclude this study by reflecting on the entire process in order to illuminate results and shortcomings.

1.9 Overview of the following chapters

We will explicate in chapter 2 what we understand by 'feedback' and what research says about its implementation. In chapter 3 we will investigate the results of studies in statistics education. In chapter 4 we will describe the methodology used to answer our research question. Chapter 5 is dedicated to the description of prototype design and development. We evaluate the two pilots conducted with this prototype in chapter 6 and formulate the adjustments to the prototype these results suggest. The evaluation of the adjusted prototype is presented in chapter 7. In chapter 8 we will look back at our study, discuss its results and shortcomings and suggest further research.

Chapter 2
The value of feedback and the potential of ICT support

In this chapter we describe the importance of feedback and the possibilities of information and communication technology (ICT) rich learning environments to support the supply of feedback. The aims we have in this chapter are threefold:

1. *to define what we mean by 'feedback';*
2. *to investigate what research says about the usefulness and efficacy of feedback for students' learning;*
3. *to investigate whether ICT can have added value for providing feedback, and if so, how.*

First, we will explore research on feedback and the related concept of formative assessment. Then we describe which types of feedback are possible and what the characteristics of these types are in order to be effective. Furthermore, we review what ICT in general can mean for the generation of feedback. Then we conclude what the possibilities are of our chosen ICT setting (see section 1.3) for the supply of feedback. This chapter will result in recommendations for the prototype that will be described in chapter 5.

2.1 Feedback and formative assessment

2.1.1 Why feedback?

Meta analyses show that, from the perspective of students' achievement, feedback has a large positive effect (Hattie, 1999; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Marzano, et al., 2001). In his synthesis of over 800 meta analyses, Hattie (2009, p. 12), relating to achievement, states that "*...the most powerful single influence enhancing achievement is feedback*". Sadler (1989) underpins the importance of feedback by stating that the acquirement of skills needs practicing, and that practicing needs feedback. This has been supported by subsequent research on the role of feedback for improving knowledge and skill acquisition (Azevedo & Bernard, 1995; Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Moreno, 2004; Pridemore & Klein, 1995). For us, this is important, because statistics education inevitably requires skills. Lovett (2001) concluded that feedback was useful to help students improve their ability to select appropriate data analyses, which is specifically interesting for this study, because in the domain of descriptive statistics analysis of data is the main goal. There thus seems to be a broad consensus that when wanting to improve student achievement, feedback is a major variable to focus on. However, its application is not trivial. In fact feedback "*is one of the more instructionally powerful and least understood features in instructional design*" (Cohen, 1985, p. 33). Nevertheless, since 1985 a lot of research has been reported.

2.1.2 Feedback versus formative assessment

Feedback and *formative assessment*, also known as *assessment for learning* (Black, Harrison, Lee, Marshall, & Wiliam, 2003), are related in such a strong way that they are sometimes used interchangeably. Black and Wiliam (1998a, p. 9) distinguish formative assessment and feedback as follows:

> " [Formative assessment] *is to be interpreted as encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged."*

In this view, in the process of formative assessment the concept of feedback takes a very central place: it is *the* means to improve the process of teaching and learning. The central place of feedback in formative assessment is also reflected by the fact that providing feedback is the central (third of five) step in the model Black and Wiliam (2009) proposed on the process of formative assessment. Their definition as cited above has some other noteworthy aspects. The goal of formative assessment is to provide feedback in order to modify teaching and learning activities. We interpret 'modify' here as 'improve', because our goal is to improve the teaching process as much as possible. Feedback can be given, and received, by both teacher and student. If received by the student, it can be used to modify the learning. If received by the teacher, it can be used to adapt the planned instruction. Hattie (2009, p. 12) notes that: "*...it dawned on me that the most important feature* [of feedback, JT] *was the creation of situations in classrooms for the teachers to receive more feedback about their teaching.*" During the intervention described in this study this creation is exactly what we try to reach, the use of feedback in two directions: from teacher (and handhelds) to students and back from students' behaviour to the teacher.

2.1.3 What is feedback and how do we use it in this study?

Shute (2008, p. 154) also emphasises the interwovenness of feedback and formative assessment when she defines 'formative feedback':

> "*Formative feedback is defined in this review as information communicated to the learner that is intended to modify his or her thinking or behavior for the purpose of improving learning. And although the teacher may also receive student related information and use it as the basis for altering instruction, I focus on the student (or more generally, the 'learner') as the primary recipient of formative feedback herein."*

In her definition, Shute explicitly focuses on the student as a recipient of feedback, in which she differs from Black and Wiliam (1998a). The main goal of feedback is to diminish the gap between the learners' factual level of knowledge, skills and beliefs, as shown in learners' work, and the level which is aimed at by the teacher or, more generally, the curriculum. This 'closing the knowledge gap' (Ramaprasad, 1983) is an important example of 'improving learning' as posed by Shute (2008). We follow Shute (2008) and Black and Wiliam (1998a) in stating that the *goal of the feedback* is the possibility of improving the learning process and result of *the student*.

We will now take a closer look on the difference between Shute's (2008) use of 'feedback' and its use by Black and Wiliam (1998a). Who is giving and who is receiving feedback? Shute's definition of formative feedback (2008) does not point out which actor is responsible for giving feedback. Hattie and Timperley (2007) use an 'agent' for the supply of feedback. As possible agents they suggest: teacher, peer, book, parent, self, and

experience. Which of these are the most important for this study? When realizing that giving feedback is part of an interaction (Rapp & Goldrick, 2004), we can base feedback on Moore's (1989) identification of three types of instructional interactions: between the *learner* and the *instructor*, *between learners* and between *learners* and *the content* they are trying to master. While focusing on the feedback aspect of interaction, all of these interactions contain interesting aspects for our study. For the use in this study, we choose these feedback types, based on Moore (1989):

1. *Teacher* feedback: given by the teacher on the work of the students
2. *Peer* feedback: from student to student
3. *Curricular* feedback: given by (parts of) the (computerised) curriculum materials to the students

Teacher feedback is particularly common as a form of feedback that it is usually abbreviated as 'feedback'. *Peer feedback* came into research's focus in the mid 1990s. Liu and Carless (2006) stated that peer feedback is the learning element in peer assessment, similar to our conclusion that feedback is the learning element in formative assessment. With respect to peer assessment, Black et al. (2003) agree with Sadler (1998) that peers are a valuable source of feedback in the learning process. *Curricular feedback* will be analysed in section 2.3. We then contend ourselves to curricular feedback provided through ICT. The students' work and behaviour are a source of feedback for the teacher in order to modify instruction. The teacher is in this case the receiver of feedback.

After these findings from research and after having explained our incentive, that being the improvement of feedback in order to enhance students' learning, essentially we can define *feedback* in this study as:

> The teacher's, peers' and ICT-generated comments on students' work in order to improve the students' learning, while the students' work and behaviour are a source of feedback for the teacher in order to modify instruction.

2.1.4 The relationship between formative assessment and feedback

In the preceding sections, we have discussed:

1. feedback as a means for improving students' learning;
2. formative assessment as a key element for the design of the intervention.

The distinction between formative assessment and feedback is not easy. Black and Wiliam (1998a) state that feedback and formative assessment overlap strongly. We do not agree fully, because using the term 'overlap' suggests that there are aspects of feedback that do not belong to formative assessment while in our view feedback is an integral part of formative assessment.

An example of feedback during formative assessment could be described as follows. A teacher defines a task and a standard (that often stays implicit) and delivers this to the students. Each student performs this task, with or without informal peer consultation. Immediately after completion, the student gets feedback from his graphing calculator. Then the performed task is delivered to the teacher who compares this performed task with the standard. This is feedback *for the teacher* on students' progress, possibly to be used to modify instruction. The teacher determines feedback and this is delivered to the student. If the student possesses sufficient self-regulation skills (De Corte, Verschaffel, & Eynde, 2000; Nicol & Macfarlane-Dick, 2006) and if the feedback is designed effectively

(see section 2.1.5), the student will adapt his or her learning. This way, the student will perform better in the learning process, mainly because of the feedback given by the teacher on the task performed. Because it is hard, if not impossible, not to learn when doing something, there is of course an effect per se when the student performs the task. But providing feedback can enhance the learning effects, as we will see in the next section. Feedback is a central concept in formative assessment, as Black and Wiliam (1998a) stated, and it is entirely encapsulated in the process of formative assessment.

We summarise that formative assessment is the process in which teacher and student are involved in order to improve the students' learning, with feedback on students' work as the central and one of the most active ingredients. The students' work, on the other hand, can be considered as feedback for the teacher in order to modify instruction. Feedback thus works in two ways between the teacher and students.

2.1.5 Function, types and aspects of feedback

In this subsection we explain critical factors in the use of feedback in order to be effective: the timing of feedback; the specificity and complexity of feedback, the initiator of the feedback, and the way feedback is addressed, presented and personalised. All of these aspects count for feedback received by the student.

Directive versus facilitative feedback

Black and Wiliam (1998a) distinguish two main functions of feedback: *directive* and *facilitative*. Shute (2008, p. 157) describes these two types as follows: "*Directive feedback is that which tells the student what needs to be fixed or revised. Such feedback tends to be more specific compared to facilitative feedback, which provides comments and suggestions to help guide students in their own revision and conceptualization*". Shute (2008) mentions both types of feedback, referring to Knoblauch and Brannon (1981) and Moreno (2004), that, perhaps unlike some constructivist thinkers would expect, research has shown that directive feedback may actually be more helpful than facilitative feedback (Hartman, 2002), particularly for learners in the early stages of mastering a certain domain.

Is there a difference in the efficacy of feedback with respect to the difficulty of the students' activities? The meta-analysis by Bangert-Drowns et al. (1991) revealed a significant correlation between the error rates in the lessons (how difficult is it?) and the effectiveness of feedback. We interpret this result practically as: be sure to provide adequate feedback in the case of difficult exercises.

Specificity of feedback

The specificity or elaborateness of feedback is interpreted as the level of information presented in feedback messages (Goodman, Wood, & Hendrickx, 2004). Specific feedback tends to be more directive than facilitative. As we stated in the previous section, this raises a surmise that specific feedback could be more effective than feedback designed with the emphasis of guidance.

> *Example*: Given a longitudinal dataset with the number of murders per year in Groningen for the period 1970-2009, a student is asked to calculate the three most common measures of centre: mean, median and mode. He gave the right answers for the mean and median, but a wrong answer for the mode of the data set. Low specificity feedback could be: 'You did not find the right answers.' Middle specificity feedback could be: 'Two of the three answers you gave are correct.' Feedback with high

specificity could be: 'Your answers for mean and median are right, but you did not find the right value for the mode.' Very high specificity feedback could be formulated as: 'Your answers for mean and median are right, but the mode does not exist as the highest frequency occurring was 7 for the years 1978, 1983 as well as 2007.'

Research shows that 'low levels of specificity' (thus: vagueness) in the content of the feedback makes this feedback less effective (Butler, 1987; Kluger & DeNisi, 1996; McColskey & Leary, 1985; Wiliam, 2007; Williams, 1997). A feedback message that contains details on how to improve is found to be more effective than 'right / wrong' feedback (Pridemore & Klein, 1995). On the other hand, too specific feedback can discourage students to engage in further explorative activities (Goodman, et al., 2004). The efficacy of the specificity of the feedback may vary depending on the type of task. Phye and Sanders (1994) investigated learning achievements depending on two different types of feedback (general advice versus specific feedback) and conclude that more specific feedback resulted in learning gains for retention tasks, but not necessarily for transfer tasks. We see that 'specificity' can interfere with 'complexity', which we will discuss in the next subsection.

We conclude that specificity to a certain degree seems to be important in feedback but it must not degenerate into too complex formulation. Nor is more specific feedback suitable for all learning tasks.

Complexity of feedback

Complexity in the structure of feedback has wide variety (Shute, 2008). Some well-known types of feedback, ordered from simple to complex, are described in table 2.1, based on Shute (2008, p. 160). Basically, when feedback gets more complex it contains more details and the information value increases.

Table 2.1 The complexity of feedback after Shute (2008)

	Type of feedback		Description
Simple	Verification (knowledge of results)		Informs the learners whether they provided the right answers or not.
↓		Correct response, (knowledge of correct response)	Informs the learner about what the correct response to the assignment was
		Try again	Asks the student to repeat until correct
		Error flagging	Locates the error in a given answer, without presenting the right answer.
	Elaborated feedback		General term. Provides an explanation about why a specific answer was correct or not. May or may not present the correct answer.
		Hints	Guides into the right direction, without explicitly presenting the correct answer.
		Bugs/misconceptions	Continuous error flagging. Requires error analysis and diagnosis ('what is wrong and why?')
Complex		Informative tutoring feedback	Incorporates a combination of verification feedback, error flagging (where has been made a mistake?) and strategic hints on what to do next.

Overall, Bangert-Drowns et al.'s (1991) meta-analysis yielded, amongst other things, that verification feedback (correct–incorrect) resulted in lower effect sizes compared to correct response feedback (i.e., providing the correct answer), which has been confirmed by Pashler et al. (2005). Hints (Pol, Harskamp, & Suhre, 2005, 2008) may be freely chosen by the student. Hints are more facilitative than directive and are specifically suitable for supporting the learning of problem solving. As they can be consulted by the students before they have answered a question or even thought themselves, hints strictly do not belong to our definition of feedback in which the comments on the students' work are ex post. However, the concept of providing hints is so close to feedback we nevertheless mention it here.

Research results caution us to be careful in designing complex feedback. Kulhavy et al. (1985) reported that when increasingly complex feedback is presented, there comes a point where the amount of learning time required increases but learner performance does not.

Because the research on the complexity of feedback with respect to learning gains is sometimes contradictory (Kulhavy, 1977; Mory, 2004), it is not guaranteed that sophistication of feedback always pays off. Shute (2008), in her review of research on formative feedback, found the relationship between the complexity of the feedback and effectiveness that this relationship to be inconclusive. Other factors, such as nature and quality of the feedback content, may be more decisive for the efficacy of the feedback according to Shute (2008).

Thus, when designing feedback, do provide information on the 'why' and not just 'right' or 'wrong', but keep the elaboration of the feedback simple.

Addressing feedback

After a thorough meta-analysis of existing research with respect to, amongst others, the addressing of feedback, Kluger and DeNisi (1996) advise with respect to the content of the feedback to keep close to the task-learning processes when it comes to addressing the feedback ('You have divided by *n*, while you should have divided by *n*-1.') and to avoid if possible the level of meta-task processes ('You have made this mistake once again!'). This has been confirmed in recent research (Wiliam, 2007). The fact that feedback on personal level, on the *self*, can be very discouraging for low-performing students would not surprise too many of us. However, the finding that its use with high-performing students ('Right! You're so smart.') can also be detrimental (Mueller & Dweck, 1998) may be somewhat more surprising.

A feedback designer should even be careful when using comparison with other students in the feedback. The research of McColskey and Leary (1985) pointed out that normative feedback (with reference to others) results in lower learning outcomes for low-achieving students than self-referenced feedback (that is, with reference to earlier work from the same student). Providing feedback this way can also be threatening for the self of the (low-achieving) student.

We conclude that if possible, feedback should be addressed at the task-level. Feedback addressed at the student's self should be avoided. Make it functional, not personal.

Timing of feedback

Timing seems to be a crucial aspect of feedback. When designing feedback oriented learning environments with the use of modern tools, one has a very important question to answer with respect to timing: 'Do we use immediate or delayed feedback?' We do not go into detail in studying the time scale used for measuring the delay but by stating that 'immediate' in this study means 'right after completion of an exercise' (delivered automatically by the student's GC) and that 'delayed' means 'after one or a couple of days' (delivered orally, based on a software supported analysis of students' work).

Kulik and Kulik (1988) report after a meta-analysis of studies into the timing of feedback that in four of the eleven classroom studies there were significantly larger learning gains when immediate feedback was used than when delayed feedback was used. The other seven studies showed no significant differences.

The power of immediate feedback in computer-based instruction (CBI) is confirmed by the meta-analysis conducted by Azevedo and Bernard (1995) of 22 CBI studies using either delayed or immediate feedback. The mean effect size in the meta-analysis was 0.80. The mean effect size from the nine studies incorporating delayed feedback was 0.35. It appears that immediate feedback in CBI is superior to delayed feedback, which nevertheless has a respectable effect.

Corbett and Anderson (2001) have used feedback in order to investigate effects of the timing of feedback on students' problem solving skills. They came to the conclusion that immediate feedback is supportive in mastering problem solving skills. In another study (Anderson, Conrad, & Corbett, 1989), the authors studied the effect of immediate feedback during student acquisition of computer programming skills (McCarthy, 1960). They concluded that immediate feedback improves student learning.

Schroth (1992) investigated the effects of delayed feedback on transfer using a concept-formation task. This resulted in the conclusion that the delay slowed the rate of initial learning, but that after seven days the transfer of the learning was enhanced. In our view, this could have been caused by the fact that for more difficult tasks, a student needs more time to think things over than in the case of a task involving only recall of knowledge.

We have to be very careful in generalising Schroth's (1992) conclusions, stating that delayed feedback on transfer using a concept-formation task is superior to immediate feedback. Schroth measured the delay in dozens of seconds, while our delay will usually be one day. But delays up to a week can still be beneficial (Butler, Karpicke, & Roediger III, 2007). Other research provides contradictory conclusions. Clariana (1999), for instance, found some evidence, in contradiction with Schroth (1992), that feedback on more complex tasks should be delivered immediately.

Concluding, we state that there is no absolute consensus on how to time feedback. However, immediate feedback seems to be beneficial in computer-based instruction and for procedural knowledge. Delayed feedback may be useful in cases of more complex (conceptual) tasks. In our intervention, we used both immediate as well as delayed feedback.

Presentation of feedback

Orally delivered feedback by the teacher in front of the classroom is perhaps the most classical presentation of feedback in the twentieth century. However Kluger and DeNisi (1996, p. 271) state that care needs to be taken in delivering feedback orally. Orally delivered feedback (from the instructor) "*may direct attention to meta-task processes because of the salience of the FI* [feedback, JT] *provider*". A student could be embarrassed by receiving feedback from the teacher in the classroom in the presence of peers. Feedback provided in writing could augment feedback effects because it is more private and thus less likely to invoke meta-task processes. The same argument of being less likely to invoke meta-task processes counts for feedback provided graphically and feedback provided by a computer, according to Kluger and DeNisi (1996).

Thus, it seems advisable to use alternatives for orally delivered feedback, for example written feedback or feedback that uses graphics.

Personalisation of feedback

The incentive for feedback personalisation is twofold: personalised feedback could be more effective than standard feedback because it better fits the learning circumstances of the student. Further, a student feels like they are being treated as an individual instead of an element of a population.

Without the aid of automated (nowadays: computerised) tools, personalisation of feedback can just be reached in situations of (almost) private tutoring. The basic idea behind customised feedback is that the student's more or less unique interacting with the learning content could offer a personalised route through the educational objectives. A very basic example: a student studies an educational unit and takes an assessment when

he thinks he is ready for it. When his results on this assessment are good enough, he can continue with the next learning unit. When they are not, he gets presented a new test on the same learning unit. When he passes this test, he has 'mastered' the level of this learning unit. Personalisation thus exists here in the fact that each learner is allowed to the next learning stage when his personal interaction with the system is regarded to be sufficient.

Intelligent Tutoring Systems (ITS) have worked these ideas out with far more sophistication (Sleeman & Brown, 1982). The idea behind ITS is that the learning path offered to the student depends on his interaction with the system: the system gets to know the student and adapts the presented learning units to the student. Acting this way, they have a certain claim to fame with respect to 'personalised feedback' (Anderson, Corbett, Koedinger, & Pelletier, 1995; Park, 1987). In section 2.2 we will come back to ITS in more detail.

Feedback and learning gains

In general, when the intervention is based on explicit student exercises, feedback strengthens the probability of correct responses and reduces the probability of subsequent incorrect responses. Students are thus improving their learning (Phillips, Hannafin, & Tripp, 1988).

Reflecting on their review of studies examining effects of formative assessment, Black and Wiliam (1998b) concluded that improving formative assessment, in the most diverging implementations, resulted in noticeable increases in student learning with typical effect sizes between 0.4 and 0.7.

As we have demonstrated, we consider 'feedback' to be the core of formative assessment. It is therefore logical that the effect sizes found in the review of Black and Wiliam (1998b) come very close to those found by Kluger and DeNisi (1996) when investigating feedback effects in a meta-analysis. Their data (607 effect sizes; 23,663 observations) showed that, on average, feedback interventions improved students' performance (mean effect size from .41), while about 38% still had a negative effect size. One of the main features of the interventions sharing a negative effect size was a focus on the learners' '*self*', giving normative comparisons with other students instead of feedback on the completed task. The wide variation in effect sizes after an intervention introducing feedback brought Kluger and DeNisi (1998) to call feedback '*a double-edged sword*'.

Hattie and Timperley (2007) cite Hattie's (1999) synthesis of over 500 meta-analyses on various influences on student achievement. This analysis included more than 100 factors influencing educational achievement. The average effect size of schooling was 0.40, and this provided a "standard" from which to judge the various influences on achievement, such as that of feedback. The average effect size of feedback interventions was 0.79 (twice the average effect). To place this average of 0.79 into perspective, it fell in the top 5 to 10 highest influences on achievement in Hattie's (1999) synthesis.

A quite recent example of enhanced learning outcomes by the employment of formative assessment in an undergraduate life science course is given by Klecker (2007). This study is of particular interest to us because the formative assessment was deployed in an ICT setting, a web-based learning environment (WLE), conceptually quite close to the one described in this study. Klecker used this environment to deploy web-based multiple choice tests weekly to a randomly chosen part of the population of students. The immediate feedback provided by the WLE consisted of (1) the total number of items correct, and (2) the item number of items not answered correctly. Students did not receive

the correct answer as feedback immediately after the test. These answers to the test items were provided through the WLE seven hours after the deadline of the test. Compared to the control group, the experimental group scored significantly better on the final exam, while student satisfaction about the course did not differ significantly between both groups.

We conclude that, from the perspective of learning gains, feedback, when implemented correctly, can be very powerful.

Some concluding remarks on feedback

Hattie noted that "*the most powerful single modification* [in the teaching process, JT] *that enhances achievement is feedback*" (Hattie in Marzano et al. 2001, p. 96). This may be true, but there are nevertheless problems to overcome. As we saw in the meta-analysis conducted by Kluger & DeNisi (1996), in more than 38% of the cases performance was reduced after the feedback intervention due to an ineffective addressing.

Introducing feedback into an intervention is not a panacea. Hattie and Timperley (2007, p. 104) state: "*Feedback, however, is not 'the answer'; rather, it is but one powerful answer."* We are, at this stage, aware of the pitfalls and hence even more careful. We would like to state that feedback *could be* a powerful answer.

Efficacy of feedback strongly depends on, for example, the specific knowledge to be learned, the timing of the feedback, the way it is addressed and presented and possible interactions between two or more of these variables.

It is remarkable that hardly any of the studies mentioned above and of others consulted for this chapter without being cited, focus on the content of the feedback. Methodologically this may be logical, for abstracting from the content makes a lot of aspects of feedback from different studies comparable, but it is likely that the content of the feedback is one of the most important aspects of feedback. How feedback can be designed for the learning domain of statistics will be explored in this study.

2.2 The potential of ICT for giving feedback

In this section we discuss the potential of ICT for the supply of feedback and we present some examples of classical ICT applications in order to do so.

2.2.1 Some classical applications of ICT to support feedback

After the conclusion that feedback could be a valuable ingredient for student learning, we will focus on the question: can information and communication technology add value to the use of feedback? Before analysing what ICT can mean for realising several relevant aspects of feedback, as distinguished in section 2.2, we will first take a look at three classical applications of ICT and their possibilities for feedback.

Drill and practice

The term 'drill and practice' (D&P) is used for the kind of practice that repeats the material to be learned until it is mastered (Alessi & Trollip, 2001). The mastery condition is the reason why D&P is sometimes also referred to as 'drill and skill' (Dede, 2008). D&P can be considered as a computerised application of mastery learning (Block, 1972; Block & Burns, 1976; Bloom, 1968). D&P is traditionally used to reinforce previously introduced learning goals.

Streibel (1986, p. 141) identifies seven assumptions about instruction that uses drill-and-practice courseware, based on Bunderson (1981) and Salisbury (1990). With respect to feedback it is assumed in D&P courseware:

> "*The best feedback by the program from an instructional point of view is an immediate check on a student's responses according to the logic of the content: positive feedback if the answer is correct, corrective (rather than judgmental) feedback when the answer is incorrect.*"

This assumption does contain some elements consistent with other previously mentioned, research. It incorporates the power of immediate feedback in computer based instruction (Azevedo & Bernard, 1995) and addresses feedback to the task more than to the student's self (Kluger & DeNisi, 1996).

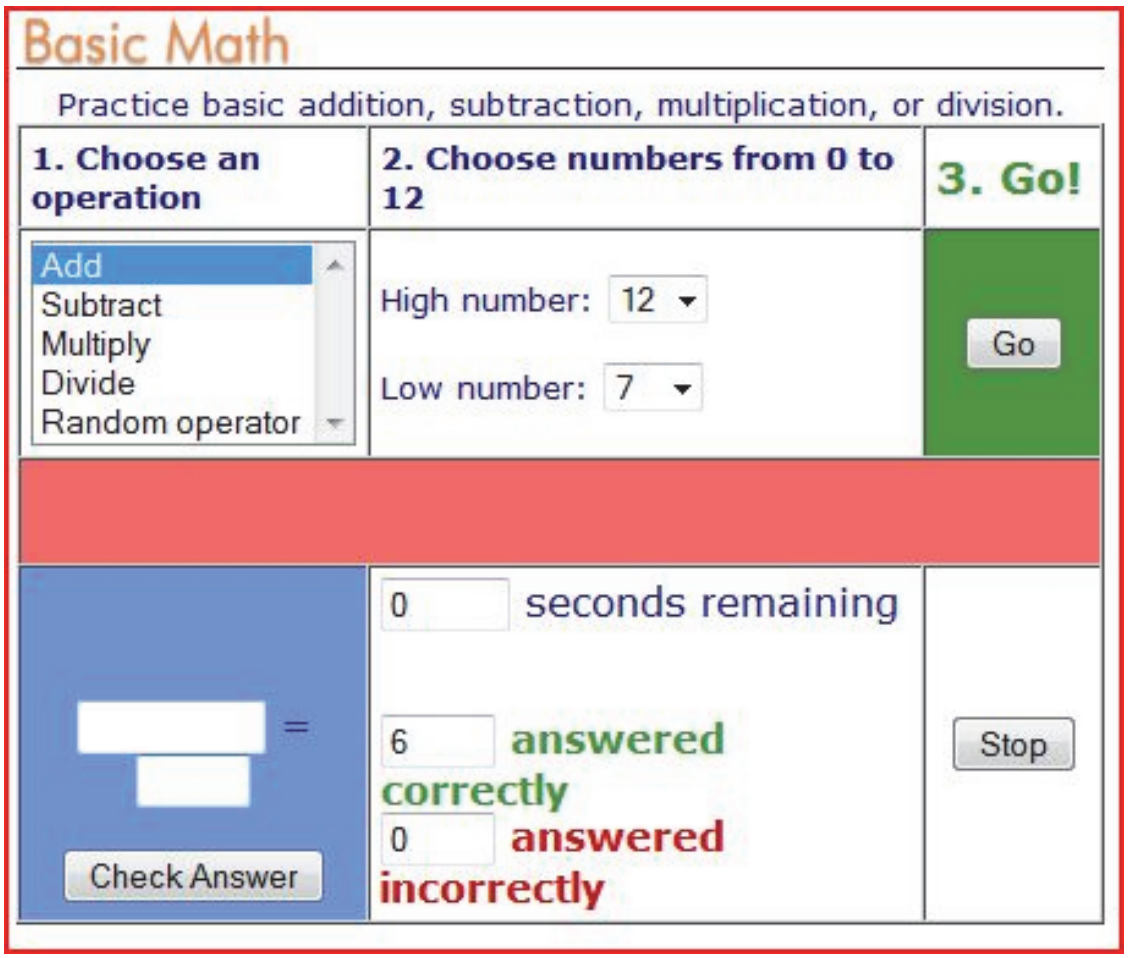


Figure 2.1 An example of a drill and practice exercise using the World Wide Web. Retrieved from http://www.math.com/students/practice/arithmeticpractice.htm, January 21st 2012.

Hativa (1988) concluded that D&P raises student achievement, especially in mathematics. But Streibel (1986), pointing at the heavily behaviouristic approach of D&P, doubted whether a real transfer of learning gains can be reached. Decoo (1994) nevertheless argues that D&P is underestimated by those who always prefer the newest developments.

In figure 2.1 we see an example of drill and practice in arithmetic using the World Wide Web. Exercises are presented for 60 seconds around one or more operations on whole positive numbers. The feedback given is rudimentary: how much time is left (organisational), how many exercises were performed correctly and how many are not (knowledge of results).

Drill and practice can be effective for the well described goals it has (Kulik & Kulik, 1991; Vinsonhaler & Bass, 1972).

Thus the drill and practice approach can be useful for reinforcement for instance because of its possibility of providing immediate feedback. Despite its critics, D&P is still a popular approach in classroom practice in general (Dunleavy, Dextert, & Heinecke, 2007).

Tutorials

When compared to drill and practice, tutorials are more complete from the courseware perspective. Tutorials have always had a firm position in education, especially in the Anglo-Saxon culture, before computers were found in the classroom. Sweeney, O'Donoghue, and Whitehead (2004) describe this process as guided interaction in a small group of students. It is the interaction aspect that makes tutorials relevant for our study.

The ICT supported tutorial is very multifaceted and still lacks a formal definition. What is the difference with D&P courseware? Tutorials usually aim to present new content, while D&P is mainly used to reinforce existing knowledge. Bork (1980) mentions on-line tests, remedial dialogues, and interactive proofs as exemplary tutorial content. All of these have obvious feedback potential. From the view of learning achievement, tutorials can be effective as, amongst others, Kromhout (1972) and Grant and Courtoreille (2007) showed.

> *Example*: As a current example, we take a look at the web based Introductory Biological Psychology Tutorials from Athabasca University (Nagel & Grant, 2007). The tutorial includes 20 ‘sub tutorials’ on biological psychology. Each sub tutorial is built around an introductory concept, for instance ‘Structure of the Neuron’ and is divided into three sections:
>
> 1. An initial part 1 in which students work with an image-mapped graphic to identify visually based information such as brain structures. In this section students also read descriptions associated with the graphical information to develop a verbal knowledge of the material.
>
> 2. Part 2 in which students do a self-test matching visually based information (e.g., brain structures) to labels.3. Part 3 consists of multiple-choice self-test items that test student knowledge of the material presented in the tutorial.

Parts 2 and 3 from each sub tutorial can be considered as formative assessment. As feedback, they provide knowledge of result (part 2) and knowledge of correct response (part 3) feedback. In figure 2.2 we see the generation of immediate feedback, after the student clicked on alternative d as, in his view, being the answer to the fill-in exercise.

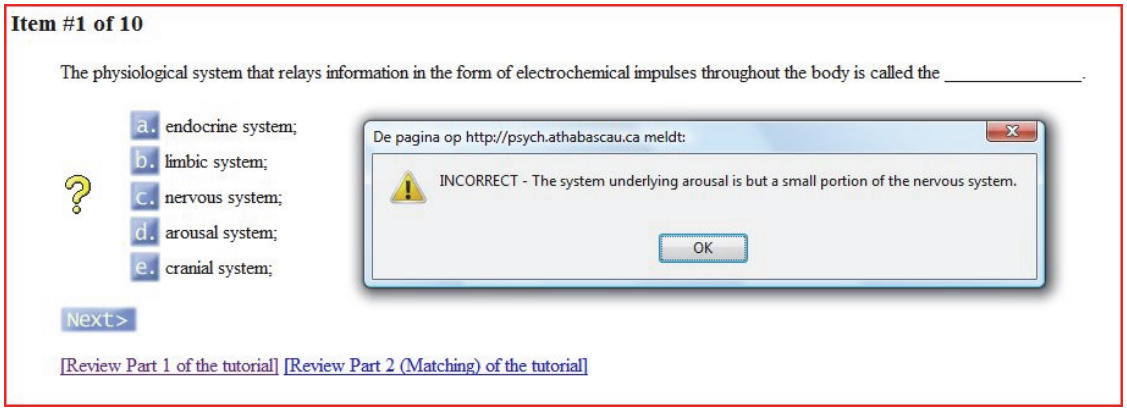


Figure 2.2 Introductory Biological Psychology Tutorials from Athabasca University (Nagel & Grant, 2007)

We see that this type of feedback is still rather simple. In terms of complexity as described in section 2.1, it contains correct or incorrect (knowledge of results) feedback and the relationship between the given answer and the correct answer (knowledge of correct response).

Bliwise (2005) introduced interactive web-based tutorials as a supplement to lectures in an introductory statistics class. Two groups were distinguished: classes with tutorials and lecture-only classes. Analysis showed that students who attended the classes with tutorials scored higher than students who attended the lecture-only classes.

We conclude that ICT-supported tutorials, despite their shortcomings when compared to human interaction, have interesting possibilities for the supply of feedback.

Intelligent tutoring systems (ITS)

Burns and Capps (1988) stated that computer-assisted instruction can be labelled as an Intelligent Tutoring System (ITS) by passing three tests of intelligence.

First, the subject matter or "domain" must be known to the computer system well enough to draw inferences or solve problems in the domain (*expert* knowledge).

Second, the system must be able to deduce a student's approximation of that knowledge (*student diagnostic* knowledge). Third, the tutorial strategy or pedagogy must be intelligent in that 'the instructor in the box' can implement strategies to reduce the difference between expert and student performance (*instructional* or *curricular* knowledge). Combining these three intelligences simultaneously gives an ITS the possibility to provide intelligent feedback.

Very roughly, we could state that 'tutorials' are the next generation 'drill and practice' and 'ITS' are the next generation 'tutorials'. Pedagogical ambitions, though, have been formulated more realistically, as Anderson et al. (1995, p. 168) stated: "*We have totally abandoned our original conception of tutoring* [in our view, Anderson refers with tutoring to Intelligent Tutoring Systems, JT] *as human emulation*." ITSs have evolved to the extent that there are even some commercial products based on ITS. Examples of ITS in mathematics education are ALEKS (http://www.aleks.com), Cognitive Tutor (http://www.carnegielearning.com) or Smart-s (http://www.smart-s.com/en/) (retrieved January 21st 2012).

> *Example*: in an ITS the student is asked to draw a frequency polygon given a set of data. There are two activities the system marks on: has the student chosen the right classification (based on two rather directive hints for this) and has the student

classified the data correctly? Figure 2.3 shows the feedback received if this last activity is not completed properly.

Rules about mastery are incorporated in the system and about the mastery hierarchy: mastery at underpinning sub domains is required before a student is allowed to enter a more sophisticated level. We see that the feedback is rudimentary and that the exercise is at a technical skill level.

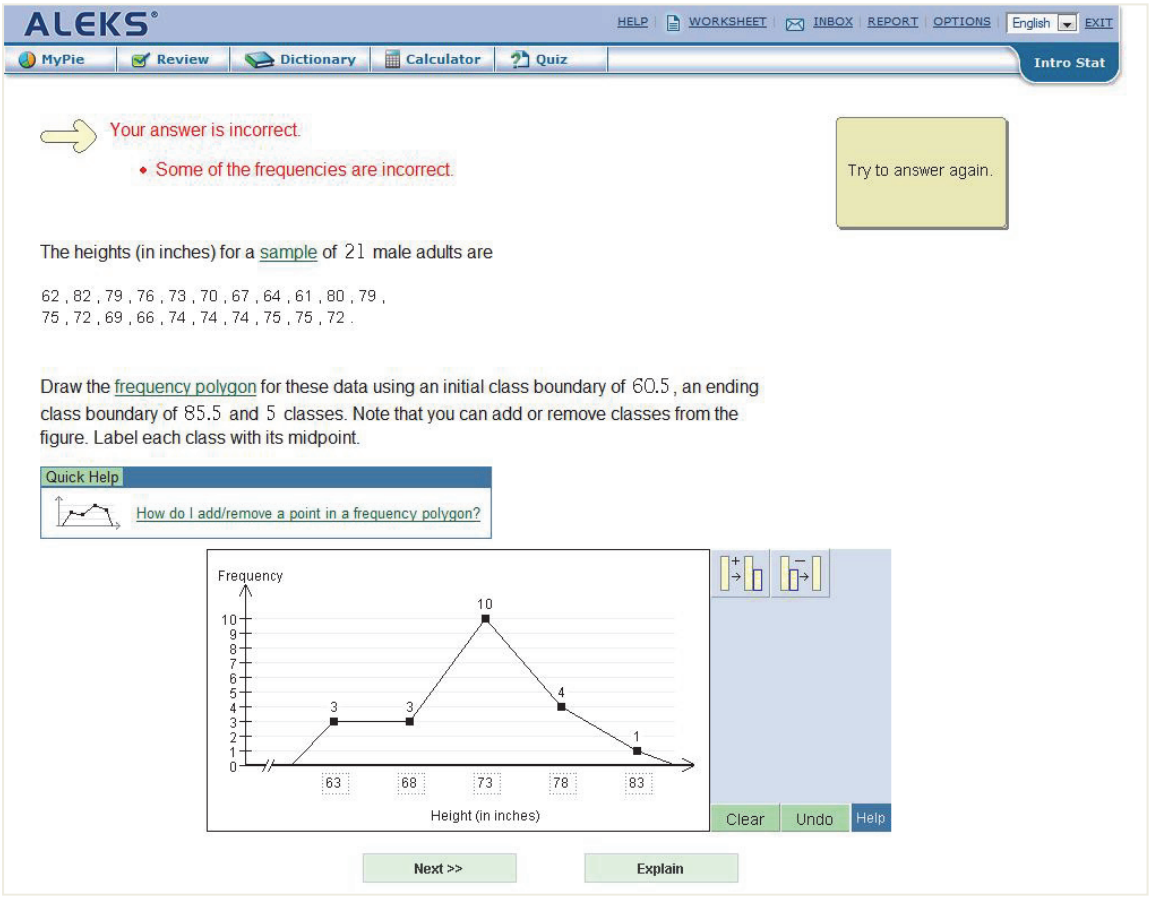


Figure 2.3 Feedback generated by an intelligent tutoring system

The gains achieved by deployment of ITS in instruction have initially been reached in highly structured, well-defined, domains such as geometry, Newtonian mechanics and system maintenance (Lynch, Ashley, Aleven, & Pinkwart, 2006; Sykes, 2007).

Experiences with this type of ICT support are usually hopeful. Schofield et al. (1990) reported large improvements in the motivation of students with students spending more time on task. Anderson et al. (1995) also report remarkable motivational gains.

What about learning gains? Anderson et al. (1995) report classroom results with a geometry tutoring system (Koedinger & Anderson, 1990). It has been subject to a preliminary evaluation (Koedinger & Anderson, 1993) in which a large positive learning gain was found but only for the teacher activities carefully integrated into the project. The fact that the tutor had its benefit only in combination with the teacher highlights the issue of integrating the tutor into the classroom. This makes it very unlikely that an ITS will replace human teachers in the short term.

Recent research, amongst others, tries to add emotional intelligence to ITS (Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008) in order to tailor instruction and feedback

even more narrowly to the student. It tries to enhance adaptive abilities of intelligent tutoring systems especially regarding problem solving modes offered to a learner and ordering of hints from the most general to the most specific (Anohina, 2007).

In short, ITS can achieve large motivational and learning gains. We will explore the implementation of ITS in education in section 2.3.3.

2.2.2 ICT or human feedback?

Bloom (1984) reported that individual human tutors can have an effect size of 2 above normal classroom instruction. Achievement gain research so far (Anderson, et al., 1995; Koedinger & Anderson, 1993, 1997) reported effect sizes for computerised tutors of about 1. Human tutors are thus very good compared with computers. But what about learner preferences when it comes to the source of the feedback they are provided?

There seems to be broad evidence that computer mediation is preferred when it comes to the source of feedback. Karabenick and Knapp (1988) conducted the following experiment. Subjects were given help after having performed a task and after having received computerised 'right-wrong' feedback they were offered further help. They could choose between help from a computer or from a person contacted through the computer network. The vast majority chose help from a computer. This is consistent with the studies of Kluger and Adler (1993) who concluded that a computer is more likely to be consulted as a source of feedback than a person. ICT-mediated feedback could be preferred relative to human feedback because of a smaller anticipated 'loss of face' (Ashford & Cummings, 1983). This means that if the feedback supplying computer is not perceived as evaluative (Harackiewicz, Manderlink, & Sansone, 1984) students would not feel computer-mediated feedback to be as confrontational as human-mediated feedback. This, again, is consistent with Kluger and DeNisi's (1996) suggestion to keep away from the students' self when providing feedback. This is illustrated, for example, by Yarnall and colleagues (2006, p. 151), examining the effect of formative quizzes in science education:

> "*A fifth-grade teacher reported in an interview that 'The nature of automatic feedback encourages students to do their work. They are not in denial when they get a low score or think that the teacher must not like them. They accept since the score is shown on the Palm' [the chosen handheld device in this intervention]*."

Does this preference for a computer as a feedback medium over a human evaluator also induce learning gains? Not in general, conclude Kluger and Adler (1993), but it does for students low in self-esteem and students high in public and private self-consciousness.

We conclude that computer-generated feedback is a non-threatening and effective way of providing feedback. We have to mention that, with the emergence of the world wide web, the distinction between computer feedback and human feedback has faded, the internet, as used for instance in case of *instant messaging*, being a computerised medium for human to human feedback. In our intervention we utilised both computer-generated feedback as well as human feedback.

2.2.3 Feedback aspects and ICT

In section 2.2.2 we concluded that the computer seems to be appreciated as a feedback supplier by students. In this section we describe which aspects of feedback, as distinguished in section 2.2, the use of ICT has advantages with.

First, we mention that *general applications of ICT* (e-mail, instant messaging, weblogs) can be suitable for providing feedback (Foley, 2002), especially when realising feedback is part of an interaction (Bloch, 2002). Generic features of ICT, such as speed and scalability, can be utilised. These applications are in fact based on human feedback, delivered in a computerised way.

On the other hand, there are *feedback specific ICT applications*, such as intelligent tutoring systems or web-based formative tests, having the generation of feedback to some extent automated without direct human interference. We could state that general ICT applications can be used for feedback, while feedback specific ICT provides the feedback more or less itself. This distinction could be useful because learners seem to prefer a non-human source for feedback on their work.

We recall from section 2.1.5 that the *specificity of feedback* is interpreted as the level of information presented in feedback messages (Goodman, Wood, & Hendrickx, 2004). Goodman and Wood (2004) found that increasing the specificity of feedback during practice, using a computer for feedback delivery, increasingly guided performers to correct responses, thus profiting from the increased level of specificity. They conclude with respect to ICT that the use of ICT has obvious advantages for providing specific feedback on a frequent basis. In our view, this has to do with the general advantages of ICT (speed and scalability). Davis et al. (2005) reported that ICT-supported feedback can distinguish the level of specificity and that high specificity was especially beneficial for students low in learning goal orientation.

What are the implications of *complexity* for the type of ICT to be used for the deliverance of feedback? In Table 2.2 we check ICT support possibilities for the feedback types we presented in Table 2.1 with respect to *complexity*.

Table 2.2 ICT support for feedback distinguished with respect to complexity

	Type of feedback	Possible ICT support
Simple	Verification (knowledge of results)	Immediateness of response (IOR)
↓	Correct response (knowledge of correct response)	IOR, Immediateness of solutions (IOS)
	Try again	IOR, IOS
	Error flagging	IOR, Visualisation
	Elaborated feedback	
	Hints	IOR, Hypertext structure
	Bugs/misconceptions	IOR, Data analysis to trace the errors
Complex	Informative tutoring feedback	IOR, Intelligent tutoring systems

It is nowadays more easy to present complex feedback in an organised way. When using a hypertext structure, like, for example, on the WWW, it is possible to provide hints in a couple of phases. The first phase is very basic feedback, with a week hint towards to correct answer. Each following phase contains stronger hints. The last phase contains the correct answer with an explanation. Nevertheless, the advice to be cautious with feedback complexity still remains. The fact that complex feedback can be easily realised with the support of ICT may never be the first reason to actually design and use complex feedback. We conclude that educational research advises not to make feedback too complex but that ICT may support every degree of desired complexity.

For the *addressing* of feedback (to the task level, to the meta-task level, to the personal level; see section 2.1.5), ICT offers no specific advantages. Its general advantages, of velocity or scalability, of course, remain.

For the *timing* of feedback, ICT offers obvious advantages. In current terminology ‘immediate’ is often called ‘synchronous’; ‘delayed’ has found its analogue in ‘asynchronous’. ICT supported immediate feedback started with multiple-choice questions because it was not very hard to program a specific feedback possibility on each of the answer alternatives.

The immediateness of personalised feedback for a whole classroom cannot be realised without the use of ICT. So when wanting to profit from the benefits of immediate feedback for each individual student (Azevedo & Bernard, 1995; Kulik & Kulik, 1988), it seems inevitable to use ICT for the feedback delivery. Numerous studies have reported learning gains in working this way (Hannafin, Philips, Rieber, & Garhart, 1987; Phillips, et al., 1988).

Delayed (asynchronous) feedback can also be delivered with the support of ICT, for example by using ‘time stamps’ when programming the feedback delivery in time (answers on the items of diagnostic test *x* are to appear at *y*-day at *z*-o’clock). But here ICT has less added value than in the case of immediate (synchronous) feedback, because

a human tutor is able to deliver delayed feedback too, in case of a delay in terms of hours, depending, of course, on the number of students.

We conclude that computers are the most suitable tools on the realisation of immediate feedback in various forms.

For the *presentation* of feedback, ICT offers specific possibilities. Think of simulations, animations or serious games. There is nevertheless no research reported in which ICT plays a leading role for the presentation of feedback. An exception should possibly be made for Rieber (1996), who conducted research on the representation of feedback using a simulation of elementary mechanical laws acting on a ball. Results showed that subjects learned more tacit knowledge (Polanyi, 1966) when provided with animated graphical feedback than with textual feedback although gains in explicit understanding of these science principles did not depend on the way the feedback was represented. Even in the field of serious games there has been no serious research conducted on the possibilities of feedback, although these possibilities may seem promising (Wong, et al., 2007). It appears that researchers are not that interested in the possibilities of ICT for the presentation of feedback.

For the *personalisation* of feedback, ICT offers possibilities. The ICT-based learning approach that specifically focuses on personalised feedback is called Intelligent Tutoring Systems (ITS) (Sleeman & Brown, 1982). It has been commented on in section 2.2. Despite progress in this area, human teachers have proven so far to be unbeatable in personalising their feedback one-to-one to individual students (Sarrafzadeh, et al., 2008; Streibel, 1986). Anderson et al. (1995, p. 168), as pioneer developers of ITS, gave up competing with human tutors. Lots of subtle attributes of human interaction, like intonation of the voice and body language, are very difficult if not impossible by means of the ICT of today to interpret, let alone to use actively when the feedback situation demands to.

Unfortunately, personalisation of feedback by human means is an intensive and thus expensive process. Personalisation can best be reached in individual interaction, and human tutors are hard to split up in order to be used by more than one student at the same time. This means that for personalisation of feedback, a human tutor needs a lot of time. Computers, on the other hand, are very good at sharing their attention to more than one user.

A relative new trend in educational technology is trying to build in not only ‘intelligence’ but also ‘empathy’. Some researchers think that ITSs would be significantly enhanced if computers could adapt their feedback to the emotions of the student (Alexander, Sarrafzadeh, Masoodian, Jones, & Rogers, 2004; Kort, Reilly, & Picard, 2001; Picard, 1997; Sarrafzadeh, et al., 2008). Systems thus striving are called ‘affective tutoring systems’ (ATS). Affective tutoring systems could be an interesting new step in personalisation of feedback, but there are very few systems developed so far and there is hardly any experience with ATS in real classroom learning reported yet (Mao & Li, 2010).

We can conclude that ICT is not as good in personalising feedback as human teachers. However, ICT can do it very fast and more or less independent from the number of students. Therefore it has a great potential. In our intervention we chose to use both ICT feedback as well as human feedback.

Summarising, ICT has been reported to add value to the timing and personalisation of feedback. It could have a yet unexplored potential for the presentation of feedback. It

could also offer its general features of speed and scalability for realising specificity, complexity, and addressing of feedback.

2.3 The feedback potential of graphing calculators in a classroom network

After having investigated what research tells us about feedback and the general possibilities of ICT for the supply of it, we change two things in the way we inspect reported research:

1. *Perspective*: we now approach research with a specific hardware/software combination focus, instead of specific feedback aspects.
2. *From global to local*: we will focus on one specific situation: the use of a wireless classroom network of graphing calculators for providing feedback.

This approach is needed because we want to zoom in on the specific ICT setting as described in this study, for the sake of providing feedback.

Abrahamson, one of the first researchers in the field of classroom networks, stated that the networked classroom can support real-time formative assessment for teachers (Abrahamson, 1999). We recall from section 2.1 that formative assessment offers a central place for feedback. Roschelle (2003, p. 263), in an overview of research, noted that: "*these systems* (referring to what he earlier calls 'classroom response systems') *enable students to receive much more feedback than normal*", thus considering it an enhancement of feedback possibilities. He continues (2003, p. 268) with "*Students can perceive receiving much more individualised assessment feedback*, …". These remarks, made after having observed early practices around classroom networks, suggest that formative assessment, feedback and classroom networks are a quite logical combination. These authors apparently see advantages in the use of a classroom network in order to establish formative assessment for the sake of feedback.

In one of the first studies on classroom networks, Dufresne et al. (1996) mention that it is not only the students who receive feedback on their work. The teacher using the classroom network to collect and analyse students' work also gets immediate feedback on every student. Supported like this, he or she is able to adapt the teaching process to the apparent needs of the students. This aspect was included in Black and Wiliam's (1998a) definition of formative assessment.

With respect to classroom response systems (CRS) Fies and Marshall (2006) mention a lack of research on, amongst others, the use of CRS for purely formative assessment modalities that scaffold learning.

Penuel, Boscardin, Masyn, and Crawford (2007) conducted a survey amongst 498 elementary and secondary teachers using student response systems in their education. They hypothesised that student learning would improve in the first place because of improved feedback to students, although this was not based on empirical classroom data.

Summarising, there does not seem to be much research reported with empirical findings about the use of classroom networks for the sake of feedback. We can nevertheless conclude from what has been reported that a combination of a classroom network with handheld computers *could be* a powerful setting for realising direct formative assessment in order to provide feedback to the students. There seems to be a lack of research on the use of classroom networks for purely formative assessment in scaffolding learning. Our study tries to meet this deficiency, albeit in just one specific domain.

2.4 Conclusions with respect to our study

In section 2.1 we have described three aims for this chapter. We will now summarise our findings with respect to these aims.

1. *What do we understand in this study by 'feedback'?*

   We consider 'feedback' to be the teacher's, content's and peers' comments on students' work in order to improve the students' learning, while the students' work and behaviour are a source of feedback for the teacher in order to modify instruction.

2. *What does research say about the usefulness and efficacy of feedback for students' learning?*

   Feedback can induce considerable learning gains when implemented correctly. How it should be designed in order to optimise students' learning is difficult, because this depends on a range of variables such as type of knowledge to acquire, learner characteristics, or stage of the learning process, and possibly on interactions between these. Some rules of thumb are nevertheless formulated, such as 'address feedback to the task, not to the learners' self' and 'in the case of procedural tasks, use a computer to deliver the feedback; in the case of conceptual tasks, let the teacher give the feedback'. In our intervention, we will therefore use both computer feedback as well as human feedback. There seems to be no absolute consensus on how to time feedback. However, immediate feedback seems to be beneficial in computer based instruction and for recall and or procedural knowledge. Delayed feedback may be useful in the case of more complex (conceptual) tasks.

3. *Can ICT have an added value for providing feedback and if so how?* Computer-delivered feedback is usually preferred above human delivery by the students and is effective too when considered from the perspective of learning gains. Besides that, ICT offers advantages of scale: personalised feedback can be provided immediately to ten students as easily as to hundred or thousand. A classroom network with handheld computers is pre-eminently suitable for providing feedback; through the handhelds immediate feedback can be delivered on procedural tasks. The teacher can supply delayed feedback on the conceptual tasks.

With these findings with respect to feedback, in chapter 5 design guidelines and principles will be formulated.

Chapter 3 **Statistics education**

In Chapter 1 we justified our choice of descriptive statistics as the educational domain for this study. In this chapter we investigate what research says about the goals and the design of this domain. Then we identify some focal points for this study and use these points to analyse two textbooks on descriptive statistics. We then describe some aspects of realistic mathematics education (RME) that are consistent with some earlier identified aspects of statistics education. After this, we describe a content analysis of the two Dutch textbooks predominantly used in secondary mathematics education with respect to the recommendations as formulated before. In the end we summarise the findings of this chapter that will serve as support for the formulation of design principles in chapter 5.

3.1 Statistics

Statistics, although heavily depending on mathematical insights and techniques, is basically an independent empirical mathematical science. This means that reality is the starting point of most statistical activities. These activities may often have a mathematical character, but in statistics reality comes first (Chance & Rossman, 2005).

It is enlightening to take a short look at the differences between the scientific domains of 'mathematics' and 'statistics'. Philosophising about the relationship between 'statistics' and 'mathematics' has a long history. Wagemann (1935, p. 20) states: "*Mathematics is the science of pure number, statistics that of empirical number.*" (In German: "Mathematik ist die Wissenschaft der reinen Zahl, Statistik die der empirischen Zahl."). A mathematician would notice that this statement must have been made by a statistician, while he/she considers mathematics to have a broader scope than just numbers. However, with respect to the treatment of numbers, Wagemann has a point. A statistician lives in the real world, a mathematician in the ideal world.

Research on statistics is considered to be a little isolated from other research in mathematics. Cobb and Moore (1997, p. 801) state that: "*Statistics requires a different kind of thinking, because data are not just numbers, they are numbers with a context."* Context is in their opinion key for the analysis of data. Mathematics is usually more abstract. Cobb and Moore (1997, p. 802) formulate it this way: "*In mathematics, context obscures structure.* […] *In data analysis, context provides meaning."* We may not fully agree with the obscuring character of contexts in mathematics, but on the fact that contexts are indispensable in statistics, we could not agree more.

Statistics can thus be seen as a science that draws heavily on mathematics, but is essentially an empirical science. Reality and thus contexts therefore play a crucial role. This is especially the case in the least mathematical part of statistics: descriptive statistics, where mathematics is least manifest.

3.2 Statistics education

3.2.1 Goals of statistics education

Roughly, we could state that, when considering learning activities in statistics education, there are learning activities with a focus on 'data and concepts'. The other learning activities concentrate on 'recipes'. In our view, 'data and concepts' are the most important goals of statistics education. They offer the possibility to reason about statistics

itself (concepts) and its application (data) in various situations (concepts are transferable and sustainable). On the other hand, there are 'recipes', algorithmic in character, needed to apply statistics in practice. Cobb (1991) advises a focus on 'data and concepts' at the expense of 'fewer theory and recipes'. We, nevertheless, consider working with these 'recipes' (thus: algorithmic activities) to be indispensable for students in order to acquire a sense of where and why 'data and concepts' are needed.

In our view, Cobb's advice particularly counts for introductory courses in statistics, where intuition should be stressed in order to make the students more receptive to more formal statistical techniques, a possible domain for their subsequent study. Cobb indicates the basic choice when it comes to determining the goals of statistics education.

Thus, the learning activities in statistics education should reflect these two sides of statistical activities:

1. reasoning and sense making (Martin, et al., 2009) with and about data, (Cobb's 'data and concepts'), later to be called *data literacy* (DL);
2. *algorithmic statistical skills* (ASS), (Cobb's recipes).

When comparing both types of knowledge to the characterisation by van Streun (2001) who distinguishes:

1. *declarative* (factual) knowledge: to know *what*;
2. *procedural* knowledge: to know *how*;
3. *conceptual* knowledge: to know *why*;
4. *metacognitive* knowledge: knowing about knowing.

We project ASS on *procedural* knowledge (with a little declarative knowledge) and DL on *conceptual* knowledge.

Preferably, theory itself should not be a goal or a starting point, but has to be incorporated into the learning activities. The textbook or the teacher could distil the theory after these activities have been completed by the students. Both categories are to a high extent disjunctive, as will be shown in section 3.4, but nevertheless interdependent: to underpin and motivate the acquisition of skills, the learner needs data literacy. But the development of data literacy is stimulated by doing exercises concerning algorithmic statistical skills.

The broadness of the scope of statistical activities co-determines which specific elements belong to each category. When performing the full statistical investigation cycle (Wild & Pfannkuch, 1999), from *Problem* to *Plan* to *Data* to *Analysis* to *Conclusion* and, if necessary, then back to *Problem*, other elements of data literacy are needed than when working through the exercises of a classical textbook for descriptive statistics.

The category 'Skills' is a rather obvious and traditionally well stressed goal when it comes to statistics education. Gal and Garfield (1997) summarised their own research and that of others on the goals of statistics education. They listed eight goals for statistical education. In table 3.1 we show what these eight are with a short description.

Table 3.1 Goals for statistics education by Gal and Garfield (1997)

Description	Explanation
Understand the purpose and logic of statistical investigations	Understand why statistical investigations are conducted and the 'big ideas' that underlie approaches to data-based inquiries.
Understand the process of statistical investigations	Understand the nature of and processes involved in a statistical investigation and considerations affecting the design of a plan for data collection.
Master procedural skills	Organise data, compute needed indices or construct and display useful tables, graphs, plots and charts (with or without the use of ICT).
Understand mathematical relationships	Intuitive and/or formal understanding of the main mathematical ideas that underlie statistical displays, procedures or concepts.
Understand probability and chance	Students need only an informal grasp of probability in order to follow the reasoning of statistical inference.
Develop interpretive skills and statistical literacy	Become able to interpret results and be aware of possible biases or limitations on the generalisations that can be drawn from data.
Develop [the] ability to communicate statistically	Strong communication skills are needed if students are to effectively discuss statistical investigations and probabilistic phenomena and processes.
Develop useful statistical dispositions	Develop an appreciation of the role of chance and randomness in the world and for statistical methods and planned experiments for decision making in the face of uncertainty.

Gal and Garfield thus explicitly mention as a goal of statistics education: 'to master procedural skills'. Which specific skills are to be mastered depends on the exact domain of statistics. In section 3.4 we will give concrete examples of Skills.

With respect to 'data literacy', we concentrate in this study on four of the eight Gal & Garfield goals:

1. 'Understand the purpose and logic of statistical investigations';
2. 'Develop interpretive skills and statistical literacy'
3. 'Develop [the] ability to communicate statistically'.
4. 'Develop useful statistical dispositions'.

The first two are typical for data literacy, the third deserves special attention because our intervention aims to improve the classroom discourse in statistics classes through teacher feedback on students' work. This is essentially 'communicating statistically'. These

dispositions are to be a somewhat implicit part of the intervention, to be made more explicit in the classroom discourse aimed to be induced by teacher feedback.

The Gal & Garfield goal 'Understand the purpose and logic of statistical investigations' is in our intervention a 'sub goal': it is covered, but not in the centre of the educational design. For our target group, this goal could, in the phase of its development, become too philosophical. It is, of course, purely a data literacy goal. The second Gal & Garfield goal 'Understand the process of statistical investigations' we try to develop during our intervention, but we choose not to let the students formulate their own research question, nor gather their own data. We stress the importance of these 'front end' activities, but consider them to be too time-consuming within the practical constraints of this project. This Gal & Garfield goal nevertheless belongs to data literacy.

Then there is the Gal & Garfield goal 'Understand mathematical relationships'. We try to serve this goal by integrating students' activities underpinning these relationships in the educational design.

The Gal & Garfield goal 'Understand probability and chance' is not really applicable in our view when it comes to the basics of descriptive statistics. In the transition to inferential statistics (Loether & McTavish, 1988), this goal gains importance.

We consider ‘data literacy’ to be somewhat less technical than the well-known concept of ‘statistical literacy’. There is some variation in interpretation of ‘statistical literacy’ amongst various authors (Ben-Zvi & Garfield, 2004; Rumsey, 2002; Schield, 2002; Snell, 1999). Gal (2002, p. 2) already noted that the most common interpretation of ‘statistical literacy’, as “*a minimal (perhaps formal) knowledge of basic statistical concepts and procedures*” has expanded with “*desired beliefs, habits of mind, or attitudes, as well as general awareness and a critical perspective*”. The Gal decomposition of statistical literacy (2002) also includes ‘statistical and mathematical knowledge’. As we would like to classify the algorithms representing statistical and mathematical knowledge into the category ‘algorithmic statistical skills’, we use the term ‘data literacy’ instead of ‘statistical literacy’. Garfield and Gal (2007, p. 3) noted that: “*students need to learn what is involved in interpreting results from a statistical investigation and to pose critical and reflective questions about arguments that refer to summary statistics or to data reported.”* Garfield and Gal thus stress, ten years after their overview of goals, the importance of reflection. We agree, and consider ‘reflection’ and ‘interpretation’ to be important elements of ‘data literacy’.

> *Example*: The statistical operations of calculating the mean, median and mode, range, inter quartile distance and standard deviation of a certain data set typically belong to the domain of ‘Skills’. But analysing the context that generated the data set, in order to determine which measure(s) of central tendency and measures of variation is/are suitable to solve the information problem, typically belongs to ‘data literacy’.

In our opinion, ‘statistical’ as an adjective is more frequently associated with algorithmic techniques than with ‘data’. We consider this to be inconsistent with the core of what Shaughnessy, Garfield and Greer (1996, p. 206) call ‘data handling’: ‘*mathematical detective work within a context and neither the context nor the principal players should be disassociated from the data*’. When stressing the importance of ‘data literacy’ we lay more emphasis on the *data*, as advised by Cobb (1991), and on the concepts ruling their organisation (Bradstreet, 1996).

Even when the student activities gain complexity and size, for instance in a more or less realistic statistical investigation process (Wild & Pfannkuch, 1999), students' activities can be categorised according to the presented dichotomy.

An emphasis on data and concepts is beneficial for the learning results, as reported by Smith (1998). Moore also stresses the importance of teaching the statistical concepts (1997, p. 130):

> "[…] *on the content side we have always wanted to emphasize conceptual understanding; good pedagogy urges us to choose conceptual explanations over proofs that are not convincing to most of our students."*

We fully agree with this statement, having the majority of the students in secondary education in mind who may lack the perseverance to fully understand and appreciate formal proofs. By putting emphasis on statistical concepts, even these students will be able to build up statistical intuition. This has also been stressed by Watson and colleagues (J.M. Watson, Collis, Callingham, & Moritz, 1995; J. M. Watson, Gal, & Garfield, 1997).

We state that 'data literacy' (DL) is the domain of statistical activities which Cobb (1991) indicated as 'data and concepts', while 'algorithmic statistical skills' can be characterised by 'recipes'. We consider data literacy to be more reflective, with activities like estimating, comparing, interpreting, reasoning, and concluding. In other words more qualitative activities that are difficult to outsource to computer technology.

'Algorithmic statistical skills'(ASS), as defined as a goal of statistics education by Gal and Garfield (1997) with 'Master procedural skills', on the other hand, is a sub domain in which the emphasis lays on numerical and mathematical activities like determining (e.g. the mode), calculating (e.g. the standard deviation), representing (e.g. a data set by a stem-leaf diagram), or constructing (e.g. a boxplot). These are practical, well defined activities, very suitable to be performed using computer technology. We will demonstrate the distinction between DL and ASS with concrete learning activities in section 3.4.

We conclude this section with the remark that the goals of the Dutch mathematics A curriculum (being non-formal and focussed on application), at the intended level, are more or less outlined with those described above. The question of whether some of these aspects are present in the implemented curriculum will be posed – and answered – in section 3.4 by the analysis of two textbooks.

3.2.2 How to design statistics education

The 'how' of statistics education

Now we have made our choices with respect to the goals of statistics education explicit, we will discuss in this section its design by describing some of its characteristic elements.

Research recommendations on the '*how*' were reviewed and summarised by Garfield and Ahlgren (1988, p. 48), resulting in these guidelines:

1. Introduce topics through activities and simulations, not abstractions;
2. Try to arouse in students the feeling that mathematics is related usefully to reality and is not just symbols, rules and conventions;
3. Use visual illustrations and emphasise exploratory data methods;
4. Teach descriptive statistics without relating it to probability theory;

5. Point out to students common misuses of statistics (say, in news stories and advertisements);
6. Use strategies to improve students' rational number concepts before approaching proportional reasoning;
7. Recognise and confront common errors in students' probabilistic thinking;
8. Create situations requiring probabilistic reasoning that correspond to the students' views of the world.

For this study, we concentrate on guidelines 1, 2, 3, 4 and 8 (in 8 we replace 'probabilistic' by 'statistical'). In our view, they all belong to data literacy. We will discuss them in the sections below. We do not integrate common misuses of statistics (Garfield and Ahlgren guideline 5) because we consider that this could be a too tedious approach for this target group. We do not follow guideline 6, because proportional reasoning is not a learning goal of our intervention. Guideline 7 considers probabilistic thinking which we try to avoid when starting statistics instruction.

Central place for student activities

The first Garfield and Ahlgren guideline can be regarded as supportive for our view of the centrality of students' activities. This guideline, to introduce topics through student activities, and not through 'theory', that is usually an abstraction delivered by teacher or textbook, finds its roots in the work of Dewey (1929) and Vygotsky (1962, 1978). It has been made operational in mathematics education, amongst others, by Freudenthal (1973) and is an important pedagogical approach. Garfield (1995) states that constructivism, the learning theory that puts student activities in a central place, is the guiding theory for much research and reform in mathematics and science education.

In our view, in domains where algorithmic skills play such important roles as in mathematics, science and statistics, student activities are key elements. Without practicing themselves, students will find it difficult to master these skills. Data literacy, on the other hand, can only be fostered in a learning environment in which students are challenged to formulate their own ideas about data and data handling and in which they get feedback on their ideas from the teacher and their peers.

Authentic contexts

In section 3.2 we saw that contexts play a crucial role in statistics. Therefore, it is very important to incorporate contexts in statistics education. But are all contexts suitable for learning statistics? Freudenthal (1983) stated that contexts in mathematics education should have phenomenological richness. Nowadays, educators prefer a context, from which the learning activity is to be deduced, that is *authentic* (Verschaffel & de Corte, 1996).

With respect to mathematics education, Wijers et al. (2004) distil the following *authenticity questions* for contexts, based on Roelofs and Houtveen (1999) and Hoefakker (2002):

1. Do they fit into the students' lives?
2. To what extent are the contexts relevant/meaningful for out-of-school situations?
3. Is there a complex problem through which knowledge construction is possible, when using an own approach?

4. To what extent are the tasks to be performed consistent with those in the professional field?

We consider the fourth question to be the criterion for *authenticity of tasks*. The first three criteria can be considered as criteria for a context to be challenging to authentic tasks.

With an eye on these four, we formulate four criteria for the authenticity of a context for the sake of statistics education. In order to be called 'authentic', a context must:

1. Be *real*. The situation the context describes is based on 'real life'.
2. Be *attractive* for students, for instance by staying close to their private lives. Lesh et al. (1997) observe that students gain enthusiasm when they are able to link their learning activities to their personal life. In our view, this aspect draws heavily on the choice of the contexts chosen.
3. Invite *meaningful* statistical operations, offering "*phenomena that beg to be organised*". From the authentic context a central question should naturally emerge. Along this central question, the statistical operations should be organised. These operations should be consistent with those in the professional field.
4. Induce activities that *cover the entire statistical process*. Witterholt, van Streun, Goedhart, and Beijaard (2007) state that formulating the research question, planning, data collection, data ordering, data analysis and drawing conclusions are all to be included. This study is just in the field of descriptive statistics, hence we use *data ordering, data analysis and drawing conclusions* to cover the statistical process.

We will use these criteria in section 3.4 in order to analyse two Dutch textbooks in statistics in order to answer the question: do they offer authentic contexts for the learning of statistics?

Visual representations of data

When wanting to develop data literacy, it seems very supportive to let the students make diagrams that represent the given data sets. This is acknowledged by Garfield and Ahlgren (1988) too. Their third guideline advises to use 'visual illustrations'. In general, students learn better from words and pictures than from words alone (Mayer, 2009). Cleveland and McGill (1984) mentioned that graphs are a vital part of statistical data analysis. Cleveland (1993, p. i) stated that: "[visualisation of data reveals]… *intricate structure in data that cannot be absorbed in any other way*." After Tufte (1983) introduced amongst others the term 'chartjunk', hardly anyone could see statistics being practised without the use of precise and elucidating graphics. Bakker (2004, 2007) and Bakker and Hoffmann (2005) showed that in grade 7 and 8 what they called *diagrammatic reasoning*: making a diagram, experiment with it and reflect on the results, led to good student reasoning. Apparently, working with diagrams has a central place in statistics education.

ICT in statistics education

When trying to stress data literacy, it is important to use authentic contexts and, thus, real data sets. However, these sets tend to be large and calculations become tedious and cumbersome. ICT, for instance a graphing calculator (GC), can offer substantial help in bringing back the time needed to perform these. For the visual representation of data, the GC offers good possibilities too.

There is a very important practical consideration that can be used to advocate the use of ICT in statistics education. When using sets of real data, as we would like to see happen, directing to the importance of data (Cobb, 1991) and the importance of attractiveness for the target group (Garfield & Ahlgren, 1988), students usually have to handle a considerable amount of data (Mills, 2002). This is almost impossible without the support of ICT. Moore (1990) stated that computers and calculators have made more complex analyses on larger data sets possible. Moore (1997, p. 130) adds that: "*Consistent emphasis on visualization and problem-solving are hardly possible if graphics and calculations must be done by hand."* Garfield (1995, p. 29) notes that "*using software that allows students to visualize and interact with data appears to improve students' understanding of random phenomena* (Weissglass & Cummings, 1991) *and their learning of data analysis* (Rubin, Rosebery, & Bruce, 1988)." Thus, in a way, the use of ICT in statistics education, with goals as we described in section 3.2.1, is inevitable.

With respect to ICT, we first look at the specific possibilities of the graphing calculator (GC) which we will utilise in our intervention. Doerr and Zangor (2000, p. 151) specify five functions for this device in mathematics education. We will shortly describe per function the implementation in statistics education.

1. *Exploration* of the data set; quickly ordering the data and establishing the minimum and maximum.
2. *Calculation* of specific measures; calculating the standard deviation of a data set of 100 observations can be done quickly if the data are entered in the GC.
3. *Transformation* of the task. Doer and Zangor suggest that tedious computations are transformed into interpretations: "Is this a reasonable value for the SD of this data set?"
4. *Visualisation*; a boxplot and a histogram can be created relatively quickly, especially with extensive sets of data.
5. *Checking*; while involved with statistical tasks, all kinds of conjectures rise in a student's head. With ICT these conjectures can be checked.

Doerr and Zangor point at the fact that when ICT is used in the classroom with students working individually there will be no discussion and, therefore, no chance for the students to learn from each other. Of course there can be, and usually will be, individual internal reflections, but these will not be made productive in the classroom discourse. As a result of this limitation, we add a sixth specific possibility when working with GCs in a classroom network:

6. *Interaction*; ICT can be used as a means of communication. In this study we are specifically interested in a classroom network of graphing calculators. The GCs are then also used for interaction with the teacher. When the teacher decides to reveal this interaction to the whole classroom, students participate in a shared social mathematical discourse (Stroup, et al., 2005).

Researchers seem to be quite optimistic about the possibilities of ICT in statistics education. Cobb (1991) states that calculations and graphics should be automated, to the maximum extent feasible. Burill (1997, p. 15) mentions that "*technology makes statistics and statistical reasoning accessible to all students*". It is interesting that she also mentions 'statistical reasoning', because reasoning is something you do not outsource to a computer. Burill does not state it explicitly, but it is likely she means that curricular (and

mental) space, created by outsourcing technical skills to technology, should be used for 'statistical reasoning'.

We will not deny that ICT can broaden the statistical learning experience. However, there are questions to be answered. How does educational design handle the black-box character all ICT intrinsically has? Which tools do we choose? Biehler (1997) discussed the pedagogical possibilities of ICT in statistics education, the requirements for software and the interesting analogy between the functional needs that have been implemented in ICT that supports statistical investigations and those needs in statistics education. He concluded, in 1997, that the ICT tools then available for professional statisticians lack the features that make these tools really suitable for the use in statistics education.

Another possible advantage of the use of ICT in statistics education could be the reduction of the *cognitive load* (Sweller, 1988) needed for technical operations (Ben-Zvi, 2000). Moore (1997, p. 131) states that "*Good software reduces the students' cognitive load, replacing complex algorithmic procedures by simpler commands, thus allowing learners to focus on higher-level understanding."*

In our view, this is not a pure example of reduction of cognitive load (Chandler & Sweller, 1991), but more an example of transition of statistical activities. However, this transition can be helpful when one of the goals of statistics education is a shift in focus from techniques to concepts. ICT can be used, for instance, to calculate very quickly the variance of a large data set, a cumbersome activity when performed by hand. Time saved by using ICT could, for example, be used to reach the Gal and Garfield (1997) goal 6 of statistics education, to 'develop interpretive skills and statistical literacy', or to put an emphasis on the data and concepts used (Cobb, 1991). Meaningful questions are thus: which part of the curriculum can be outsourced to ICT, how is curricular space used to attain more conceptual understanding and does this work in classroom practice?

Successful use of ICT in statistics education has been reported. Morris and colleagues (2002) reported that activities involving the direct, computer-supported, manipulation of data significantly improved students' understanding of measures of central tendency.

Making diagrams is one of those statistical skills where ICT can have a specific efficacy advantage (Ainley, Nardi, & Pratt, 1998; Moore, 1990). Using ICT to visualise and to interact with data appears to improve students' understanding of random phenomena (Weissglass & Cummings, 1991) and their learning of data analysis (Rubin, et al., 1988).

We conclude that there are considerable pedagogical possibilities when using computers in *statistics* education. Additionally, the application of ICT could be a result of *general pedagogical* choices, for example through the motivational aspects of ICT for students' learning (Passey, Rogers, Machell, McHugh, & Allaway, 2004) or changing work force requirements (Trilling & Fadel, 2009).

Garfield and Ben-Zvi (2007) mention besides this that "*technological tools should be used to help students visualize and explore data, not just to follow algorithms to pre-determined ends*". Educational software can support students to understand abstract ideas. Developing understanding of the Central Limit Theorem (stating that, and the conditions under which, the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed) can be reached by using ICT to construct various populations, calculate some statistics of samples from these populations (e.g. the mean) and then observe the distributions of these statistics (Ben-Zvi, 2000). ICT can also be used to explore statistical models, to change

assumptions and parameters for these models and to analyse data generated by applying these models (Biehler, 1991; Jones, Langrall, & Mooney, 2007).

We conclude that ICT can play an essential role in the support of teaching and learning of statistics, provided a thoughtful pedagogical integration in the learning environment is achieved.

3.3 Realistic mathematics education (RME)

In section 3.1 we stated that, despite the differences, mathematics and statistics have a lot in common. As a working hypothesis we take mathematics education as a potential source of inspiration for the pedagogy of statistics education in general. Of interest is the question: is there a view on mathematics education that is particularly fruitful for statistics education?

In section 3.2.1 we cited Cobb (1991) stating that in statistics education there should be more focus on the data and less on theory and recipes. Further, we saw in section 3.2.2, that Garfield and Ahlgren (1988) state, amongst others, as guidelines for the design of statistics education the need to:

1. introduce topics through activities and simulations, not abstractions;
2. try to arouse in students the feeling that mathematics is related usefully to reality and is not just symbols, rules and conventions.

The first guideline states that student activity should be a starting point. In this study we call that the *questionising approach*: try to represent as many of the learning and teaching activities as possible as questions to the students. The second guideline explicitly highlights the bridge between the real and the mathematical world, as has been advocated by, amongst others, Freudenthal (1973). When trying to design statistics education according to Cobb's recommendation and to the two guidelines cited above, *realistic mathematics education* (RME) comes into sight.

RME "… *can be characterized by a complete break with the traditional approach, which goes from 'theory' (principles, rules, theorems) to 'practice'*" (Korthagen & Kessels, 1999). This traditional approach is usually called a deductive approach. In Freudenthal's view, however, mathematics is "*not a created subject*" to be transferred to children, but "*a subject to be created by children*" (1978, p. 72). In his view, this nature of mathematics has a profound impact on mathematics education: students should be offered the guidance to 'reinvent' the mathematics themselves. This roughly means that RME builds up from practice to theory: an inductive approach in a process towards abstraction called 'progressive mathematization' (Treffers, 1987). This approach fits in with the first design guideline by Garfield and Ahlgren (1988, p. 48) with respect to statistics education: "*introduce topics through activities and simulations, not abstractions*".

Treffers (1993, p. 89) characterises the discriminating aspect of RME with respect to mathematics education with a more formal approach as follows:

> "*New in Freudenthal's* (1971, 1983) *views was not only that he wanted to incorporate everyday reality emphatically in mathematics education, but especially also his fundamental idea to let that rich context of reality serve as a source for learning mathematics*."

Didactical phenomenological analysis (Freudenthal, 1983) can be used to identify potentially suitable reinvention routes. From the viewpoint of didactical phenomenology,

the designer of mathematics education should search for phenomenological rich contexts, in which "…*phenomena beg to be organized*…" (Freudenthal, 1983, p. 32). In statistics education, in which data sets and the contexts from which they are deeply rooted into the fascinating world called *reality*, it should be possible to find contexts that are phenomenological rich. The *authenticity* of contexts was discussed in section 3.2.2.

Concluding, we state that the insights about how to design statistics education and the ideas behind RME are similar. Therefore, we consider RME in the rest of this study to be a background inspiration.

3.4 Analysis of textbooks

In this section we discuss the way the textbooks current in 2006 (during the preparation of the pilot of the first prototype of the intervention) adopt the main recommendations with respect to the goals and design of statistics education as summarised in section 3.2.

3.4.1 Motive and goal

Since possible answers to the main research question that leads this study seem to be appropriate to be explored by an educational design study, we had to develop or perhaps reuse teaching materials. In order to select useful materials from existing sources, we decided to do a content analysis of two textbooks.

The goals we have with the analysis of textbooks are threefold:

1. To determine whether textbooks are in line with the results of recommendations as summarised in sections 3.2 and 3.3;
2. To select thus identified materials in order to reuse them in our intervention;
3. To determine statistical learning goals to be covered.

The (parts of) textbooks we have chosen to analyse are:

1. Moderne Wiskunde (Modern Math), 7th edition 1998, MW Havo mathematics A1,2 part 2, chapter S2
2. Getal & Ruimte (Number & Space), 1st edition 2003, Havo mathematics A - part 2, section 7.1, 1st edition 2005, 3 Havo-2, section 9.1

We have chosen these textbooks for two reasons:

1. They together cover about 85% of the market, meaning they are very influential in Dutch classroom practice.
2. They have a reputation of complementary approaches: Moderne Wiskunde is said to be more realistic, Getal & Ruimte is supposed to be more formal.

We have chosen to analyse these specific parts of the two textbooks because they roughly cover the topics we plan to be the topics of our intervention.

3.4.2 Focal points, analysis questions and method

Our study takes place in the domain of descriptive statistics. In our intervention we will be utilising graphing calculators in a wireless network, being a specific form of ICT. From the curricular focal points formulated in section 3.2 we restrict ourselves here to the four most important ones. The analysis of textbooks would otherwise become very lengthy and out of the primary focus of our research. We concentrate on:

1. *authentic contexts*; in section 3.2.2 we described four criteria for authenticity of contexts;
2. *data literacy*; in section 3.2.1 we described this type of statistical activity;
3. *algorithmic skills*; in section 3.2.1 we described this type of statistical activity;
4. *use of ICT.*

In section 3.2.1 we described the distinction in statistical activities between 'data literacy' and 'algorithmic statistical skills'. In order to be more concrete about what kind of activities belong to either of these categories, we will present in this section examples of both categories. Therefore, we have studied all the elementary exercises from the Dutch textbook Moderne Wiskunde (Modern math). This is one of the two textbooks to be analysed and by reputation perhaps the one that best fits the guidelines emerging from research. When we consider the distinction in activities between DL and ASS clear enough, we adopt one textbook with respect to this.

How did we proceed? For each exercise, we described the (intended) statistical activity using a verb. After that, we counted the number of times each verb occurred. We then assigned them to either the category of 'data literacy' or 'algorithmic skills', based on the descriptions given in section 3.2.1. After summing up all verbs in both categories, we try to describe again the nature of each category.

We have chosen to conduct the analysis from the perspective of the curricular focal points of our study, based on the conclusions from research. These points (authenticity of context, the ordering of the data, data literacy-statistical skills and ICT use) are to be investigated in the analysis of the textbooks. The resulting analysis questions, with their corresponding methods, are presented in table 3.2.

Table 3.2 Questions and method for analysis of statistics textbooks

No.	**Analysis question**	**Method**
1.	Are the contexts used authentic?	Determine per assignment whether it is based on an authentic context.
2.	How does the textbook try to develop data literacy?	How many exercises can be regarded as aiming to develop 'data literacy'? What kind of tasks are the students to undertake?
3.	How does the textbook try to develop algorithmic statistical skills?	How many exercises can be regarded as aiming to develop 'algorithmic statistical skills'? What kind of tasks are the students to undertake?
4.	What role does ICT play in the textbook?	Is there a role for ICT in the textbook (per exercise)? Which role: exploratory, technical supportive, to check the results? What kind of ICT is used?

We have analysed both textbooks from the perspective of these questions. We concentrated on the 'basic part' of the textbooks, covering the educational goals.

3.4.3 Results

One of the most important aspects of the textbooks with respect to our study is: how do they try to develop data literacy (DL) and how do they try to develop algorithmic statistical skills (ASS)? In order to specify DL and ASS we inspected the text book parts on descriptive statistics and we assigned all the exercises, with their key student activity as represented by the verb, to either DL or ASS. This resulted in:

Data literacy (number of activities: 32):

1. Estimating (4)
2. Comparing (3):
3. Interpreting (9):
4. Reasoning (10):
5. Concluding about (6):

Algorithmic statistical skills (number of activities: 79):

1. Determining (19):
2. Calculating (30):
3. Representing an ordered data set as a histogram (1)
4. Reading off (6):
5. Constructing (23):

Generally the distinction between DL and ASS was unambiguous although there were some cases where crossover occurred. 'Reading off' was classified as a 'Skill', although this is an activity similar to 'Interpreting' in the DL category. The main difference between both activities is the quantitative aspect: an algorithmic skill is usually performed with a quantitative goal. However, one of the main goals of data literacy is to understand statistics at a conceptual level, for instance to understand a specific feature of a certain way of visualising a data set and to be able to articulate this. Another borderline case is the skill 'estimating'. We reasoned that this is a heuristic, not aiming at 'exact' quantification, but at general quantitative intuition, thus belonging to 'data literacy'. On the other hand, it would have been plausible to categorise it as a statistical skill, usually following a certain pattern and having a quantitative goal.

What can we conclude about these activities? We see that activities belonging to 'data literacy' are: estimating, comparing, interpreting, reasoning and concluding. These activities are characterised by *more or less qualitative thinking*, often with a reflective character.

'Algorithmic statistical skills' consists of the activities: determining, calculating, representing, constructing and reading off. These activities are characterised by *practical, well defined actions*, often with a quantitative character or goal.

We have to note that by this classification, the activities are formulated abstracted from their statistical nature; almost all of them are employed elsewhere in mathematics and science education too.

For reasons of reliability, the classification of 'data literacy' versus 'algorithmic statistical skills' and of 'authentic context' versus 'artificial context' has also been done by a second expert. Before she started categorising, we sent her the definitions of DL and ASS as formulated in section 3.2.1 and we had a twenty minute talk about these definitions in order to clarify the concepts, which were at that time new for her.

When working independently, 81% of the exercises with respect to DL/ASS were scored the same way (Cohen's kappa coefficient 0.62, $n = 181$) and 86% of the contexts with respect to 'Authentic/Artificial' were scored the same way (Cohen's kappa coefficient 0.73, $n = 54$). After deliberation on the categories, we could agree on 100% of both categorisations. We conclude that the categorisations 'Data literacy' versus 'Algorithmic statistical skills' and 'Authentic context' versus 'Artificial context' are sufficiently reliable.

Finally, we got the results presented in table 3.3.

Table 3.3 Results of the analysis of statistics textbooks

Analysis Question	**Modern Mathematics**	**Number and Space**
1	Authentic : Artificial = 5 : 20 (20% real)	Authentic : Artificial = 2.5 : 26.5 (9% real)
2	33 exercises focus on 'Data literacy' (34%)	25 exercises focus on 'Data literacy' (30%)
3	64 exercises focus on 'ASS' (66%)	59 exercises focus on 'ASS' (70%)
4	No computer software is needed. In 3 of 25 assignments the GC is used.	ICT plays an explicit role. This role is mainly exploratory and checking. Explicit attention is paid to the GC instruction; Lists (an essential concept in automated data analysis) are explained and used in examples. The GC is presented as a natural partner.

Textbook: Moderne Wiskunde (Modern Mathematics)

A typical example of focus on data literacy is given in assignments 16 and 17. The given data set represents the weights of forty bags of Bintje potatoes and forty bags of Doré potatoes (Bintjes and Dorés are both well-known species of potatoes in The Netherlands). In assignment 16 two boxplots are constructed and based on these diagrams the central question is asked (in the fifth and last exercise of this assignment): which potato species has the largest variation in weight? In assignment 17 similar exercises are to be performed on the same data set, but now with the average deviation from the mean as a measure of spread. Again in the last exercise the central question is posed: what potato species has the largest variation in weight? Although we consider these assignments to be good examples of data literacy, there are two points we would like to add, from the perspective of data literacy:

1. As noted before, we consider it to be pity that in an 'almost authentic' context, the central question is in both assignments only posed in the last exercise. Students would be able to think about statistical procedures more by themselves if these two assignments would *start* with the central question ("What kind of potato species has the largest variation in weight?") and with some explanation about the procedure (using two different ways of presenting variation).
2. When working out two different ways of the presentation of variation, the development of data literacy would be enhanced when the ultimate questions are posed: how do you compare both methods and their results and what is your final conclusion about relative variation in weight between Bintjes and Dorés? Unfortunately, the textbook does not pose these questions. It is therefore doubtful whether the textbook aims at developing data literacy and if so, if it succeeds.

In a typical example of an assignment aiming to develop algorithmic statistical skills a certain data set is given. Students have to calculate respectively: the mean, the deviation of each observation from the mean, the square of the deviation of each observation from the mean, the sum of the squares of the deviation of each observation from the mean, the mean square of the deviation of each observation from the mean and the square root of the mean square of the deviation of each observation from the mean. Then the students are asked why this last statistic is a good measure for the variation in a data set. Following this they are asked why a positive or negative deviation from the mean does not matter. These last two exercises, of course, do not belong to the category 'algorithmic skills', but are examples of data literacy. The students are asked to find out how the calculation of the standard deviation is to be performed on the graphing calculator, an example of an algorithmic statistical skill (in an ICT environment). Following this, four context-less sets of data are presented with the students required to calculate the standard deviation.

Textbook: Getal & Ruimte (Number & Space)

A typical example of an assignment that aims to develop data literacy in this textbook is the following: for employees of the Nijha Company the mean weekly salary (apparently non-authentic, because nowadays in the Netherlands salary is usually paid and calculated per month) is €610 with a standard deviation of €75. Assume all employees get a raise in salary of €50. This is another non-authentic element of the context, because this is very unlikely to happen in real life. The students are then asked how this uniform raise in salary influences the mean salary and how it influences the standard deviation of the salary. Assume all employees get a raise in salary of 5%. What happens to the mean salary and what happens to the standard deviation of the salary?

We consider the presumed dominant activities to belong to data literacy. We interpret these two exercises to aim at reasoning about the transformation of the data and the shape of the distribution. However, students could consider it to be just technical exercises. These students just adapt the list in their graphing calculator representing the salaries and recalculate mean and SD in both exercises. They then conclude that in the first exercise the mean rises to €660 and the SD does not change, and in the second exercise, the mean rises to €640.50 and the standard deviation rises to €78.75. These students are not asked explicitly to indicate that thus both the mean and the standard deviation rise 5%. They do not have to give comment on the fact that both mean and standard deviation rise with the same percentage and they do not have to compare the two different ways of salary rise (with a fixed amount or with a fixed percentage). In our view, these are chances to develop data literacy that are missed, and it is doubtful whether this assignment will reach its goal.

Assignment 12 is a typical example of an assignment aimed at developing algorithmic statistical skills concerning the introduction of standard deviation. Assignments 1-11 are all about boxplots and variation. In assignment 12 three histograms are presented, representing the results on a test from three parallel groups in a certain school. Students are asked to order these groups with respect to the size of the variation. Then the standard deviation is algorithmically explained. Some guidelines for calculating the standard deviation on the two most popular graphing calculators are given. After that, a worked example is presented, based on a frequency table containing the errors of a certain group with respect to a writing test. Exercises 13-16 are all very similar: a one sentence context with data in a frequency table in order to calculate the mean and the standard deviation.

3.4.4 Conclusions from the textbook analysis

Authentic contexts: both methods actually fail when it comes to the use of authentic contexts. Data originating from the contexts are not delivered as 'raw data', but are usually presented somewhat abstracted. In neither of the methods is ordering mentioned as an important operation in descriptive statistics. The textbooks very rarely present real contexts. There are no central questions posed explicitly at the start of an assignment to make the statistical operations relevant from the start. In its best assignments, one of the textbooks presents such a question; unfortunately this appears in the last exercise. In the other textbook, central questions are completely lacking.

Data literacy: about one third of the exercises presented in the textbooks aim to developing DL. One textbook uses the power of the data sets for reflection, the other textbook focuses more on reflecting on the statistical techniques themselves. The first method is in our view more attractive for students, because we expect them to be more interested in the real life phenomena that statistics studies than in the techniques of statistics. However, this method draws rather heavily on the didactical quality of the data and the authenticity of the contexts used. Unfortunately both textbooks fall short with 20% and 9% respectively of the contexts used obeying the criteria of authenticity.

Algorithmic statistical skills: Both textbooks differ in the way they aim to develop ASS. One textbook uses contexts to make the usefulness of the skills plausible for the students. Skills are thus more or less integrated into the statistical activities. The other textbook usually abstracts from the context then focuses rather narrowly on the skills in a somewhat repetitive way, and does not return to the context. This causes an impression of 'statistics for statistics', which is unlikely to appeal to the target group.

The role of ICT: is bigger in the textbook published in 2003-2005 (Getal & Ruimte) than in the 1999 textbook (Moderne wiskunde). But when trying to stress the role of real data more than is done in these two textbooks, and using ICT for interactional purposes, the position of ICT should perhaps become even more central.

All in all, we conclude that both textbooks do not really meet the requirements as we formulated in section 3.2, so we decided to design our own teaching materials from scratch.

3.5 Conclusions

We now summarise the conclusions from sections 3.1-3.4. With these conclusions, in chapter 5, design guidelines and principles will be formulated.

The scientific disciplines of mathematics and statistics differ, but their commonalities legitimate the use of the domain of mathematics education as a source of inspiration for the education of statistics.

The way statistics education is advised to be designed, as summarised in section 3.2.2, resembles RME. RME recommends starting with 'practice' and building up to 'theory' (inductive approach). The practice should consist of phenomenological rich contexts, from which the central statistical question naturally arises. These contexts should be '*authentic*': be real and attractive for students, inviting to meaningful statistical operations and the activities to be undertaken should cover the entire statistical process.

Taking these authentic contexts as a starting point for exploration and analysis, a statistics education emerges that fosters:

1. the students' understanding of the purpose and logic of statistical investigations and their development of interpretive skills and statistical literacy; we call this '*data literacy*' (DL);
2. the students' development of procedural skills and their understanding of mathematical relationships; we call this '*algorithmic statistical skills*' (ASS).

These goals are recommended to be reached by introducing topics through activities and simulations, not through abstractions. Students' activities should be at the very core of the intervention. The students should realise that mathematics is related usefully to reality and is not just symbols, rules and conventions. With the use of visual illustrations and emphasis on exploratory data methods the students could improve their imagination for data. ICT should play an essential role, because the outsourcing of tedious and time consuming calculations on large data sets frees mental capacity for reasoning about these data sets. ICT should be used to: explore, calculate, visualise, check and communicate about statistics. To support the communication, we will use a classroom network.

An analysis of the two market-leading Dutch textbooks on mathematics education, on the subject of statistics education, shows that with respect to the attributes that are most important for this study (support of data literacy, use of authentic contexts and the use of ICT in order to facilitate the feedback process) they do not offer enough possibilities. Hence, we decided to design our student activities from scratch.

Finally we note that the body of research with respect to the instruction of descriptive statistics is in nature rather descriptive itself. Research that is supported by empirical data does not seem to be leading. This study tries to supplement this type of research.

Chapter 4 **Research methodology**

In this chapter we describe the design research methodology chosen and argue why it is suitable in the context of our study. We continue with a description of how the methodology has been applied in this study. We distinguish stages and phases in the study, relate it to the Analysis, Design, Development, Implementation, and Evaluation *(ADDIE) components and depict the iterative developmental nature. After that, we describe each component as it was used in this study. Finally, we discuss some issues of validity, reliability and participant anonymity.*

4.1 The main research question specified by four subquestions

In section 1.8 we formulated our main research question: "What are the potentials of a classroom network in supporting teachers with providing feedback?" We now formulate four subquestions, specifying the potentials.

With respect to the *feasibility* of the chosen approach, we ask ourselves whether the chosen technological setting is adequate for supporting feedback and whether a mathematics teacher was able to utilise this support in order to provide feedback as intended. Thus two subquestions arise from the main research question:

1. *Is technological support by means of the classroom network adequate for the intended feedback in the lessons?* (Conditional question)
2. *Is it possible for a mathematics teacher to implement the prototype in accordance with the intentions?* (Existence question)

After having distinguished the statistical learning goals in section 3.2.1, we ask ourselves whether the realised feedback was both on ASS as on DL activities. Therefore we specify a third subquestion with respect to the general research question:

3. *Is the feedback support of the classroom network equal for ASS and DL?* (Didactical question)

While realising that teacher behaviour is the first focus of this study, we would like to identify the situations in which feedback was realised as intended and in which situations it was not. Thus a fourth subquestion emerges:

4. *Which teacher characteristics promoted/hindered the implementation of the CN as intended?* (Identification question)

This subquestion is rooted in our sense of the complexity of the intended intervention. A teacher will have to be able to manage a lot of actions simultaneously: the statistics itself and the learning goals, the students' input as collected and rudimentary analysed by the CN, the ICT environment, and the flow of the classroom discourse as initiated by the teacher feedback. A close observation of teacher behaviour should result in data with which this subquestion can be answered.

4.2 Educational design research

In this section we discuss the educational design research (EDR) approach and its appropriateness for this study.

4.2.1 What do we mean by educational design research?

Activities aiming at improving education by the use of specifically designed educational artefacts are perhaps as old as education itself. However, a systematic approach to these design activities and a thorough evaluation of them do not have a that long a history. In a well-known publication about *design experiments,* Brown (1992) describes how she developed from a classical educational psychologist, using control group laboratory-based research techniques, to an educational design scientist, developing materials, using them in real life classrooms and then evaluating the classroom results systematically in order to answer the research questions. All techniques Brown describes were well known, if not from learning science then from other design-oriented sciences, like for instance the engineering sciences. This approach was called *educational design research* (EDR), also known as development(al) research; although we prefer the adjective 'developmental' for our specific study, because we actually developed an intervention, which is more concrete than just designing it, we conform ourselves to the term 'design research' being most common. During the 1990s, educational design research gained momentum amongst educational researchers (Gravemeijer, 1998; Richey & Nelson, 1996; van den Akker, 1999) which was extended during the first decade of the twenty-first century (Collins, Joseph, & Bielaczyc, 2004; Kelly, 2004; Kelly, Lesh, & Baek, 2008; van den Akker, et al., 2006). Although there are different interpretations and articulations of the principles and methods to be applied when conducting EDR, some key elements frequently show up. Five of these are as follows (van den Akker, 1999):

1. *Interventionist*: a particular intervention is designed and implemented in an actual setting of 'real life education';
2. *Iterative*: a cyclic process of design, evaluation and revision is followed;
3. *Process-oriented*: aimed at understanding the process and improving the intervention;
4. *Utility-oriented*: aimed at being as practical as possible;
5. *Theory-oriented*: based on existing theory and aiming at contributing to theory building of poorly understood contexts.

The first element, that of an EDR study having an interventionist character, we followed closely in this study. The developed intervention is used and evaluated during real life mathematics education. The second element of EDR, being iterative, was also a key in this study. Each time we piloted our prototype we evaluated the experiences with using this prototype and adapted the prototype. We conducted an initial study (see section 1.7) to determine the key concept, which turned out to be teacher feedback on students' work. We piloted the developed prototype during three stages. The third element of EDR, the orientation on the process, we tried to meet very closely too. In fact, we constantly evaluated the whole context of the pilot of the prototype, in order to be able to research the teaching process, prompted by the prototype. The fourth element, EDR being oriented on utility, we followed in this study by concentrating on an urgent problem in real life education: a lack of teaching time in upper secondary mathematics education. By cooperating with practicing mathematics teachers, we constantly validated whether the relevance of our goal is felt in educational practice and we could evaluate the usability. This is the most important reason why EDR is considered by educational practitioners (like teachers) to be the most useful type of research (Vanderlinde & van Braak, 2010). The fifth element as formulated by van den Akker (1999), the orientation on theory, we had express intent to meet in this study by examining theoretical reviews of the concept

of feedback, of statistics education and of the use of information and communication technology (ICT) into design principles. These principles were confronted with the analysis of the collected data and evaluated, in order to be able to contribute to a theoretical understanding of how to support teachers in statistics education utilising a wireless classroom network. We aim thus, in short, to build on, as well as to contribute to, theory.

In the next section, we will illustrate these elements in more detail within the context of our study.

4.2.2 Why is EDR suitable for this study?

We started this study with an initial study (Tolboom, 2005) serving a very exploratory goal: what do we observe when we utilise a wireless classroom network in mathematics education (see section 1.7)?

After observing that this yielded lively interaction around mathematical objects and processes (Sfard, 1991), we concluded that a classroom network could be a fertile addition to the learning environment, because of *enhanced feedback possibilities*. Then the question arose: "What are the potentials of a classroom network in supporting teachers with providing feedback?" which is posed in section 1.8 as the main research question guiding this study. This question starts with '*what*', implicating a *how*, that is: researching a *mode*, eventually trying to find out *why* (an intervention succeeds or fails with respect to its goal). Educational design research (EDR) is supposed to be a logical paradigm for this research. Kelly (2007) states that design research is most appropriate for '*open* or, more appropriately, *wicked* (Rittel & Webber, 1973) problems". Our problem is *open* since the technology we are exploring and its use are very new. They are that new we presume we will need several iterations in order to create a teaching setting specific enough to offer us data that could possibly lead us to an answer to our research question. Our problem could be *wicked*, because feedback, although a classical theme in learning science, is still not completely understood (Cohen, 1985; Shute, 2008) and not very well structurally implemented in classroom practice. Hattie (2009) calculates, in his synthesis of over 800 meta-analyses relating to achievement, a mean effect size of 0.72 for controlled experiments using feedback. This score is one of the highest effect sizes that were collected with the more than 800 interventional studies. If feedback is so powerful, why is it not the basic mechanism of all education worldwide?

Interventions drawing heavily on the use of educational technology are in general suitable for a design research approach (Reeves, 2006; Wang & Hannafin, 2005). Cobb and colleagues put it in more general terms: "*Design studies are typically test-beds for innovation*" (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003, p. 10). While we see educational technology as a potential pivot for innovation of instruction, we consider the last two arguments to be consistent. Since, with this study, we try to improve feedback in statistics education, utilising an innovative technology, we consider design research to be an appropriate methodology.

Gravemeijer (1994) introduced the metaphor of the design researcher as a tinkerer: "*a handy person who uses as much as possible those materials that happen to be available*" (Gravemeijer & Cobb, 2006, p. 21). Simon (2009, p. 488) stated that "*mathematics education researchers* [...] *need to (when appropriate) engineer pragmatic coordination of analyses done from different theoretical perspectives*". This post-modern approach suits us well, because in this study we need a variety of ingredients such as:

1. Methods and results from research on feedback in education, a topic in educational psychology dating from the early 1960s (Waimon, 1962).
2. Domain specific learning theories about realistic mathematics education (1980's) and statistics education (1990's).
3. Research on the use of educational technology that gained momentum from the introduction of the personal computer (Xerox Alto, 1975) and accelerated with the breakthrough of the World Wide Web (1990's).

The design research approach adopted the insight of triangulation of theory and methodology (and of data and investigators) (Denzin, 2006). This allows us to be eclectic and lets us use the elements above to build a consistent framework of methods. Consistency, of course, cannot be reached by just applying the methods available. Design researchers always have to underpin the choice of the subset of methods for each design study again. In this chapter, we describe the considerations behind our choices.

Van den Akker (1999, pp. 4-5) states that *design research* in the domains of teacher education and didactics was at the end of the twentieth century already relatively well-established:

> "*The primary goal is usually to contribute to the teachers' professional learning and/or bringing about change in a specific educational setting (Elliott, 1991; Hollingsworth, 1997). In the area of didactics, the emphasis tends to be on 'developmental research' as an interactive, cyclic process of development and research in which theoretical ideas of the designer feed the development of products that are tested in classroom settings, eventually leading to theoretically and empirically founded products, learning processes of the developers, and (local) instructional theories…*"

This study is an interactive, cyclic process of development and research, using theoretical insights to develop an intervention to be tested in classroom settings, eventually contributing to a specific theory of teaching, by bringing change in a specific educational setting. It therefore seamlessly fits into van den Akker's description of design research. Nieveen, McKenney and van den Akker (2006) distinguish the nature of EDR in two, more or less complementary, approaches:

1. *validation* studies, aiming at developing, elaborating, and validating theories about both the process of learning and the resulting implications for the design of learning environments;
2. *development* studies, aiming at the derivation of design principles for use in practice.

The study we describe falls into the second category.

4.2.3 The use of case studies

We piloted the intervention in successive stages. Each stage contained one or more cases. The evaluation of each case implicated an adaption to the prototype, yielding a next case.

Yin (2003) formulates three conditions to be met when applying a case study methodology:

1. The type of research question: typically to answer questions like 'how' or 'why'.

2. Extent of control over behavioural events: when the investigator has a little/no possibility to control the events.
3. General circumstances of the phenomenon to be studied: contemporary phenomenon in a real-life context.

We meet the first condition by the fact that our research question "what are the potentials of a classroom network in supporting teachers with providing feedback in statistics education?" can be paraphrased as "*how* can we utilise a classroom network in order to support teachers in statistics education?" The second condition is met because we lack complete control over the events we are studying: it is real life education, guided by its own complex rules. The third condition is obviously met: our intervention draws on innovative technology to be implemented in real life education.

Yin (2003) further distinguishes three types of case studies in terms of their outcomes: *exploratory* (as a pilot to other studies or research questions), *descriptive* (providing narrative accounts), and *explanatory* (testing). This classification of case study outcomes is consistent with the one Merriam (1988) proposed: *descriptive* (narrative accounts), *interpretative* (developing conceptual categories inductively in order to examine initial assumptions), and *evaluative* (explaining and judging). Our research used case studies gradually shifting from *exploratory* to *descriptive* to *interpretative*, in order to answer our research question.

From an interventional perspective, we strived to '*successive approximation*' of the '*ideal intervention*' (van den Akker, 1999, p. 2) and thus evaluated and improved (each prototype of) the intervention from case to case, until we reached our *best guess*.

A completely generalisable theory is not the goal of EDR, as is mentioned, amongst others, by Berkvens (2009), and is very hard, if not impossible, to reach by the use of case studies (Yin, 2003). By a "*thick description*" (Ryle, 1971) of the contexts of the case studies, we facilitate other researchers to translate our results to their specific research contexts.

4.2.4 Research components in this study

In figure 4.1 we depict our study in terms of the well-known ADDIE components (Gustafson & Branch, 2002; Molenda, 2003; Molenda, Pershing, & Reigeluth, 1996) Analysis, Design, Development, Implementation, and Evaluation. These components are more or less standard ingredients of any design oriented product. We stress that reflection, *in-action* as well as more retrospective *on action* (van den Akker & Kuiper, 2008) with respect to the quality in our view should complete each educational design research study. This results in the following representation of this study:

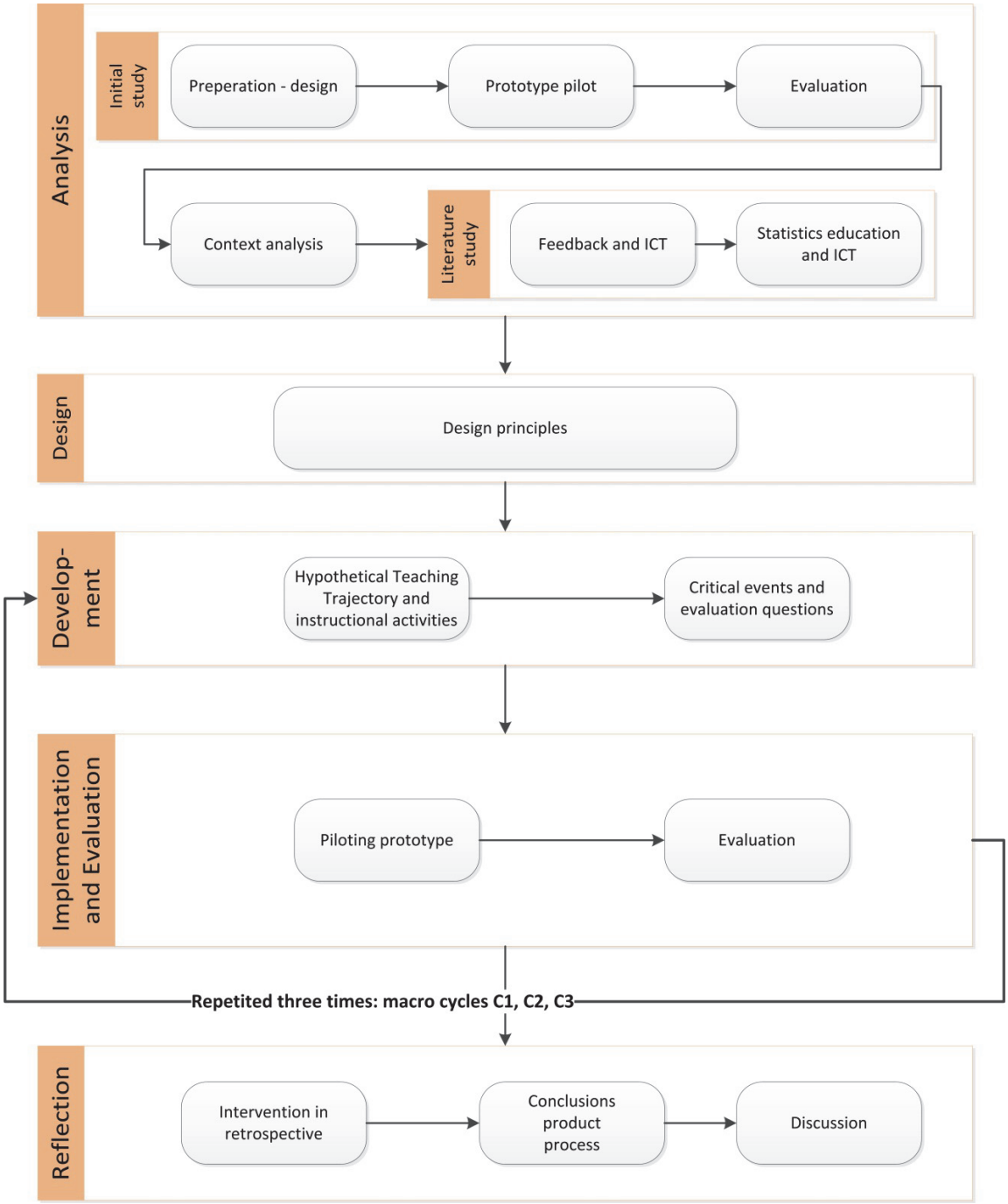


Figure 4.1 Research components of and stages in this study according to the ADDIE framework

In sections 4.3-4.6, we will describe each of the ADDIE stages this study contains in more detail. The macro cycles C1, C2 and C3 are the stages in which the three prototypes of the intervention are piloted.

4.3 Analysis

Initial study

We started this research with an *initial study* (see section 1.7) in one group with one teacher. We technically converted a chapter of the textbook used (about the normal distribution and its applications) into the teaching setting with a classroom network. The

teacher was trained in working with the technology. When observing the lessons we notified a remarkably intense classroom discourse centred around the teacher feedback on the students' work. After a questionnaire and interviews with three students and the teacher, we came to the conclusion that there was a possibility to strengthen feedback in statistics education when using a classroom network.

Context analysis

Having gathered this basic idea for further research, we conducted an analysis of the context in which we conducted this research. This context is at a national (Dutch) level and is politically sensitive, hence we also analysed official governmental publications like acts, newspapers and professional publications in order to get a complete contextual picture.

Literature study

In chapter 2, we described what the vast body of research on feedback says about those features of feedback we presume to be relevant for our study. During this part of the research, we had a specific eye for the possible role of ICT in the supply of feedback.

In chapter 3, we investigated the main goals of statistics education and how it is supposed to be designed, as reported in the research literature. We were particularly interested in the possible role of ICT during this study. We analysed two textbooks with respect to our learning domain. The starting point of this content analysis was formulated by the most important findings on what research says about statistics education, as considered in the research question. The conclusions of chapter 2 and 3 are used as input for design and development of the intervention, by formulating research based design choices and design principles, guiding the development of the intervention.

4.4 Design & development

4.4.1 Design principles

The results of the research on feedback and statistics education are used in chapter 5 to formulate design principles. The design principles are those recommendations by reported research that will guide the development of our intervention. These principles consist of product specifications and development guidelines. The design *principles* are the pivot between research questions and reported research on the one hand and the intervention design on the other.

4.4.2 Hypothetical teaching trajectory

In order to formulate our evaluation questions, investigating our hypotheses, from the perspective of our research question, we use the construct of a hypothetical learning trajectory (HLT). Basically, a hypothetical learning trajectory describes the aimed students' learning (realised curriculum). To illustrate the relationship between the intervention to be implemented, the HLT and the realised concrete instruction, we use the movies industry as a metaphor. The prototype can be interpreted as the screenplay upon which the movie is based. The movie itself represents the actual realised instruction in practice. We regard the HLT as the complete script for the movie: it is a director’s preview of how she sees what the actors have to say, have to do, where on stage, and how

this should look and sound. In this study, the researcher is the director and the teacher is the actor; as in producing a movie, the director uses more than the complete script in order to prepare the actors. Director and actor constantly discuss about how the script is to be translated into a performance before the cameras. We regard this discussion as a vital part of the *procedural specification* (Doyle & Ponder, 1977; McIntyre & Brown, 1979; van den Akker, 1988).

For a more precise description of the hypothetical learning trajectory we follow Simon's (1995, p. 136) approach, distinguishing three components of HLT:

1. the *learning goal* that defines the direction;
2. the *learning activities*;
3. the *hypothetical learning process*: a prediction of how the students' thinking and understanding will evolve in the context of the learning activities.

Simon (1995, p. 136) illustrates his implementation of the HLT as an instrument with the mathematics teaching cycle (abbreviated) as shown in figure 4.2. While in his scheme Simon consequently uses 'Teacher's', in this study, we use 'Researcher's' for indicating whose learning goal, whose learning activities and whose hypothesis of the learning process are to be taken into account. Executing the prototypes of the intervention with this HLT as a guideline, live, in the classroom, is of course a responsibility of the teacher. In addition, the almost daily fine-tuning of the details of the intervention is obviously discussed with the teacher.

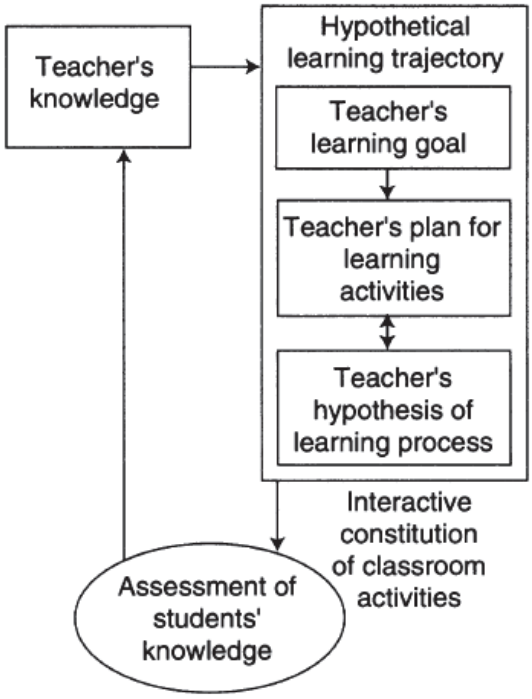


Figure 4.2 Mathematics teaching cycle after Simon (1995)

In this research, the HLT is regarded both as a research instrument (in order to make the educational expectations explicit and to prelude the formulation of the analysis question, with which the realised instruction is to be examined) as well as an instrument for the procedural specification, with which the participating teachers are prepared.

We mention here that for the HLT, we primarily focus on the teacher behaviour: when and how does she or he give feedback on which aspect of the learning process? We formulate our HLT therefore in terms of how students might answer the exercises and the way the teacher is expected to give feedback. This is different from focusing on the 'students' thinking and learning' in the first place. Considering the HLT as a research tool from the perspective of our main research question "How does the teacher utilise the possibilities of a classroom network during statistics education in order to enhance feedback?", primarily focussing on the teacher behaviour, it is a logical step to introduce an analogical concept 'hypothetical teaching trajectory' (HTT). We used an HTT to

'describe the ideal teacher acting', mainly with respect to the supply of feedback on the students' work. We consider an HTT to be an HLT with the focus on the teaching aspect.

When abstracting from the supporting teaching activities, the first prototype piloted (see chapter 5) and conducted (chapters 6 and 7) can be considered as built up in twelve episodes. An episode we see as a coherent teaching sequence, mainly consisting of exercises, together serving a specific learning goal.

Note that in Simon's scheme there is an important place for the 'Assessment of students' knowledge', which is, in a formative way, a main pillar in our study. For, as we stated in chapter 2, formative assessment is the framework we chose in order to give teacher feedback a central place in the classroom discourse. Simon suggested this assessment is to be proceeded by 'Interactive constitution of classroom activities', which is the key goal of working with a classroom network: using teacher feedback on students' work as a stepping stone to a more student-centred interactive classroom discourse.

With a formulation of HTTs for the subsequent episodes of our intervention, we can start to formulate the evaluation questions.

4.4.3 Materials and resources

The content analysis of the two predominantly used textbooks convinced us of the need to design our own teaching materials from scratch. With the selected design choices and principles in mind, we started the development of these materials, to be implemented in the prototype pilots, keeping the learning rationale – developing DL supported by a fluent ASS – constantly in mind.

A major guideline was to design the student activities as much as possible as exercises, in order to get the students active in statistics and to maximise chances for feedback.

The prototype, as was piloted during the macro cycles C1, C2 and C3 (see section 4.5), consists of teaching means with respect to descriptive statistics. Let us inspect these means from the perspective of a student; they will be described in more detail in chapter 5. The student activities mainly consist of exercises, aggregated in sections. Each section serves a specific statistical learning goal. The exercises are formatted in such a way they can be used in the wireless classroom network. In figure 4.3, we see the process of feedback depicted for an individual student, for an individual exercise.

The deliverance of exercises ('tasks'), from teacher to student and from student to teacher, is executed by the classroom network. First comparison of the performed task and the standard is rudimentarily done by software on the teacher's computer. With this information, the teacher can decide whether or not to start a feedback session with the group on this specific exercise.

Thus, there are two ways feedback on the students' performed task is delivered: immediate elementary feedback by the GC and delayed teacher feedback in the classroom. The immediate elementary feedback is programmed in the tasks that are delivered to the students, with the use of the classroom network. The teacher feedback is more sophisticated. The classroom network software analyses the tasks performed by the students and represents these results in an orderly way. The teacher inspects these results and decides whether or not to give plenary feedback on them. This feedback can be seen as the start of a classroom teaching dialogue.

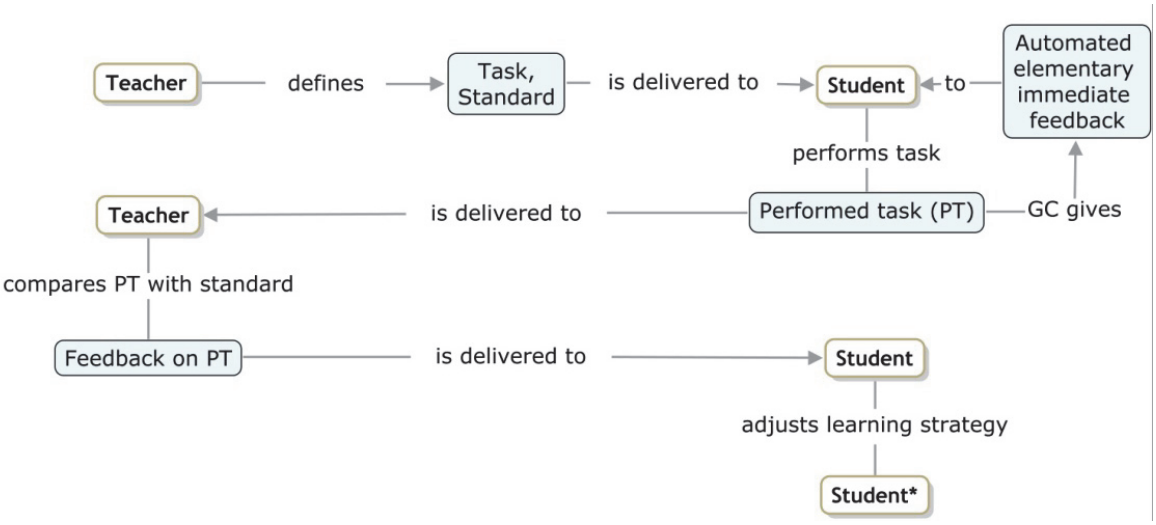


Figure 4.3 The process of feedback for an individual student for an individual exercise

An external mathematics and science education expert reviewed the developed student activities, as prepared for macro design cycles C1 and C2 (see section 4.5), with respect to the research goals. According to his comments, we adapted the activities. An external mathematics education expert reviewed the developed teaching activities, as prepared for C3, with respect to the research goals. According to his comments, the activities again were adapted.

4.4.4 Critical events and format for the evaluation questions

After having formulated for each episode a hypothetical teaching trajectory (HTT), we were ready to identify the intended 'critical events' to be emerging from the teaching practice we aim to establish. Each critical event will be accompanied by an evaluation question, aiming roughly at questioning whether the specific expectations as formulated in the HTT were met in classroom practice.

Using the adjective 'intended' in combination with 'critical events' we do not in fact follow the original description of 'critical event' as a construct that only can be called critical *after* it had happened (Kelchtermans, 1993). We are also perhaps less ambitious in our interpretation of the construct critical event than Woods (1993, p. 357) who states: "*They are flash-points that illuminate in an electrifying instant some key problematic aspect of the teacher's role and which contain, in the same instant, the solution."* Rather, we see critical events as those moments, as seen from the perspective of the research purpose, where a '*key step*' is to be set (*a priori* [without actual experience] critical event) or has been set (*a posteriori* [with empirical evidence] critical event). As a guideline for identifying a priori critical events, we use our hypothesis that teacher feedback will occur after each exercise that contains data literacy (DL) or a *new* algorithmic statistical skill (ASS). This is because DL and ASS are the discriminants of the student activities.

Then we had to formulate for each *a priori* critical event an evaluation question, investigating whether empirical evidence justifies this a priori critical event to be called an *a posterior critical event* after data analysis. In other words: Do the most important things the prototype aimed at really happen?

Thus, with the evaluation questions we wanted to evaluate the quality of the intervention. Nieveen (2009) distinguishes four quality aspects on which this evaluation can be based:

1. Relevance: what is the urge for the intervention?
2. Consistency: is the structure of the intervention logical and cohesive?
3. Practicality:
    a. Expected: do we expect the intervention to be usable for its goal?
    b. Actual: was the intervention really usable for its goal?
4. Effectiveness:
    a. Expected: do we expect that deploying the intervention is to result in desired outcomes?
    b. Actual: did the intervention yield the desired outcomes?

The relevance of the intervention was underpinned in chapter 1. The consistency of the intervention was described in Chapter 1 and evaluated in Chapter 6 and Chapter 7. The practicality was investigated by interviewing the teachers, as described in chapter 6 and 7. With the evaluation questions we want to explore the relevance, consistency and practicality of the prototype with respect to feedback on DL and ASS. Specifically, we want to explore:

- when and what kind of feedback does the teacher give, after having investigated the results of the students' work, when compared with what was intended;
- the efficacy of this teacher feedback in terms of the subsequent classroom discourse.

We thus have a 'two stage interest' in the feedback. Therefore, we use the following format for each evaluation question:

> (How) does the teacher use the classroom network regarding exercise[*x*] to give feedback[*x*] on the learning objective of exercise[*x*]?
>
> (How) does this feedback[*x*] prompt students to contribute to the classroom discourse around the learning objective of exercise[*x*]?

For example, when filling in the attributes in these two evaluation questions with respect to exercise 1.1 we come to:

> (How) does the teacher use the classroom network regarding exercise 1.1 to give feedback on the students' conceptions of the concept 'mean maximum temperature'?
>
> (How) does this feedback prompt students to contribute to the classroom discourse around the concept 'mean maximum temperature'?

For this intervention we thus see teacher feedback as a propelling force to an interactive classroom discourse.

We used the evaluation questions for the analysis of the data that we collected during the pilots C1, C2 and C3 of the prototypes of the intervention. We will describe results of this analysis in Chapter 6 and Chapter 7.

When did we intend plenary teacher feedback? The teacher could have different motives for giving feedback, after having analysed the students' work with ClassroomAnalysis. As a rule of thumb, we could state that we intended plenary feedback to be given by the teacher:

- with respect to ASS exercises: just after the introduction of a new skill or after the repetition of a complex skill;
- with respect to DL exercises: after all of these exercises.

In the HTT we made explicit, for the whole prototype, when we intended, based on this rule of thumb, teacher feedback for the whole class.

After having formulated the critical events and the format for the corresponding evaluation questions, we were ready to pilot our prototype in real life education.

4.5 Implementation

4.5.1 Terminology and implementation schedule

We now briefly describe the pilots of the prototypes we conducted in real life classrooms in order to collect empirical data for testing the conjectures in this study and to answer the evaluation questions. We understand by 'prototype pilots ' more or less the same as what other authors (Bakker, 2004; Doorman, 2005; Drijvers, 2003; Gravemeijer, 1998) call ‘teaching experiments’. These authors do not interpret the term ‘experiment’ in their studies as referring to a research design testing the effect of a certain treatment on an experimental group versus a control group. With the adjective ‘teaching’, a ‘teaching experiment’ has become a method in design research that has detached the interpretation of ‘experiment’ from its original methodological meaning (Steffe & Thompson, 2000). Rather, 'experiment' refers to the character that designing new teaching activities intrinsically has, to the actual deployment of these new activities in real life education and to the systematic evaluation of these activities. Nevertheless, in order to avoid the association with experiments in the classical sense, we use the term ‘prototype pilot’.

Regarding the pilots, in Table 4.1 we distinguish three research stages and their main organisational characteristics namely: what was implemented, when, with which teachers, in how many groups, with how many students and with how many observers in the classroom.

Table 4.1 Organisational characteristics of the three research stages.

What	When	Teacher_ID	#_groups	#_students	#_obs
C1	February 2006	A	1	31	3
C2	May 2006	B	1	25	3
C3	February-June 2010	C, D, E, F, G	6	128	1

The cycles C1, C2 and C3 were divided by a ‘begin to end’ redesign, based on the collected experiences. After C2, we also took data from interviews and questionnaires into account. Each cycle, of course, had its own subdivision in the three phases that are characteristic for design research: preparation/design, piloting and evaluation. We stress that in each cycle, in each phase, with each teacher, there was an intense communication about the learning goals, the way to achieve them, the materials, and the structure of the lessons and so on. Each individual lesson was discussed between teacher and researcher before implementation and was revised afterwards. It is therefore noteworthy to cite

Hattie (2009, p. 12): *...it dawned on me that the most important feature* [of feedback] *was the creation of situations in classrooms for the teachers to receive more feedback about their teaching.* The intense communication between the researcher and the teachers during the pilots can be regarded as feedback about their teaching.

The use of six cases (with five teachers) during C3 should be considered as a first up-scaling of the second prototype. After C2, we concluded to have addressed organisational and technical problems with the prototype and to be ready for comparing its pedagogical use with respect to feedback. Therefore, we needed several cases, while striving to "*successive approximation*" of the "*ideal intervention*" (van den Akker, 1999, p. 2). We evaluated and improved from case to case, on a not so fundamental level as we did between complete cycles.

4.5.2 Teacher characteristics and preparation

In this subsection we present characteristics of the two teachers participating in C1 and C2. In Chapter 7 we provide these characteristics before the evaluation of each successive case study.

The teacher in C1 was male, 34 years old, with 10 years of experience. The teacher of C2 was male, 52 years old, with 25 years of experience. These teachers were to be technically capable of performing ICT-based innovations in their classes. Besides this, we had an organisational condition: the intervention had to take place in the second semester of the school year 2005-2006. The intervention took place in two grade 10 groups of senior secondary education, both having about 25 students. Both teachers were enthusiastic about participating in this project.

To make sure that the teachers would be able to handle the classroom network, with respect to both the hardware and the software needed to utilise this and all the accompanying teaching activities, we organised a one day course three months before the start of the pilot. This may seem a long period, but both teachers wanted to be sure of having time to become familiar with the technology. They both participated in this training. For preparation at this training, we sent them manuals and some literature on the first findings on working with this technology in mathematics education.

After this training day, the teachers were given access to the hard- and software needed to build a classroom network to practise with. We visited the schools for a technical check-up. The prototype was discussed with them, first at a general level, then into the details per section. The moments of feedback were presented and discussed as 'suggestions' and not as 'mandatory' elements. It was crucial in this research phase to investigate in which situations meaningful, rich classroom discourse emerged from the feedback. Therefore, we did not want to oblige the teachers to follow a too narrow path, but wanted to give them, as much as possible, the same freedom in the organisation of the classroom discourse as they are used to. After these visits, the teachers seemed prepared well enough to step into the real classes, in which they have the possibility to give feedback while using the classroom network.

4.5.3 Cycle C1: pilot of the first prototype

The pilot of this first prototype was conducted in February 2006. The target group was a 10^{th} grade senior secondary education class. The topic was descriptive statistics: measures

of central tendency, measures of spread, and the graphical representations boxplot and histogram (frequency diagram). The prototype was evaluated with respect to the extent it achieved the goal of 'a smooth educational process' in statistics: was the technological infrastructure reliable and did it result in meaningful statistics education? Success of the intervention with respect to higher, domain specific pedagogical goals was to be investigated in successive macro cycles.

However, because it was possible the prototype worked out exactly the way we intended (as formulated in our HTT), we observed the lessons with a strong delegation: two research assistants and the principal researcher attended all lessons. One research assistant was responsible for the camera with transmitter microphone that focused on the teacher, the other research assistant operated the camera that was to record the students. The principal researcher attended the lessons with an observation form that was derived from the form used during the preliminary study, guided by the HTT. The three observers shared their personal interpretations in order to get a more reliable 'first impression'.

We used a questionnaire before intervening in order to determine the attitude of the students with respect to mathematics and how they perceived the feedback they got from the teacher. After the prototype pilot, we used the same questionnaire in order to determine whether the students considered working with a classroom network as a way to improve receiving feedback on their learning of statistics. After each lesson, we had a short evaluation with the teacher around the core topic of feedback. We asked the teacher to select three students: one with weak, one with average and one with good competence with respect to mathematics. With these students, we organised individual interviews in order to discuss the results from the post questionnaire and their personal opinion about a classroom network in statistics education. Finally, we interviewed the teacher about his experiences with respect to the feedback possibilities of a classroom network in statistics education.

We abandoned the investigation of learning gains by using a pre-test/post-test quasi-experimental/control group approach. This was because of our belief that the relevance, consistency and usefulness of our intervention should be investigated before its contribution to possible learning gains is studied.

4.5.4 Cycle C2: pilot of the second prototype

The pilot of the second prototype was carried out in May 2006. The target group was again a 10th grade Senior Secondary Education class. The recommendations from the first prototype pilot, as formulated in section 6.2.2 , were implemented as far as possible into the new prototype. Roughly, we formulated three different types of recommendations:

1. technical-organisational;
2. instruction of the teacher;
3. teaching activities.

Although it may seem obvious, from the perspective of the research question, we were primarily interested in recommendation of the second and third type. We nevertheless realised that those of the first type were very important too. By implementing them correctly, we tried to optimise chances to establish a classroom setting and climate that can foster feedback.

Data analysis was carried out in approximately the same way as during the next macro cycle C3. These methods will be addressed in section 4.5.6. This resulted in a couple of recommendations for adaption of the prototype, in order to conduct the pilot of the third and final prototype.

4.5.5 Cycle C3: pilot of the third prototype

The pilot of the third and final prototype was carried out during the period January-March 2010. The target group was again a 10th grade Senior Secondary Education class. The recommendations from the second prototype pilot, as described in chapter 6, were implemented as far as possible into the intervention. The prototype was reviewed, with the new state of technology in mind. Technology drastically changed in the time between cycle C2 and cycle C3 of our prototype pilots, partly because of our recommendations to the manufacturer of the handhelds and the classroom network, based on our experiences in cycles C1 and C2.

There were five teachers, from five different schools, participating with six groups in this cycle of prototype pilots. We scheduled them successively, in order to optimise chances for "*successive approximation*" of our "*ideal intervention*" (van den Akker, 1999, p. 2). We make use of selected video-captured 'critical events' in order to provide subsequent teachers with specific information, based on real life education. We tried to arrange school visits for this purpose too, but besides one exception, the timetables of the participating teachers, all from different school organisations, were incompatible. We used online video in order to facilitate communication about the experiences in the classrooms. In this way, teachers were able to have a look into each other's classrooms, without actual on site visits. We tried to organise these visits as well, but the traveling distance and the teachers' incompatible timetables made this, with a few exceptions, impossible. Thus, we tried to make the teachers even more 'problem owner' of the project goals.

We tried to integrate experiences from each case into the subsequent one, while keeping a keen eye for each unique context. Basically, we used in cycle C3 the same research instruments as in the cycles before: videotaped observation, questionnaires, individual teacher and student interviews, the students' answers on the exercises and the results on the summative test. Nevertheless, because we were convinced that each pilot of the prototype has a context that was unique, we added an extra instrument, in order to be able to compare the experiences of the participating teachers with respect to the intervention goals. After a first analysis of all of the individual teacher interviews, we designed another questionnaire to be completed by the teachers. Analysis of the collected questionnaires was the input for a final group interview, aiming for an overarching consensus, if possible (Bogdan & Biklen, 1992).

4.5.6 Evaluation

A characteristic of design research is that evaluation in the sense of 'reflecting and trying to understand what happened' is a continuous ingredient. After each lesson, in the pilot of each prototype, in each macro cycle, the researcher reflected on the possible gap between expectations and reality, individually and in consultation with the teacher. This sometimes resulted in minor changes in the prototype, or in a slightly different articulation of it. This redesign is called a *micro cycle* (Gravemeijer & Cobb, 2006). In C3, we found reason to adapt the prototype in a more profound way twice. Besides this,

we evaluated each macro cycle rigorously, leading to adaptions of the prototype. The most thorough evaluation was the one concerning the third prototype during C3, which will be described in Chapter 7. After having collected all the results of C1, C2 and C3, with an emphasis on the results of the last, we tried to come to a plausible answer to our research question "What are the potentials of a classroom network for the support of feedback in statistics education?" In order to do so, we carried out a cross-macro cycle and a cross-case (C3) analysis, from the viewpoint of all analysed data. The results had to be interpreted within the context of each case in itself and from the viewpoint of successive cases.

First, the data had to be analysed. How did we do that?

In Figure 4.4 we depict the relationship between the data sources.

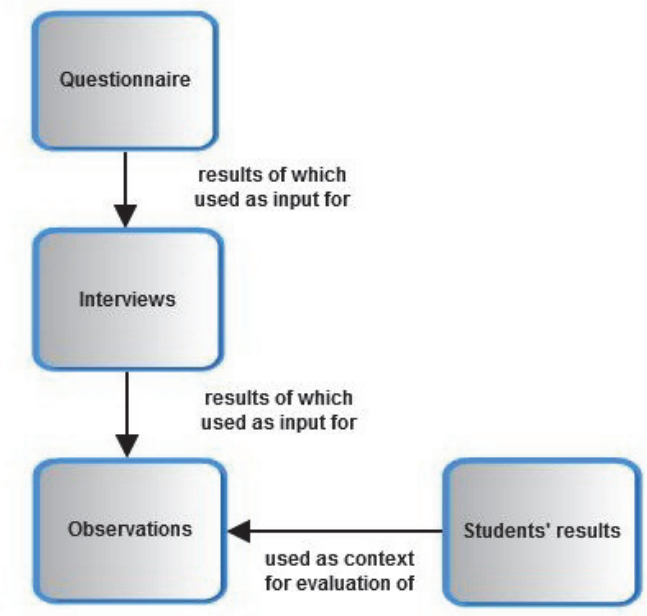


Figure 4.4 The relationship between the data sources

The primary data source was the observation of the lessons conducted during the pilots of the three prototypes. The interviews with teachers and students and the questionnaires completed by the students were used as input for the analysis of the observations. The students' answers on the exercises and the results on the summative test were used as a context for the evaluation of the observations.

Evaluation of the observed lessons was an iterative process itself. First, we inspected the video-taped lessons integrally. Then we made a selection of the fragments with respect to relevance to the research question. We selected those fragments corresponding with the a priori (hypothesised) critical events that gave rise to the evaluation questions as described in section 4.4.4. Then we had a first walkthrough of the results, asking ourselves two questions:

1. Does it give rise to a change of the coding scheme that we used in the analysis of the results of the preceding macro cycle? In a constant comparative approach (Glaser & Strauss, 1967), iterative in its nature like a lot of methods used in EDR, we tried to optimise the coding scheme.
2. Do these fragments have enough persuasiveness, positive or negative, with respect to our evaluation question? If not, we noted why. If so, we transcribed the fragment into text. In the evaluation of the C2 prototype pilot, we transcribed all of the lessons completely. However, this was time-consuming (about two and a half months full time), so we decided to do a relevant selection first in C3.

In the analysis of the C2 and C3 data, we used specific software for qualitative analysis for video and audio data.

We started coding the fragments with the scheme based on the corresponding evaluation questions. The coding was partially done by a colleague researcher as well. We informed

him about the intentions of our coding scheme, and about the general perspective of our research question. After this, we discussed our collective results, in order to retrieve a more objective interpretation of the coding scheme. With the adapted scheme, we recoded the fragments. It took several 'fresh starts' of the coding process before we were able to perform it consistently. After the coding, we aggregated the results on two levels: that of a single case study and that of all the case studies together. Then we tried to determine which patterns there were at both levels. Then we tried to find a plausible explanation of these patterns. Finding a plausible explanation was supported by the analysis of the data that were collected using the other primary research instruments: the interviews and the questionnaires.

In C2, we used pre and post student questionnaires, with mainly Likert-scale items about the students' perception of feedback. Thus, we tried to identify significant changes in the students' perception on feedback, induced through the way of working during the intervention. Because in the school organisation and the resulting difficulty of non-response items, especially when using the school's web based learning environment for the questionnaires, when the actual lessons had not started, we decided to use in C3 just one questionnaire. The items were put in a comparative way. For example: “Did you receive, when working the way we did the last three weeks, more feedback on your work than before this experiment?”. We handed the questionnaires out on paper at the end of the last lesson of each episode of the prototype pilot. We asked the students to fill it in and to hand it in. We entered the data into a computer program for quantitative data analysis and calculated whether the students had perceived any effect from the intervention.

Which interviews were conducted during C2 and C3? Each participating teacher was interviewed at the end of each case (i.e. an episode of a pilot) using an open semi-structured interview scheme (Bogdan & Biklen, 1992). The aim of the teacher interviews was twofold, namely recording:

1. how the teacher perceived his or her feedback possibilities;
2. how he or she perceived the teaching context.

The same approach was used when interviewing from each participating class three students, appointed by the teacher, together representing the students with respect to mathematical skills: a weak student, an average student and a good student. In C1 and C2, the students were interviewed face to face. In C3, because of the reasonable number of interviewees (18) in combination with the travelling time (approximately 45 minutes on average), we chose telephone interviews. An explicit input into the student interviews were the results of the analysis of the questionnaires: first, of those in the group of the interviewed student and second, of all of the students that had at the moment of interviewing participated in that stage of the study. The main goal of the student interviews was to give context to the Likert-scale based opinions about the perceived feedback by the whole group. All of the interviews were audiotaped. We selected and transcribed all of the fragments pointing at a positive or negative characteristic of the feedback as perceived by the students.

In the evaluation of C2 and C3, the last data source consisted of the students' answers on the exercises and the results on the summative test. We developed this summative test, with open exercises, according to the learning goals: through algorithmic statistical skills to data literacy. We calculated the mean scores on the exercises and the test. We discussed the results with the responsible teacher, asking “Are the achievements

satisfactory and how do they relate to the feedback-based classroom discourse?" We used the students' performances with respect to the exercises and the test as an important part of the context of interpreting the results of our primary data sources.

4.6 Validity, reliability and participant anonymity

4.6.1 Terminology

In this study, the primary data sources were observations (from the realised statistics lessons), student-questionnaires, and interviews (with students and teachers). The plausibility of the interpretation of the data from these sources determines the persuasiveness of this research. We now shortly describe how we tried to promote validity and reliability of our methods and instruments and how we secured the confidentiality of the participants of the empirical parts of this study.

In EDR, validity and reliability are in general issues that should be taken particular care of (Lincoln & Guba, 1985). Guba (1981) formulates some criteria with respect to the trustworthiness of naturalistic inquiries, like the one reported here. He therefore more or less translates dimensions as used in rationalistic (quantitative) research into appropriate terms in naturalistic inquiry (qualitative research). He thus comes to criteria for judging qualitative studies:

- Internal validity: Credibility
- External validity: Transferability
- Internal reliability: Dependability
- External reliability: Confirmability

We consider these translations to be meaningful. For instance, the 'confirmability' of a study expresses the likeliness that when the intervention is repeated by other researchers under conditions as described by the one that is to be replicated, evaluations of this repeated intervention will be similar as those of the original study. That is, the possibility to confirm. We consider the Guba terminology to be better understandable in educational design research contexts than the traditional terminology. Nevertheless, we note that in the reported EDR, starting roughly with Brown's (1992) autobiographical essay on her transformation from a traditional quantitative oriented cognitive psychologist into a more qualitative oriented design researcher, the traditional terms of validity and reliability are still used. We therefore conform to these traditional terms.

4.6.2 Validity

Internal validity tries to underpin the reasoning within a study with the actual gathered data. The internal validity of this study has been addressed by data triangulation, i.e. using multiple measures to investigate an event (Cohen, Manion, & Morrison, 2007; Denzin, 2006; Yin, 2003). Through conducting three cycles of the prototype pilots (C1, C2 and C3), during a period of six years, with the cooperation of six teachers, in a continuous process of briefing and debriefing, it was possible to test, refine, and retest our conjectures. We have deployed reasoning that has been continuously exchanged with colleagues, teachers and researchers. Thus, we tried to maximise the chance that the gathered data really underlie the phenomenon to be explained: teacher feedback in statistics education supported by a classroom network.

The issue of *external validity* in qualitative research like this study is mostly a concern of the generalisability of the research findings. We tried to secure this by precisely

describing the dependency of our conclusions with respect to the research context. Attributes of teachers, students, disciplinary content, and technology were taken into account when interpreting our data. These can be reused when extrapolating our findings to other educational contexts (Barab & Kirshner, 2001). In our view, *ecological validity* (Bronfenbrenner, 1976) is an appearance of external validity that focuses on the transferability of the research setting to 'everyday classroom practice'. This transferability is one of the goals of EDR, being conducted in real classrooms, with real teachers and real students, allowing the researchers to use rich descriptions of the context (Barab, Baek, Schatz, Scheckler, & Moore, 2008) so that other researchers can use our results in other real life education.

4.6.3 Reliability

Internal reliability refers to the reliability of the methods used in a study. During this research, we chose different methods for the different prototype pilots. In our initial study (Tolboom, 2005), each lesson was observed by two researchers, using the same observation scheme, aiming to detect what the main added value was of the classroom network for the classroom discourse. Both researchers exchanged their completed observation schemes, discussed each other's findings and came to an agreed global conclusion. Besides this observation form, during C1 and C2 (the first and second prototype pilots) we had two cameras to record each lesson. The principal researcher carried out the transcription of the videos. The coding was discussed with the two research assistants who did the recordings and thus were able to observe the lessons too. During C3, we did not transcribe and code all of the recorded material, but restricted ourselves to those parts relevant for our evaluation questions. A colleague researcher carried out a part of the scoring of the results. The third prototype pilot (in C3) was carried out with five teachers in six teaching contexts (one teacher taught two groups), making the results more stable and, therefore, more reliable (Denzin & Lincoln, 1994). We tried as much as possible to make our coding based on simple and neutral acts, such as counting the number of times a certain well-defined behaviour (this being well-defined actually belongs to the internal validity of the study) occurs.

External reliability expresses the concern that it should be possible for the reported research to be reconstructed by other researchers. Gravemeijer and Cobb (2006) cite Smaling (1990, p. 6) in connecting external reliability with what they call "*trackability with virtual replicability*". Smaling states that trackability could be established by reporting on "*failures and successes, on the procedures followed, on the conceptual framework and on the reasons for the choices made*". In this study, we try to meet these criteria by being explicit in making choices, illustrated with our reasoning and by showing, systematically, how our insights developed during the research trajectory. Through this process we try to come to a "*thick description*" (Ryle, 1971) with which we hope to approximate the classical description of design (or developmental, which we consider, in this context, to be equivalent) research by Freudenthal (1991, p. 161): "*Developmental research means 'experiencing the cyclic process of development and research so consciously, and reporting on it so candidly that it justifies itself, and this experience can be transmitted to others to become like their own experience."*

4.6.4 Minimising main potential biases

Role bias

While conducting EDR, there is a potential bias when the roles of developer and researcher get entangled. In the empirical part of the study, the role of researcher has to dominate. Standing behind the video recorder in the classroom, meanwhile observing the practice of the teacher of the prototype, is sometimes hard because at those moments interference is not allowed. We nevertheless tried to be as neutral as possible in the name of 'the developing stage has finished, research stage has begun'. By using standardised procedures and instruments we restricted the developer's interference with the researcher during strictly research activities.

Response bias

When interviewing teachers and students, thus collecting reflective data, it is always hard for the interviewees not behave in socially desirable ways. We explicitly stated, at the start of each interview, that only the real opinion of the interviewee, however negative his or her experiences might have been during the intervention, would help the study. We asked the interviewee consequently whether he or she was aware of this fact.

Selection bias

When conducting EDR with an empirical component, as is the case in this study, a selection of schools and of teachers has to be made. For this selection, we made use of *convenient sampling*: we had a couple of practical criteria, like geographical accessibility (the school being within an hour's drive from the residence of the researcher and willing to cooperate with this research). Eight teachers were selected from the personal network of the researcher, mainly on the basis of whether they were teaching the topic of descriptive statistics to the target group of our study. There was just one teacher who withdrew from the project before the start, because he was not able to invest the energy needed during the period we planned to pilot the prototype in his lessons. During cycles C1 and C2 we selected ICT committed teachers. We dropped this criterion during cycle C3, because we hypothesised the prototype was that robust that even teachers not as committed to ICT would be able to work with it without problems.

4.6.5 Anonymity of participants and materials used

Due to the use of tape recorders to record the interviews and video recorders to record the lessons in which we piloted our prototype, we had to deal with maintaining confidentiality and the anonymity of the participants: teachers and students. We discussed this with them, assuring that we would only publish parts of the data (sound and video) in strict scientific settings, such as presentations at scientific conferences. In addition we told them that we discussed the data with colleagues, using online video, but that we would use the 'private mode' of this video, so that it could only be watched by invitees, and the video could not be approached from the public World Wide Web.

On the other hand, we chose for openness about the choice of the materials used: the network infrastructure from Texas Instruments (Navigator) and the chosen handhelds (TI 84 (intitial study, C1, C2) and TI Nspire (C3), the textbooks we analysed (Moderne wiskunde (Modern math) by Noordhoff publishers and Getal & Ruimte (Number & Space) by EPN). Of course, in our conclusions we abstract from concrete materials used and formulate in terms of functionality.

Chapter 5
Prototype design and development

In this chapter we describe the design and development of the first prototype we used in this intervention study. Adaptions to this prototype, based on the evaluation of C1 and C2, will be described in chapters 6 and 7. In this chapter we start by translating the recommendations from research on feedback and ICT (chapter 2) and statistics education (chapter 3) into design principles. In order to get an overview of the intervention we developed, we then describe this intervention to be designed, developed and piloted from a curriculum perspective. We provide typical examples of specific feedback types for concrete exercises as an operationalisation of the feedback principles. After that, we present an overview of the teaching materials, consisting of twelve thematic units. Finally, we formulate the tenets of hypothetical teaching trajectories and corresponding evaluation questions. In chapters 6 and 7, these questions will be used in order to compare the realised feedback during classroom practice with the planned feedback as formulated in the hypothetical teaching trajectories.

5.1 Design principles

In this section we formulate the design principles with respect to feedback, statistics education, and ICT. These principles consist of product specifications and development guidelines. We discuss their role and formulate the feedback matrix with respect to statistics education.

5.1.1 Role of the design principles

The main research question for this study is:

> "*What are the potentials of a classroom network in supporting teachers with providing feedback in statistics education?*"

As regards feedback we are interested in the question:

> "*Can feedback be a pivot to an interactive classroom discourse?*"

While keeping these questions in mind, the function of the intervention was to provide a fruitful environment for the instruction of descriptive statistics that fosters both teaching 'data literacy' as well 'algorithmic statistical skills', while utilising the feedback possibilities of an ICT infrastructure, meant to facilitate interactive classroom discourse fostering reflection.

In chapters 2 and 3, we described what research says about the core topics of this study: feedback and statistics education combined with ICT. Using these findings as an input, in the next section we formulate the *design principles* that guided the intervention with respect to function, form and content. Following this, we then chose those principles that we explicitly used for the evaluation of the prototype having brought in classroom practice.

We have chosen to incorporate the design principles with respect to ICT into those with respect to feedback and statistics education, because we consider the use of ICT in mathematics education to be serving, not leading.

5.1.2 Design principles with respect to feedback

With respect to the formulation of the design principles, our format has been inspired by the one proposed by van den Akker (1999) and is formulated as:

If you want to <*goal formulation*> you are best advised to <*procedure formulation*> because of <*argument*>.

With respect to feedback we formulate the following design principles:

F1. If you want to provide feedback to students on a task that mainly concerns *procedural and/or declarative knowledge*, you are best advised to provide immediate feedback supplied by a computer, because this improves the efficacy of the feedback (Azevedo & Bernard, 1995; Corbett & Anderson, 2001; Kulik & Kulik, 1988).

F2. If you want to provide feedback to students on a task that mainly concerns *conceptual* knowledge, you are best advised to provide delayed feedback by the teacher after one or some more days because this improves the efficacy of the feedback (Butler, et al., 2007; Schroth, 1992).

F3. If you want to address feedback effectively, you are best advised to address feedback to the task, not to the learners' self, as was distilled from a meta-analysis on feedback efficacy by Kluger and DeNisi (1996).

F4. If you want to optimise learning gains from feedback, you are best advised to consider the formulation of feedback to be specific but neither too directive, nor formulated in a complex way (Shute, 2008).

F5. If you want to improve students' learning from the supplied feedback, you are best advised to provide information on the 'why' and not just 'right' or 'wrong', but keep the elaboration of the feedback relatively simple (Shute, 2008).

F6. If you want to improve the acceptance of the provided feedback, you may supply the feedback at the students' GC, because most students prefer computers above humans as a feedback source (Karabenick & Knapp, 1988; Kluger & Adler, 1993; Yarnall, et al., 2006).

F7. If you want to provide feedback to students efficiently, you are best advised to automate the supply of feedback by using a computer network, because this offers advantages of scale: it takes as much effort to provide feedback to one student as to a whole class (economies of scale when using ICT).

When aiming at a focus in research, one has to make choices. In this study, we followed all the above design principles during the design and development phases. However, in the systematic evaluation of the intervention we concentrated on the first two principles. The main reason for this is that they can be combined very well with the specific goals of statistics education, making them most interesting for our study. The other design principles can be implemented in a relatively straight forward manner. Further, when observations are the main source of data, extensive evaluation criteria cannot be handled at the same time.

5.1.3 Design principles with respect to statistics education

With respect to statistics education we formulate the following design principles:

SE1. If you want to improve the students' conceptual knowledge of statistics, you are best advised to include exercises that aim at developing 'Data literacy' by systematically focussing on the students' understanding of the purpose and logic of statistical investigations and their development of reflective and interpretive skills, because this approach appeals more to students having a non-statistical focus (Cobb, 1991; Roback, 2003; Wiberg, 2009).

SE2. If you want to improve the students' declarative and procedural knowledge of statistics, you are best advised to include exercises that aim at developing 'Algorithmic statistical skills' by systematically focussing on the students' skills with respect to practical, well defined statistical procedures, often with a quantitative goal or nature because this improves the efficacy of the feedback (Azevedo & Bernard, 1995; Corbett & Anderson, 2001; Kulik & Kulik, 1988).

SE3. If you want to promote student engagement in statistics education, you are best advised to design the teaching activities as much as possible as exercises by transforming every teaching element into a question or end every element with a question so students are forced to be active and the teacher has the opportunity to supply feedback.

SE4. If you want to design appealing statistics education, you are best advised to build up instruction inductively from 'realistic situations' to 'theory' as is proposed by Treffers (1987) as 'progressive mathematization'.

SE5. If you want to build up statistics education inductively, you are best advised to use realistic situations starting with phenomenological rich contexts making the central statistical question arise naturally.

SE6. If you want to utilise contexts as a starting point for statistical inquiry, you are best advised to use *authentic* contexts which are real and attractive for students, and inviting to meaningful statistical operations, adopted from an advise by Wijers, Jonker and Kemme (2004).

SE7. If you want to give data a central place in statistics education, you are best advised to use *real* data sets (see 3) and use ICT to perform statistical operations making use of the students' GC, because real data sets tend to be that big that calculations by hand take too much time.

SE8. If you want to give data a central place in statistics education, you are best advised to use visual illustrations and emphasise exploratory data methods making use of the GC, so the students can improve their imagination for data.

SE9. If you want students to communicate statistically, you are best advised to deploy a classroom culture around discussing statistics in which teacher feedback and computer generated feedback are key elements, because this guarantees students' commitment (while their work is in the centre) and their voices are heard.

In developing our intervention, we used all of the above design principles. For a systematic evaluation, we concentrated on the first two. The main reason here was that we see a consistent, most synergetic combination between these two and the two selected

from the feedback principles, in relation with the specific possibilities of our chosen ICT environment (see section 1.3).

5.1.4 Feedback matrix for statistics education

Although research is not completely consistent about *timing,* we nevertheless consider it to be an essential feedback feature. With respect to this variable, we have chosen *two specific types of feedback* to be used in the intervention:

1. *Immediate* feedback, to be displayed on the display of the students' graphing calculators.
2. *Delayed* feedback, delivered by the teacher after an automated analysis of students' work, aiming at a classroom discourse fostering reflection. This can be in the same lesson, but also in the following lesson.

In section 5.3.1 we visualise the process of feedback as implemented in our intervention.

In chapter 3 we distinguished two domains of students' activities in statistics education: d*ata literacy* (DL) and *algorithmic statistical skills* (ASS). Both categories differ in nature. DL is more conceptual and thus (because of feedback design principle 2) could profit more from delayed feedback. Delayed feedback in our intervention is provided by the teacher. Delayed means in this study: in the same lesson after about half an hour, but usually during the next lesson. Reflection on the concepts can thus be incorporated more easily. ASS is more procedural and this requires immediate feedback. Immediate feedback, to the whole class, can most efficiently be generated using ICT.

Nevertheless, we sometimes specify in our hypothetical teaching trajectory (HTT) delayed teacher feedback on ASS exercises, especially when the exercises pertain to a skill that is new and thus has not been performed before. On the other hand, we design immediate feedback on DL exercises in case of a 'self-section': a section of exercises where the student can check the right answer themselves on their GC (see section 7.2).

We now combine the two distinctions, 'immediate' versus 'delayed' with respect to feedback, and 'data literacy' versus 'algorithmic statistical skills' with respect to the nature of the learning goal. This defines the feedback spectrum addressed by the prototype as follows in a *feedback matrix for statistics education*:

Table 5.1 Four feedback types in the feedback matrix for statistics education

Timing of feedback / Learning goal	Immediate	Delayed
Data literacy	Type I	Type II
Algorithmic statistical skills	Type III	Type IV

In section 5.3 we present examples of these four feedback types.

5.2 The intervention from a curriculum perspective

In order to provide a broad overview of the intervention we developed, we describe the main aspects of the intervention from a curriculum perspective. Interventions aim to

investigate a certain curricular aspect and, if possible, improve this aspect. An intervention, from a curriculum perspective, should therefore be well outlined. In order to do so, we use van den Akker's (2003) curricular spider web. This offers us a good framework of curricular components and their mutual relationships.

In defining the curricular *level* of this intervention we follow van den Akker (2003) in posing that we intervene with respect to teacher behaviour (stronger emphasis on feedback), and on the students' activities (more DL), thus mainly at classroom (*micro*) level. We expect that at this level there will be effects, as well as at the student (*nano*) level, as each individual student had to learn in the new situation with the new materials. The teachers involved had to adapt their didactics to the new possibilities. Our intervention, and the accompanying evaluative instruments, thus mixed micro- and nano-level: we studied the teacher behaviour, the students and their interaction.

In curriculum theory it is quite common use to observe different curricular levels and perspectives (Goodlad, 1979). Van den Akker (2003) distinguishes some *representational forms* of the curriculum. In Table 5.2 we identify those representations, specifications and corresponding actors that are important for our intervention.

Table 5.2 Curricular representations and specifications with corresponding main actors

Curricular representation	**Specification**	**Main actor(s)**
The *intended* intervention	*ideal* and *written*	Designer - researcher
The *implemented* intervention	*perceived* and *operational*	Teacher
The *attained* intervention	*experiential* and *learned*	Students

In addition, we use the distinction of ten curriculum *components* van den Akker (2003) suggests in his curricular spider web approach, to show that our intervention, although primarily aiming at teacher behaviour and student activities, covers many curricular components. We describe the intervention in terms of: **curriculum component** with *corresponding curricular question* and the way this component was designed in our intervention, compared with the usual textbook setting and non-feedback focussed teacher behaviour during the lessons. Between brackets we denote the corresponding design principle. For example, F7 stands for the seventh design principle with respect to feedback, SE9 for the ninth design principle with respect to statistics education.

1. **Rationale**: *Why are they learning?*

   The rationale is not influenced by the intervention.

2. **Aims & Objectives**: *Toward which goals are they learning?*

   The goal as far as statistics education is concerned is tuned towards better student skills in data literacy (SE1), while still developing algorithmic statistical skills (SE2). However, as often with ICT mediated interventions, there is an additional goal: a better preparation for a digitalised society. This additional goal, in this study, is nevertheless not problematised structurally.

3. **Content**: *What are they learning?*

   There is a stronger emphasis in the prototype on DL (SE1) and on authentic contexts (SE6) than in the current textbooks. Form these contexts the central statistical question arises naturally (SE5). The contexts generate real data sets to

be explored (SE7). There is more emphasis on exploration and visualisation of the data (S8).

4. **Learning activities**: *How are they learning?*

   The content has been designed with 'questionising' as a leading principle (SE3), meaning that that almost all of the teaching activities consist of exercises to be completed by the students. This content is built up along the principle of progressive mathematisation (SE4).

5. **Teacher role**: *How is the teacher facilitating learning?*

   The teacher manages the sending-receiving process from the assignments files. He uses the classroom network to digitally analyse the collective classroom performance and gives specific, not too directive, feedback, not too complexly formulated (F4) on the topics if the analysis in his view appears to be 'feedback worthy' (F2). He does so focussing on the 'why' (F5) in a task addressing way (F3).

6. **Materials & Resources**: *With what are they learning?*

   All student activities are designed from scratch. The students are still using a graphing calculator but use this GC in a more crucial way: they do their exercises with the GC, they receive elementary feedback on their GC and the teacher uses their GC to collect their work in order to start a possible feedback session. A hardcopy version of the assignments is supplied to the students. In addition, a hardcopy study manual is provided. The digital material, supplied by the classroom network (F7), provides immediate feedback (F6) on the students' activities (F1) in a task addressing way (F3).

7. **Grouping**: *With whom are they learning?*

   Students learn more from each other through a bigger input in the classroom discourse and a stronger communication on statistics (S9).

8. **Location**: *Where are they learning?*

   This is more variable than before the intervention. The students just need their graphing calculator to do their exercises. This makes their learning more place-independent.

9. **Time**: *When are they learning?*

   This is more variable too than in a more traditional learning environment because the graphing calculator is a real mobile device.

10. **Assessment**: *How far has learning progressed?*

    Through the immediate feedback from their graphing calculators, and the teacher feedback in the classroom, students are more aware of their progress. The teacher has a lot more data of students' progress at hand.

Van den Akker (2003) advises to develop all these curricular dimensions in coherence with each other, keeping the rationale constantly in mind. Besides that, he explicitly warns for a too narrow curricular approach of interventions based on ICT "*A striking example is the trend toward integration of ICT in the curriculum, with usually initial attention to changes in* [the component] *materials and resources. Many implementation studies have exemplified the need for a more comprehensive approach and systematic attention to the other components before one can expect robust changes."* (2003, p. 5).

By preparing changes for this intervention in eight of the ten components, we consider it from a curricular perspective fairly robust.

We briefly summarise the three stages the curricular potential of ICT in education goes through (Itzkan, 1994; Voogt, 2003):

1. *Substitution* stage: ICT is solely being used as a replacement for the teacher.
2. *Transition* stage: ICT also requires that educational practices and content are changed as well. ICT applications not only structure the learning process, but students themselves structure their own learning process.
3. *Transformation* stage: ICT also requires the change of rationale of education.

Interesting here is that in stage *x* the characteristics of stage *x*-1 are not per se vanished. When ranking our intervention in this model, stage 2 of transition comes most close. Nevertheless, the intervention also has a substitution character: the assignments formerly stated in the textbooks are now embedded in the students' graphing calculators. However, the transition in teaching practice is the most dominating characteristic of this intervention. Teacher and student come pedagogically and didactically into a different relationship. The rationale of the educational process remains unchanged, so transformation is one stage too far.

When we try to characterise the pedagogical impact of our intervention, we refer to Voogt's (2003) overview of pedagogy in the industrial versus the information society. In our case, we concentrate specifically on:

1. *Supporting each other*: when students experience the crucial role of their handheld device, they are likely to act like a professional community.
2. *Productive learning*: the authentic assignments they have to make and the double feedback they receive will give the students a sense of being productive themselves.
3. *Integrating theory and practice*: the questionising principle challenges the student to be an actor in their own learning process. Therefore, we even try to design theoretical elements as a question for the student.
4. *Diagnostic*: the very core of the intervention. Through the feedback of the handheld, a student knows where he stands after having completed a specific assignment. The centralised feedback by the teacher after having collected all students' work gives both teacher and students a chance for updating their 'how do we do' knowledge. Moreover, the classroom network offers the teacher the opportunity to study individual and collective progress.

From the list of ICT capabilities for enhancing learning Dede (2000, p. 282; Cognition and Technology Group at Vanderbilt, 1997) reports, we specifically highlight "*centring the curriculum on 'authentic' problems parallel to those adults face in real world settings*". Statistics education with a central role for ICT offers a unique possibility to apply the learned concepts to data sets from the real world, if possible with an appealing context for the target group.

With this curricular positioning, we were ready to develop our intervention more concretely, starting with a specification of the feedback process and the types of feedback in section 5.3 .

5.3 Feedback process and feedback types

In this section we describe how the feedback is to be processed in classroom practice and which types of feedback we distinguish. This process and these types are the very core of the prototype. Therefore we begin by describing them before we sketch the structure and a further specification of the prototype (section 5.4).

5.3.1 Feedback process

Figure 5.1 depicts the way the classroom network is built up.

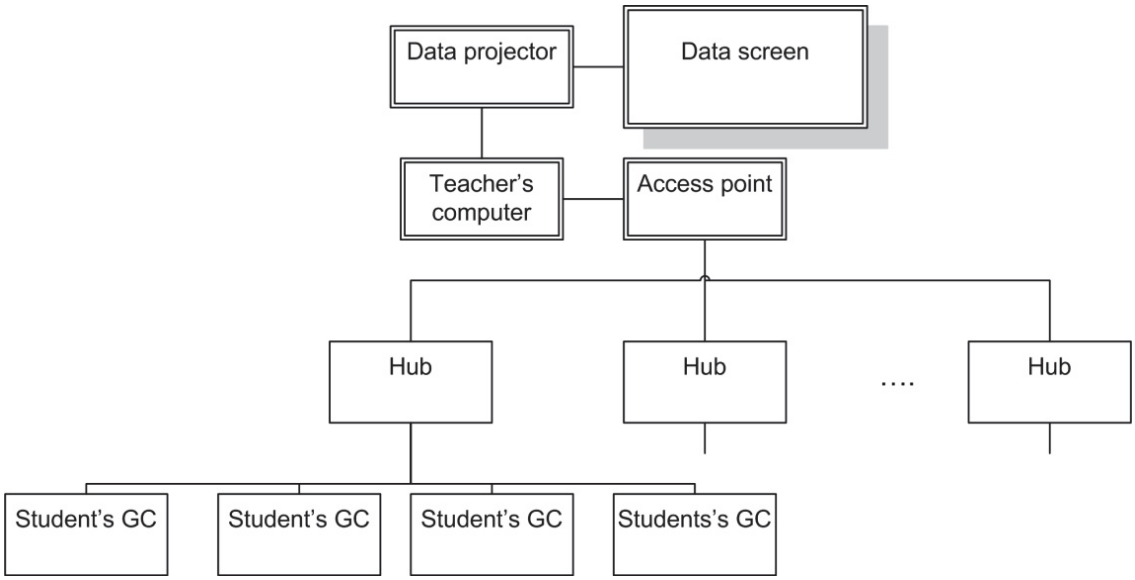


Figure 5.1 The structure of the classroom network

The teacher's computer takes a central place. Through an access point and hubs it can communicate with the students' GCs. Each hub is connected to four GCs. On the other end, the teacher's computer is connected to a data projector that shows on a data screen the results and instructions the teacher chooses to be useful.

When evaluating the third prototype, technology was advanced in such a way that the students' GCs were able to communicate directly with the access point connected with the teacher's computer so hubs were not needed any more.

The process of how feedback per exercise was delivered is depicted in figure 5.2.

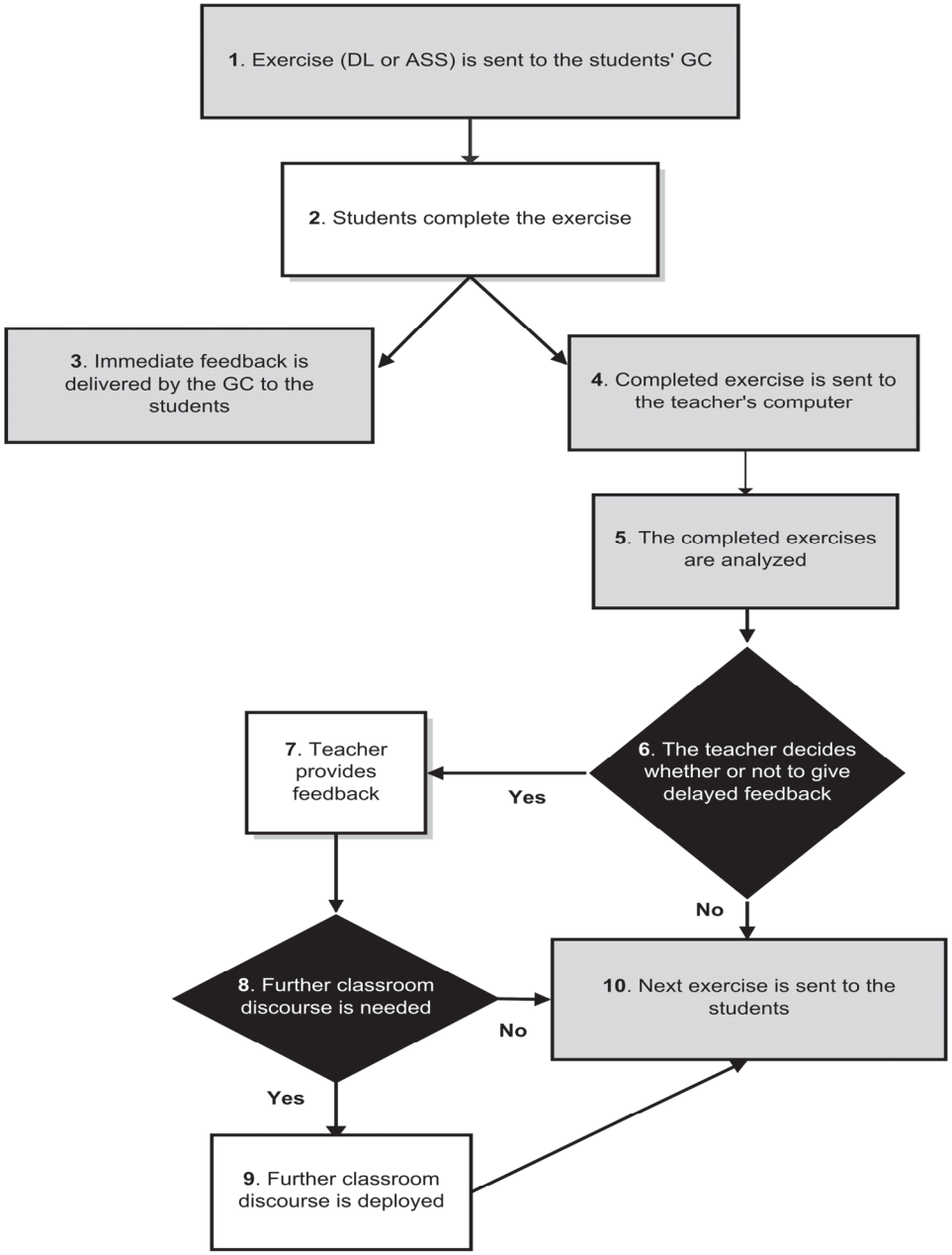


Figure 5.2 The process of feedback in the classroom network

We note that more than one exercise can be sent concurrently through the network.

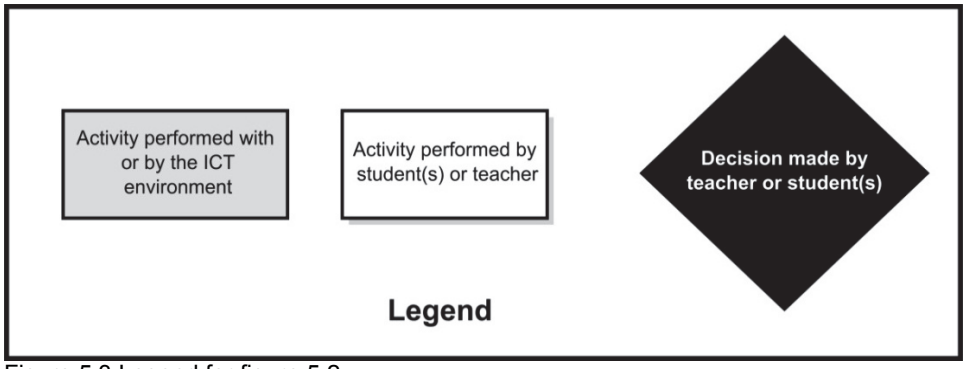


Figure 5.3 Legend for figure 5.2

What do we see happening here? An exercise is sent from the teacher's computer to the students' GC (step 1). The students complete the exercise (step 2). When they consider this completed, they press the 'Check' button on their GC. The GC then provides immediate, simple, feedback (right/wrong; step 3). When the teacher thinks the time is right, he collects the students' work. Through the classroom network this work is then sent to his computer (decision point: step 4). This work is then ordered and basically evaluated (step 5; by ClassAnalysis, a software tool on his computer). If the teacher considers this appropriate (decision point: step 6), he uses the data projector to show the students' results on which he gives feedback (step 7) that probably (decision point: step 8) starts an interactive classroom discourse (step 9). He can also decide, usually in the case of an exercise that requires use of a more complicated algorithmic skill, by using a tool that emulates the GC on his computer, how the GC had to be used in order to do the exercise correctly. If no feedback is needed, the teacher can send the next exercise to the students, or collect the results from the subsequent exercise and take up the feedback process again (at step 6).

It is important to notice that immediate feedback is always provided to the students on each completed exercise (step 3) and that the teacher can choose (step 6) to deliver delayed feedback. How does he make this decision (step 6)? He is informed by his computer how an exercise has been worked out by the students. As a rule of thumb, we stated that when more than two thirds of the students completed the exercise correctly, no (delayed) teacher feedback is needed. If less than two thirds of the students did so, (delayed) teacher feedback can be supplied. But we did not formulate this in a too compulsory way, because we considered the teacher's judgement of the situation of great importance too. In chapter 7 we describe why the experiences during the first two case studies of C3 gave us reason to change this, by splitting up the prototype in 'self-check sections' in which the immediate feedback is still implemented and 'plenary sections', in which no immediate feedback is implemented and the students, with respect to feedback, fully depend on the teacher.

The other decision that deserves some illustration is taken at step 8. Does the teacher consider his feedback to be that conclusive that a further classroom discourse is not needed? In some cases (see the HTT), we do not recommend further discussion. We call that 'Just answer checking' (JAC). But in the case of an exercise aiming at a new skill or at fundamental data literacy, we strongly encourage further discussion.

5.3.2 Exercise types

In the design of statistics education we focus, amongst others, on student activities. Therefore, we try to formulate the content as much as possible as concrete exercises. By consequently questioning the students, through these concrete exercises, we optimise

feedback chances for the teacher. The environment we chose to build the content in supported the most well-known question types. The basic dichotomy in statistics education 'Data literacy' versus 'Algorithmic statistical skills' (see section 3.2.1) can be supported by using these *question types* as follows. We note that we try to find as many possibilities as possible for realisation of feedback on data literacy activities.

1. True - False (TF): suitable to force students to choose sharply, after having been confronted with a statement ("The median is more robust with respect to outliers than the mean"). The teacher feedback can empower a powerful classroom discourse about the statement. A TF-exercise is suitable for evoking activities focusing on *data literacy*, when the statements consider qualitative reasoning.
2. Custom choices (CC): quite similar to TF, but with an adjustable number of alternatives. Used when the exercise has more than two plausible different answers. Suitable for both *data literacy* as well as *algorithmic statistical skills* activities. We denote this type as MC(number of choices), for example MC(3) for a multiple choice exercise with three alternatives.
3. Fill in the blank (FITB): used to check an exercise with a numerical answer. Highly suitable for assessing *algorithmic statistical skills*, very often having a technical nature, leading to numerical output.
4. Open Response (OR): Used to assess exercises in which students have to formulate their own ideas in their own words on statistical phenomena. Highly suitable for assessing *data literacy*, in which reflection takes a central place.

5.3.3 Examples of intended feedback ordered by type

In the next sections we will describe the four types of feedback, as defined in the feedback matrix for statistics education (see section 5.1.4). We choose for an illustration of the feedback types from the perspective of the first context: 'Was September 2005 a warm month?'

The format we present below is the format used in the formulation of the HTT and is in this presentation communicated with the teachers in the procedural specification by means of a manual and personal discussion.

Feedback type I: Immediate feedback on data literacy

As an example of immediate feedback on data literacy we choose the last question from unit 1, in which the central question was: 'Was September 2005 a warm month?'

Exercise 1.9, type: True/False (TF)

September in the year 2005 was significantly warmer than average.

Learning objective

This exercise is supposed to be the finale of the exercises 1.1-1.9: after calculation of some statistical measures the students should now (1.9) conclude whether or not September 2005 was exceptionally warm. This is very interesting, for instance, because it preludes to a very important domain of mathematical statistics: testing of hypotheses. This domain neither belongs to the scope of this intervention in technical detail, nor to the curriculum of Dutch senior secondary education in general, but thinking about the procedure can be done without the techniques without losing its relevance. We consider that 1.7 degrees warmer (20.0 – 18.3) is enough to judge the statement as 'True', but the answer to question 1.9 is not as important as is the motivation.

Expected students' answer

Students discuss whether this 1.7 degree gap is significant. Perhaps a single student will come up with the idea to take a look at the long term development. We expect that the vast majority of the students will answer 'True' to this statement.

Immediate feedback by the GC

There are two immediate feedback possibilities: 'Incorrect' in the case of a wrong student answer and in the case of a correct answer a random choice from the set {'Way to go!', 'Awesome!', 'Correct!', 'You got it!'}.

Feedback type II: Delayed feedback on data literacy

As an example of delayed feedback on data literacy we choose the feedback on the very first teaching activity of our prototype. It is an Open Response question, primarily suitable for DL exercises.

Exercise 1.1, type: Open Response

Which data would you collect when you were asked: was September 2005 a warm month? Which calculations would you perform?

Learning objectives

To think about the role of data.

To realise that data is needed as the basis for an opinion.

To choose a statistical technique that is suitable to perform on the desired data in order to give a reasonable answer.

Formulate thoughts about the two main topics in statistics: data and algorithms.

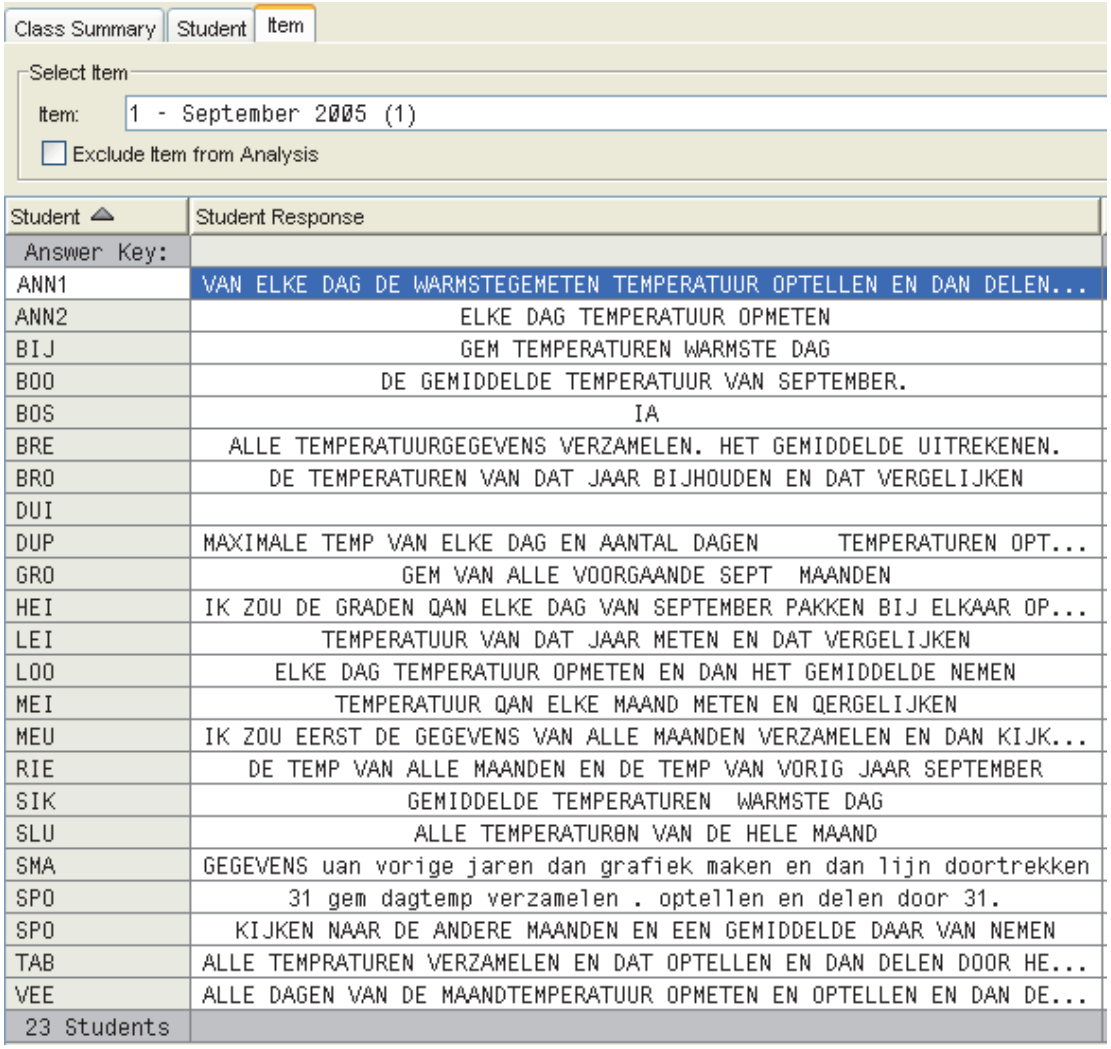
Expected students' answers

Because of the fact that this is an Open Response item, students are completely free in formulating their ideas. In the case of this exercise, their answers might strongly diverge. We expect that most of them will produce some kind of average temperature as a statistic to characterise the temperature of a month. But what exactly do they want to calculate the mean of? They are most likely to want to calculate the mean temperatures of each separate day. This is a good idea, but rather difficult to implement. The following problem then is: how do you calculate the mean temperature of a day? When reasoning very mathematically, this comes to the continuous operation of integration as the limit case of the discrete operation of summation. This is a notion way beyond the mathematical imagination of the students in this group. The mean of the maximum temperature of each day of September 2005 would be a good statistic and not too hard to calculate. Good students will come with this suggestion.

Intended teacher feedback

Input for the teacher feedback is an overview of the students' answers to this exercise. ClassAnalysis is not able to analyse students' work that is posed in natural language, such as is the case with Open Response questions. Therefore, the teacher has to scroll through a table with answers, as is shown in figure 5.4.

We intend teacher feedback to be focussed on the question: When do you consider something to be 'warm'? This should result in the conclusion that 'warm' is a relative concept and that therefore there should be chosen some kind of a reference object. For September 2005 it seems logical to compare it with other September months in the past. Good starting feedback could be: who has his birthday in September? Who remembers what kind of weather it was on your last birthday? Based on the possible answers, a naïve answer on the central question could be suggested.



Student	Student Response
Answer Key:	
ANN1	VAN ELKE DAG DE WARMSTEGEMETEN TEMPERATUUR OPTELLEN EN DAN DELEN...
ANN2	ELKE DAG TEMPERATUUR OPMETEN
BIJ	GEM TEMPERATUREN WARMSTE DAG
BOO	DE GEMIDDELDE TEMPERATUUR VAN SEPTEMBER.
BOS	IA
BRE	ALLE TEMPERATUURGEGEVENS VERZAMELEN. HET GEMIDDELDE UITREKENEN.
BRO	DE TEMPERATUREN VAN DAT JAAR BIJHOUDEN EN DAT VERGELIJKEN
DUI	
DUP	MAXIMALE TEMP VAN ELKE DAG EN AANTAL DAGEN TEMPERATUREN OPT...
GRO	GEM VAN ALLE VOORGAANDE SEPT MAANDEN
HEI	IK ZOU DE GRADEN QAN ELKE DAG VAN SEPTEMBER PAKKEN BIJ ELKAAR OP...
LEI	TEMPERATUUR VAN DAT JAAR METEN EN DAT VERGELIJKEN
LOO	ELKE DAG TEMPERATUUR OPMETEN EN DAN HET GEMIDDELDE NEMEN
MEI	TEMPERATUUR QAN ELKE MAAND METEN EN QERGELIJKEN
MEU	IK ZOU EERST DE GEGEVENS VAN ALLE MAANDEN VERZAMELEN EN DAN KIJK...
RIE	DE TEMP VAN ALLE MAANDEN EN DE TEMP VAN VORIG JAAR SEPTEMBER
SIK	GEMIDDELDE TEMPERATUREN WARMSTE DAG
SLU	ALLE TEMPERATUR0N VAN DE HELE MAAND
SMA	GEGEVENS uan vorige jaren dan grafiek maken en dan lijn doortrekken
SPO	31 gem dagtemp verzamelen . optellen en delen door 31.
SPO	KIJKEN NAAR DE ANDERE MAANDEN EN EEN GEMIDDELDE DAAR VAN NEMEN
TAB	ALLE TEMPRATUREN VERZAMELEN EN DAT OPTELLEN EN DAN DELEN DOOR HE...
VEE	ALLE DAGEN VAN DE MAANDTEMPERATUUR OPMETEN EN OPTELLEN EN DAN DE...
23 Students	

Figure 5.4 Collected students' responses to an OpenResponse (OR) exercise.

The second problem is: How do you characterise the temperature of a specific month? The teacher could mention that this is related to the characterisation of the temperature for a set of days. The calculation of something like the mean for this set brings the problem back to: how do you characterise the temperature of a specific day? This could, for example be done by taking the maximum temperature for each day.

Feedback type III: Immediate feedback on algorithmic statistical skills

To demonstrate the way immediate feedback is delivered by the graphing calculator, we choose an exercise posed in a question type typically suitable for ASS: 'fill in the blank' (FITB) to check a numerical calculation performed on the GC.

Exercise 1.5, type: Fill in the blank (FITB)

Inspect the list sep05. You can find here for each day of September 2005 the maximum temperature. Calculate the mean maximum day temperature in degrees Celsius in September 2005.

The mean maximum day temperature in September 2005 was …. Round off at 1 decimal.

Learning objective

Working with lists is essential when using a GC. It is the most basic representation of numerical data. Thus, statistical operations will usually be performed on lists. This exercise is meant to train skills in using the GC as a device for statistical operations, to start with one of the most well-known statistical concepts: the arithmetic mean.

Expected students' answers

Calculating a mean with a GC is a skill quite different from calculation by hand or with a scientific calculator. It is possible that some students will use the wrong list to perform the calculation on, or that they will perform the needed acts in a wrong order.

Intended feedback

Again, there are two immediate feedback possibilities: 'Incorrect' in the case of a wrong student answer and in the case of a correct answer a random choice from the set {'Way to go!', 'Awesome!', 'Correct!', 'You got it!'}.

Feedback type IV: Delayed feedback on algorithmic statistical skills

As an example of feedback type 4, we choose the feedback on the first time the students had to calculate a mean on their GC. This is exercise 1.5, from which the immediate feedback was described previously.

Exercise 1.5, type: fill in the blank

Inspect the list sep05. You can find here from each day of September 2005 the maximum temperature. Calculate the mean maximum day temperature in degrees Celsius in September 2005.

The mean maximum day temperature in September 2005 was …. Round off at 1 decimal.

Learning objective

Working with lists is essential when using a GC. It is the most basic representation of numerical data. Thus, statistical operations will usually be performed on lists. This exercise is meant to train skills in using the GC as a device for statistical operations, to start with one of the most well-known statistical concepts: the arithmetic mean.

Expected students' answers

This exercise is not too exciting, although calculating a mean on a GC is a skill quite different from calculation by hand or by a scientific calculator. There is a manual in which the required skill is described, but we do not expect that the students will have this manual at hand, let alone that they will inspect it. Therefore it remains possible that some students will use the wrong list to perform the calculation on, or that they will perform the needed acts in a wrong order.

Intended feedback

Before the teacher possibly decides to provide delayed plenary feedback, the students received immediate feedback on this exercise. The teacher again gets informed about the

students' work by ClassAnalysis. Here he finds all the students' answers, with grades and with a mean grade for the class.

Although we expect the vast majority of the students to do this exercise correctly, we could imagine that the teacher would nevertheless decide to give delayed feedback, to ensure everyone has seen the correct operations performed. He would therefore use the GC simulator on his computer and do the exercise himself, the results being shown on the projection screen.

The exact key order can also be displayed so the students are able to concentrate on what the teacher does without caring about imitating him at the same time.

5.4 The structure and a further specification of the prototype

After having sketched the feedback process and the feedback types in section 5.3 we describe in this section the general structure of the prototype and the learning goals and the approach per learning unit.

5.4.1 General structure of the prototype

Derived from a quote from the poem The Rock (1934) by T.S. Elliot:

'Where is the wisdom we have lost in knowledge?

Where is the knowledge we have lost in information?'

we called our prototype 'Data and meaning' (hoping that wisdom will be generated by the classroom discourse initiated by the teacher feedback). In order to determine the central topics in descriptive statistics education, we inspected textbooks on statistics education (see section 3.4) and we discussed these topics with experts. Then we decided that the most important statistics topics for the target group are successively:

1. Introduction and measures of central tendency (arithmetic mean, median, statistical mode/modal class);
2. Measures of variation (range, standard deviation, inter quartile range);
3. Histogram and order;
4. Boxplot and order.

We chose this order because of the need to know at least what the arithmetic mean is to understand how the standard deviation can be calculated. We had two reasons to plan 1 and 2 before 3 and 4. First, we wanted a low profile authentic context as a starting point. September 2005, a month which had ended just a couple of days before the start of the development of the prototype, happened to be a very warm month. The natural question: 'Was September 2005 a warm month?' led logically to the concept of measures of central tendency, more than to the concept of range, that underpins topics 3 and 4. We then considered it to be logical to continue with measures of variation, because with both measures of central tendency and variation, histograms and boxplots can be analysed. A second argument to choose this order was more practical: the skills needed to perform the calculations on the GC in section 1 and 2 are easier than those required for sections 3 and 4. We expected that with the introduction of the classroom network, the students might experience a technical overload when they are at the same time confronted with complicated algorithms on their GC to solve the statistical problems.

The sections are themselves divided into units. A unit is a set of consistent exercises around an explicit learning objective. In the following sections, we will describe the content per unit rather than per section, because of the consistency per learning objective. Each exercise itself also had a specific micro learning objective, contributing to the learning objective of the unit it belongs to. In the HTT, these micro learning objectives are all explicitly stated, together with the exercises themselves.

We note that we describe here the characteristics of the prototype as used in C1. In chapter 6 we formulate the modifications needed because of the experiences during C1 and C2. In chapter 7 we discuss the modifications that were needed during the piloting of the prototype in C3.

In the next section 5.4.2 we discuss the characteristics of the exercises at the level of units.

5.4.2 Units

Unit 1

Exercises: 1.1-1.9 (= section 1, exercises 1 to 9)

Learning objective: To understand how the arithmetic mean functions as a concept of a measure of central tendency.

Due to the introductory nature of this unit, we focused on exercises with a data literacy nature around a more or less obvious question that rises from an authentic context. We expect that a DL nature will motivate the students (in general, but especially those having a non-statistical or mathematical intrinsic focus) more strongly than an introduction with an ASS focus that can have a recipe book resemblance.

The central question in unit 1 is: 'Was September 2005 a warm month?' Related problems are: When do you consider a month to be warm? Do you have to compare it with other months in the same year or with other September months in other years? Which data do you need to determine whether a month is warm or not? Which statistic as a representation for the temperature in a month would be appropriate? How much difference would be enough to decide that a month can be considered 'warm'?

Students are confronted with these questions and then they are given a real data set in order to make a decision themselves.

With respect to ASS: the algorithm for calculation of the arithmetic mean on a GC is introduced and applied a couple of times.

Unit 2

Exercises 1.10-1.17

Learning objective: The central theme in this unit is the choice between two measures of central tendency. In unit 1 just the most well-known of these measures (the arithmetic mean) has been used. Now the 'median' is introduced and its functionality is compared with that of the mean: when is it appropriate to use each method?

These questions arise from the context that is about the number of MP3 music files students have on their computer. With respect to ASS: the algorithm of determining a median from a given data set is introduced and repeated in some exercises.

With respect to DL, the most important difference in nature between mean and median (sensitivity with respect to an outlier) is to be discovered by the students, while exploring a given data set.

In this unit, having a more technical nature than unit 1, there is a mix between exercises having a DL and an ASS nature.

Unit 3

Exercises 2.1-2.6

Learning objective: In this unit, the third and last measure of central tendency to be discussed in this prototype is introduced: the statistical mode. What is the specific reason to use the mode?

With respect to ASS, the application of the mode to a data set is practiced and the mode as a base for *modal class* is to be discovered by the students. Then a direct relationship between mode and histogram arises, which will be worked out in unit 7.

Although the statistical mode is not associated with a very complicated algorithm, it has some subtlety in its use hence we chose to put the emphasis on DL with respect to these exercises.

The unit ends with a recapitulation around the question: Which actions are involved in a statistical inquiry?

Unit 4

Exercises 2.7-2.13

Learning objective: Does a measure of central tendency represent all of the information from the data set? Students are thus to discover in this unit that when using a statistical metric (here, we consider a measure of central tendency: mean, median or mode) you will lose some information from the original data set. Besides that, just a measure of central tendency does not tell you anything about the variation in the data set and therefore not about the distribution of the data. This idea of distribution can be considered as an overarching concept for statistics education (Cobb & McClain, 2004). Students are asked to reflect on the representation of a data set by just a measure of central tendency. And they are asked to suggest a measure of variation, likely to result in some raw idea about ‘range’.

With respect to ASS, determining the mean, median and mode of a data set is repeated.

Unit 5

Exercises 2.14-2.18

Learning objective: To operationalise the naive idea about variation developed in the previous unit. The students are given a real data set, representing the heights of 388 Dutch 12 -year-old boys, and are then asked to calculate mean and median (repetition) and then minimum, maximum and range from this data set. This is a step towards further understanding of variation.

The nature of the exercises is focused on ASS, needed for an understanding of the concept of standard deviation, which is to be explored in the next unit.

Unit 6

Exercises 3.1-3.8

Learning objective: To become familiar with the basic properties of the standard deviation as a measure of spread. In unit 4, a naive measure of variation was suggested, which was operationalised in unit 5. In this unit a more sophisticated measure of variation is introduced: the standard deviation. It is introduced embedded in the repetition of previously developed skills (calculation of range and mean, concepts that are necessary in understanding the concept of the standard deviation) and uses the data set of the heights of 12-year-old Dutch boys and a data set that contains the amount of pocket money these boys receive. This part has an obvious ASS nature.

With respect to DL, implicitly, the coefficient of variation is used to explain the difference between two data sets. Finally, reflection is required on the use of estimation of SD.

Unit 7

Exercises 3.9 – 3.11

Learning objective: For this very short unit, the goal is twofold:

to introduce the concept of histograms, as a graphical representation of data, to students;

to use the histogram as an immediate identifier of a formerly introduced concept: the modal class.

In this unit, the same data set as in unit 6 is used. The students are asked to construct a histogram on their GC themselves, as well as to reason with them. Thus, the exercises have elements both from DL and ASS.

Unit 8

Exercises 3.12 – 3.15

Learning objective: Understanding the relationship between ordering markers like percentiles, deciles, quartiles and median and the percentage of an ordered data set they represent.

There is no concrete data set to be studied in this unit. It can be regarded as having a DL nature because reasoning is the required student activity although the objects with which this reasoning occurs belong to factual knowledge.

Unit 9

Exercises 3.16 – 3.23

In this unit we study data representing the computer use of the Dutch 12-year-olds. We specifically look at the differences between boys and girls.

Learning objective: To calculate a measure of central tendency as well as some measures of variation for two data sets (use of the computer by boys, use of the computer by girls) and then to draw a conclusion about the populations behind the data sets: do boys use the computer more often than girls? What can be said about the variation in both data sets?

This can be seen as an extension of unit 1, in which a conclusion was based on a measure of central tendency alone.

The first part of the unit stresses ASS, repeating previously trained skills, while the conclusion with respect to the outcomes of these ASS activities belongs to DL.

Unit 10

Exercises 4.1 – 4.10

Learning objective: To understand key characteristics of the boxplot. In this unit the students are first asked to calculate those characteristics of a data set, representing the height of boys, that characterise the boxplot: the minimum, the first quartile (Q_1), the second quartile (Q_2 = the median), the third quartile (Q_3) and the maximum. The boxplot represents these characteristics from this data set graphically. After this, the students have to interpret a given boxplot. Finally, using the data set representing the amount of time spent by girls using the computer, they have to construct a boxplot themselves.

This unit has a strong emphasis on ASS because we expect students to become more self-confident when they are in control of their GC with respect to this new concept of a boxplot.

We chose to use the (more simple) definition of the boxplot used in Dutch secondary education. This simple definition of the boxplot has been implemented in the GC we use. We thus do not use the more sophisticated definition originally given by Tukey (1977) in which there are rules concerning whether or not the minimum and/ or maximum is to be included in the boxplot (they are not allowed to be more than 1.5 times IQR (inter quartile range) away from Q1 (in case of the minimum) or Q3 (in case of the maximum).

Unit 11

Exercises 4.11 – 4.15 (repetition)

Learning objective: To reason qualitatively about a large set of concrete data linked to real life through the use of descriptive statistics. This unit has as a central topic the population pyramid representing the Dutch population and population change during the twentieth century. Because constructing a histogram on a GC is not trivial, the unit starts with a repetition of this ASS. After the introduction of data set and context, some characteristics from the data set are determined. Then an answer to the natural question that concerns demographical extrapolation of the situation sketched by the previous exercises is given. Giving this answer is a DL activity.

Unit 12

Exercises 4.16 – 4.19 (repetition)

Learning objective: To formulate a conjecture and to research it by the use of descriptive statistics. Students are challenged to formulate an a priori conjecture: '13-year-old boys will use the computer more than 12-year-old boys'. Then the students are asked to construct the boxplots needed for the comparison of the two data sets. Then they should draw a conclusion, based on their calculations, with respect to the conjecture.

This unit thus has a classical structure: confronted with a 'real life problem', use DL to formulate the problem in statistical terms, then use ASS to perform the needed calculations, before using DL again to draw a conclusion, based on previous calculations. Freudenthal (1991) called this switching between the world of real life and the world of mathematics *horizontal* mathematization. Performing mathematical operations (within the world of mathematics) is in his terms *vertical* mathematization.

Chapter 6
Evaluation of the first and second prototype

In chapter 5 we described the design and development of the prototypes of an intervention aiming at supporting teachers in providing feedback in statistics education facilitated by a classroom network. This prototype includes a hypothetical teaching trajectory, in which we formulate the feedback the teacher is expected to provide to the students on the tasks they performed. In this chapter we describe the planning and the results of the evaluation of two piloted prototypes during the cycles C1 and C2. For each prototype we formulate the adaptions that are suggested by the successive evaluations. The adaptions led to a third prototype that was also used and piloted in classroom practice. The evaluation of this prototype will be described in chapter 7. We report separately about C1+2 and C3 because of the different time interval between the cycles C1 and C2 (two months) and the cycles C2 and C3 (three and a half year).

6.1 A coding scheme for teacher feedback

In this section we describe a coding scheme for the quality of teacher feedback. In terms of quality we mean: how successful was implementation of the feedback (in classroom practice) compared with the intention (in the HTT)? As we stated in section 4.4.4, we use the following format for the evaluation questions:

> (How) does the teacher use the classroom network regarding exercise[*x*] to give feedback[*x*] on the learning objective of exercise[*x*]?
>
> (How) does this feedback[*x*] prompt students to contribute to the classroom discourse around the learning objective of exercise[*x*]?

We recall that, for this intervention, we see teacher feedback as a propelling force to an interactive classroom discourse. Therefore, when classifying teacher feedback, we consider the teacher reactions on the students' contributions during this classroom discourse as follow up feedback.

When studying the quality of teacher feedback in the context of this study there are two other phenomena to bear in mind. First, we focus on the students' contributions themselves. Do they show 'emergence of mathematical meaning' (Cobb & Bauersfeld, 1995)? In our view, 'mathematical meaning' can emerge with respect to data literacy (DL) and to algorithmic statistical skills (ASS). In the case of the latter, this can, for example, be a matter of applying the algorithms in such a way that insight in goal and context appears. Second, we have to be aware of the way the classroom network (CN) facilitates teacher feedback and subsequent classroom discourse: which parts of the feedback and subsequent classroom discourse are made possible through what kinds of teacher actions with the CN? The CN will be limited to ClassAnalysis (CA; the teacher uses the CN to analyse the collected students' responses to the exercises), SmartView (SV; the teacher uses the CN to present his work on his own GC) and GetStudents'Screens (GSS; the teacher uses the CN to present the students' work on their GCs). We will score per exercise the type of feedback by how it is supported with the use of the CN. Therefore, we use the following codes:

CAF:	CA Feedback
SVF:	SV Feedback
CASVF:	CA then SV Feedback
CA-SV-GSS-F:	CA, then SV with GSS Feedback
NF:	No Feedback
ND:	No Data

We will further score the relevance of the use of the CN per exercise from 0 to 3.

0: *no* use at all

1: *minimal* use; for example using CA in order to present the exercise without using its results

2: *some* use; for example more or less randomly picking out some of the CA results and using them to initiate classroom discourse

3: *productive* use; for example using CA to conclude that a skill has not been mastered by the students yet, then using SV for an interactive demonstration (interactive/authoritative discourse as classified by Scott, Mortimer & Aguiar (2006)) following which of the students is able to perform this skill, while using GSS to monitor student progression, possibly using students' results in order to demonstrate this skill

6.2 Practicality of the first prototype (C1)

6.2.1 Results

During this pilot, we did not have the goal of testing the content and the structure of the content substantially. Instead, we wanted to know whether the teacher and students were able to work smoothly in this learning environment. Additionally, we wanted to know how the teacher and students perceived working with a classroom network as a way to provide (teacher) and receive (student) more feedback during the teaching process. Thus, the focus in this pilot is on *practicality* (Nieveen, 2009).

Observations

We observed six lessons of 60 minutes each. In none of these lessons were all students present with, on average, 3-4 absent in each lesson. Even in the very first lesson, dedicated to technical affairs and therefore of course messy, students exhibit non-motivated behaviour such as having an earplug of their MP3-player in or making a phone call behind the teacher's back. In the fourth lesson, there was even a student who exchanged the batteries from the GC he borrowed from us with those in his iPod. There was a student during one of the lessons, starting a petition for an extra test retake because in the preparation period the teacher had been ill for a week. Almost every lesson one or more students were sent away because of unacceptable behaviour or because they did not complete any work (as was shown by ClassAnalysis).

Nevertheless, we observed that in all lessons there were moments where students were engaged, usually by a plenary discussion of the ClassAnalysis of their work. The third lesson (out of six) was the best one. For example, when the teacher said that he was to collect the students work with the CN, a student shouted 'Wait, wait!, apparently because he wanted to finish something quickly and wished it to be collected. It was remarkable

how fast the whole class could change from doing almost nothing in an uncoordinated atmosphere into attentive followers of feedback on their work.

The use of a tool that simulated the GC on the teacher's computer (not included in the standard CN software suite), and thus offering through the use of a data projector the possibility to demonstrate classroom-wide how the GC should be used, proved to be very useful in the explanation of, and feedback on, the exercises focusing on Algorithmic Statistical Skills.

Besides the lack of mutual confidence in the classroom, there were a number of technical problems. First of all, there were memory problems with the GCs.

1. We chose for CellSheet, a spreadsheet application for the GC, as a medium for the data sets used. The main reason for this choice was that a spreadsheet is the primary tool for handling numerical data. However, the use of CellSheet caused problems with the memory management on the GCs. The use of *lists* could offer a solution. Lists (TI jargon for 'one dimensional arrays') are intrinsic variables and therefore do not take extra memory.
2. Another problem with respect to memory capacity was caused by the size of some of the used CellSheets (three columns, 388 rows). We were not able to solve this problem because we chose to use real data sets of considerable size. Among other reasons, this size requires the use of ICT. Calculations performed by hand on data sets with these sizes would consume too much time.
3. There were principal shortcomings of the GCs operating system. Switching between different modes of the GC, for instance the network (communication) mode and the calculation mode, was very cumbersome. This problem is intrinsic, because of the principal 'single tasking' design of the operating system that runs the GC. Due to the choice of a certain classroom network infrastructure (TI navigator), we had to choose this type of GC (TI-83 Plus and TI-84 Plus Silver Edition), because at that time other types were incompatible with the classroom network infrastructure. Therefore, we were not able to solve this problem.

Student questionnaire

When the post-questionnaire results are compared with the pre-questionnaire results, some interesting results appear. Although the appreciation of mathematics as a school subject did not grow significantly (at $\alpha = 0.05$) during the intervention, students considered it useful to put effort into improving working with a CN in mathematics education ($p = 0.032 < \alpha$; $\alpha = 0.05$). Apparently, they saw some kind of potential. When asked about *how* they exactly experienced working with a CN, there were about as many positive answers as negative. There were two 'principal groups': one answering that working with technology is almost always better than working the traditional way and the other who did not really like working with ICT at all. Some students had substantive comments. On the positive side it was mentioned that it was a good way of working together with the teacher on an exercise, that the projected answers helped students in learning mathematics, that there was fast feedback on whether an answer was right or wrong, that it was useful to work with all of the students on a problem and that there was more explanation. On the negative side, all of the substantive comments were focussed on failing of the technology (memory problems, slowness of the network, a too complicated way of working).

Interviews

We evaluated the questionnaire results and used these in an interview with three selected students (a good, average , and weak student with respect to mathematical competence according to the teacher's opinion). The results of these interviews were used as an input for the concluding interview with the C1 teacher. The main result is that students and the teacher more or less agreed on the fact that feedback and interaction possibilities were enhanced. The teacher added that his overview of the students' performance and thinking was much better, especially with respect to the introverted students. The students mentioned that the teacher had a better overview, which was important for better feedback on students' work. In their opinion, this increased the students' attentiveness during the evaluation of their work in a plenary discussion. Students and the teacher agreed on the fact that technology was still not working well enough. All of these findings are consistent with our observations.

6.2.2 Conclusions

After evaluation of the prototype during C1, we improved the prototype by undertaking the steps below.

Technical problems

Cell sheet

As a result of too strong a demand on the memory resources we had to replace CellSheet (external application) as the main tool for data representation by List (internal application of the GC). This rather fundamentally changed the concrete skills needed (ASS), because CellSheets works two-dimensionally (like matrices) and List just one-dimensionally (like arrays).

File size

Since the files representing the data sets were large, fine tuning was required with respect to memory management on the GCs. This fine tuning is described in the new version of the student tutorial, so the students are expected to perform the required actions independently from the teacher.

Switching between calculator mode and communication mode

We unfortunately could do nothing about this problem. The chosen GC had an architecture, dating from the early 1980s, that was not suited for multitasking. Preparing the students to deal most efficiently with this shortcoming is the best we could do.

GC emulator

We added the software tool that emulates a GC (SmartView) on his computer to the tools to be mastered by the teacher when aiming at utilisation of classroom networks.

Perceived feedback by students and the teacher

Although we have not seen many convincing units of statistics education, the students' and teacher's opinion about the feedback possibilities when working with a CN was fairly positive. This does not preclude that the technology had to mature in order to fully utilise this potential. In our observations we saw sufficient behaviour, both from students and from the teacher, illustrating their opinion, in order to make these opinions reliable.

6.3 Overview of the realised teacher using the second prototype (C2)

Classification of the realised feedback

We now recall the codes used to classify the teacher's use of the classroom network (CN) in order to give feedback:

CAF:	ClassAnalysis feedback
SVF:	SmartView feedback
CASVF:	CA then SV feedback
CA-SV-GSS-F:	CA then SV then Get Students' screens feedback
NCN:	No classroom network used
NF:	No feedback
ND:	No data

For the sake of this overview, we mapped the dominating type of feedback as occurred during the classroom discourse to one of these categories. The teaching materials were organised in twelve units. We define these teacher motives for the intended feedback: a=new ASS, b=complex ASS (done before, needs some extra attention), c=DL.

In table 6.1 we depict these moments-motives combined with the coded type of feedback. Between brackets we denote for each feedback session the relevance of the input of the classroom network. Three CN tools will be considered: ClassAnalysis (CA), SmartView (SV) and Get Students' Screens (GSS).

Table 6.1 Exercises with feedback motives and the realised feedback

Unit 1					
1.1 (c)	1.3 (c)		1.5 (a)	1.9 (c)	
CAF(1)	NF(0)		CAF(2) no SV	CAF(1)	
Unit 2					
1.10 (c)	1.11 (a)		1.16 (c)	1.17 (c)	
CAF(2)	CAF(2) no SV		CAF(1)	CAF(1)	
Unit 3					
2.2 (c)	2.3 (c)	2.4 (c)	2.5 (c)	2.6 (c)	2.1 (a)
ND	ND	ND	ND	ND	ND
Unit 4					
2.10 (c)	2.12 (c)	2.13 (c)			
ND	ND	ND			
Unit 5					

2.18 (a)					
ND					
Unit 6					
3.2 (a)	3.7 (c)	3.8 (c)			
CASVF(2)	CAF(2)	NF(0)			
Unit 7					
3.9-I (a)		3.9-II (c)			
SV-GSS-F(3)		NF(0)			
Unit 8					
3.12-3.15					
ND					
Unit 9					
3.18(b)	3.23 (c)				
ND	ND				
Unit 10					
4.5-I (a)		4.5-II (c)		4.10 (b)	
CA-SV-GSS-F(3)		CAF(2)		ND	
Unit 11					
4.11 (b)	4.13 (c)	4.14 (a)	4.15 (c)		
SVF(3)	NCN(1)	SVF(2)	CAF(1)		
Unit 12					
4.16 (c)	4.19 (c)	4.19 (c)			
ND	ND	ND			

What do we conclude when we compare the realised feedback with the expected feedback?

Reasons for 'No data' events

First of all, we notice the high percentage of 'No data' events (50%). This was due to a combination of reasons.

One reason was technical failure. Memory management on the students' GCs continued to be a time-consuming problem. This was due to letting the students use their own GC loaded with RAM absorbing applications. Our argument for this choice was that this made the intervention more 'everyday', the workflow being influenced enough by the networking element. In hindsight, this was not a convenient choice. It would have been better to supply the students with clean GCs, instructing them not to install non-disciplinary tools and taking the GCs back after the intervention. In addition, during lesson 3 in which unit 3 had to be discussed, the interactive whiteboard (IWB) showed defects. This led to 6 'No data events'.

Further, scheduling in a school organisation is always tighter than is favourable for a research project. When (parts of) lessons were lost due to technical failure it was difficult to schedule them later on. Finally, we planned the intervention too optimistically. Facing the above problems should have been incorporated into our planning. Unit 12, for example, could not be recorded because, due to the timing problems, the contracts of the research assistants responsible for the two cameras and the reservation of the equipment had ended. We then chose to continue the observation using the observation form as had been used by the principal researcher during previous lessons. On reflection we consider that to be the wrong decision and conclude that it would have been better had the lesson been recorded with a private camera.

In the evaluation below we leave out these 'No data events'.

Learning activity and teacher feedback

In section 5.3.2 we decided to model Data literacy (DL) activities and Algorithmic statistical skill (ASS) activities with specific question types. We described the intended feedback on the students' activities with respect to these question types in our Hypothetical Teaching Trajectories (HTT). Combining these aspects led to the following chain from learning activity to expected feedback:

Activity		Question type		Feedback type
DL	→	OR / TF / MC	→	CAF
ASS	→	FITB / MC	→	(CA)SV(GSS)F

Feedback on data literacy activities

According to the coding scheme in section 6.1 , the DL activities are to be provided with feedback supported by CA. What do we see when we compare the expected feedback type on these activities in table 6.1 (denoted with a 'c') with the feedback type that was realised?

On eight out of eleven DL exercises the teacher used CA in order to give plenary feedback. On two out of eleven DL exercises he decided not to give feedback. On one DL exercise he gave feedback without using CA. We concluded that in most cases he did use CA. What can we conclude when we look at the relevance of the classroom network for the realised feedback? From the eight events during which the teacher used CA in order to give feedback, for five events the CA use was scored with a '1', thus in a minimal way. For three events the use was scored with a '2', standing for 'some use' of CA. We consider a mean score of 1.38 on a scale of 0 to 3 to be unsatisfactory. In order to be a reasonable success this mean score had to be somewhere around 2. In chapter 7 we will discuss what in our view could be done in order to improve the use of CA for the benefit of developing students' DL. Is there something we can say about the situations in which

the use of CA for the supply of feedback was partly relevant (scored with a '2')? This was the case during the feedback on the exercises 1.10, 3.7 and 4.5-II. There seems to be no intrinsic reason why the feedback on these exercises succeeds more than that on the other exercises.

We conclude that the teacher's use of CA for the supply of feedback in order to develop students' DL is rudimentary. He mostly used it for recalling the exercise. However, because the majority of the students completed the exercises before they were discussed in the classroom, this recalling usually was enough for initiating a meaningful classroom discourse. Sometimes the teacher used the students' answers to feed the classroom discourse. This led to a focussed classroom discourse.

Unfortunately, there is a principal shortcoming when wanting to take the students' answers on an OpenResponse item (preferred question type for DL activities) as a starting point for teacher feedback and further classroom discourse: the keyboard on the GC. Evaluation of the students' answers points out that their contributions on OR items are considerably poorer than their activities on DL that were modelled for instance as a TF or MC item. This usually did not obstruct productive teacher feedback and subsequent meaningful classroom discourse. However, the true potential of a classroom network with the students' work as an explicit and transparent starting place for teacher feedback leading to a substantive classroom discourse was not completely utilised.

Feedback on algorithmic statistical skill activities

What do we see when we compare the ASS activities in table 6.1 (denoted with an 'a' (new) and 'b' (complex)) with the feedback type that was realised?

When evaluating the CN support of the teacher feedback, the situation is somewhat less simple than in case of the feedback on the DL activities. This is because all three CN tools utilised during this intervention (CA, SV, GSS) play a role. We describe these roles below.

On five out of seven ASS exercises the teacher used SV in order to give plenary feedback.

- During two of these five feedback sessions, the teacher first used CA and then decided that a SV demonstration was needed.
    - During one of these two feedback sessions, the teacher used CA-SV-GSS (scored with a '3' with respect to the relevance of the support by the CN).
    - During one of these two feedback sessions, he just used CA-SV (scored with a '2' with respect to the relevance of the support by the CN).
- During three of these five feedback sessions the teacher did not use CA results to start SV feedback.
    - During one of these three feedback sessions, the teacher used SV-GSS (scored with a '3' with respect to the relevance of the support by the CN).
    - During two of these three feedback sessions, he just used SV (scored with a '2' and a '3'with respect to the relevance of the support by the CN).

During two out of seven feedback sessions the teacher first used CA and decided not to use SV for a demonstration, presumably considering the asked skill not to be complex enough and/or the students' results being satisfactory enough that a plenary demonstration was not really needed. In both cases the p-values of the students' scores

were good (0.92 and 0.63 respectively when giving the students who didn't answer, 2 – 9 respectively, a 0 score) so this teacher's decision could be justified by the students' level of skill mastering. Both feedback actions were scored with a '2' with respect to the relevance of the support by the CN.

What can we conclude when we look at the relevance of the classroom network for the realised feedback and subsequent classroom discourse?

From the five feedback sessions during which the teacher used SV, for three sessions the CN use was scored with a '3', thus in a productive way. For two sessions the use was scored with a '2', standing for 'some use' of the CN.We consider a mean score of 2.43 (the two CAF no SV events included) on a scale of 0 to 3 to be satisfactory. In chapter 7 we will discuss what in our view could be done in order to further improve the use of SV for the benefit of developing students' ASS.

When we return to the realised teacher feedback (section 6.1) we generally see that the feedback during the ASS events supported by the CN led to a productive classroom discourse. Even without intensive use of CA (that would have made it possible to make every student accountable for his or her own work) there were a lot of student contributions to the classroom discourse. Eight or nine students voluntarily discussing with the teacher on one exercise was not an exception. An especially inspired classroom discourse was structured by this series of activities (see for instance the classroom discourse on exercise 4.5-I):

1. Teacher uses CA to determine whether the skill is mastered.
2. He decides that the students' results are not good enough yet and starts SV.
3. He interactively uses SV to demonstrate the skill.
4. He then gives the students a couple of minutes to perform this skill themselves on their GC.
5. After a couple of minutes he captures the students' screens to check whether they have succeeded and gives feedback on these results.

The interactive and 'live' character of the activities seems to be motivating for the students, presumably because their initial mastering of the skill was shown to be insufficient so they embrace this workflow.

It appears that the teacher feedback and the subsequent classroom discourse on ASS activities were more specific than in the case of DL activities modelled by OR items.

6.4 Feedback during C2; some exemplary results

In section 6.3 we concluded that the realisation of the feedback with respect to DL (1.38 on a scale from 0 to 3) shows there is quite some potential not utilised. In this section we show two examples that nevertheless do show some of the potential that we described in the HTT: one example on data literacy (DL) and one example on algorithmic statistical skills (ASS).

6.4.1 Feedback example with respect to DL

Feedback on exercise 1.10 (DL)

This exercise (OR) has a DL character, focussing on the reasoning behind the calculation of the median from a data set with an even amount of numbers. The learning goal is to develop awareness of a new concept with respect to central tendency.

Classroom discourse

Lesson 2 [13:17]

Teacher:
"*Until now we have only summarised data collections with the mean. Sometimes it is convenient though to use the median. In the ninth grade, this has been addressed by the teacher, hasn't it*?"

Students: "*Yes*", *"Yes", "Did we too?"*

Teacher:
"*Yes, because it is stated, the median is the middle of a collection of observations, ordered at size. And er ... what 's the problem if you have an even number of observations, suppose you have 10 observations and you want to determine the median, how do you do that?*"

He waits and looks into the classroom. Students call out inaudible things. The teacher then points at student A.

Student A: "*Add the third and the fourth number and then er divide by 2*"

Teacher, after a pause of a split second: "*Yes, or the forth and the fifth perhaps."*

Student A: "*Yes, if it's about six numbers."*

Teacher: "*Yes, if it's about six, or about seven numbers, eight numbers…*"

Student B: "*It has to be an even number*."

Teacher: "*It has to be an even number. If you have an even number, then you have two numbers who are in the middle and those you can calculate the mean of. That's the way you did it in the ninth grade, maybe you can still remember that."*

[14:11]

The teacher touches the IWB and a screenshot of the right answer on a GC display appears.

Students: "*Wow!*"

Teacher: "*This is a display…*" He reads aloud this model answer: " '*A row with an even number has no middle observation, so take the mean of the middle 2 observations'. Yeah, that's the way we did it, go to the middle of the middle 2*."

He clicks on the IWB and the students' results appear on the screen. They are all presented in purple bars, because this is an Open Response exercise, which cannot be evaluated by CA.

Student B: "*All of them are wrong!*"

Teacher: "*Well no… 'Add up the two middle and divide by 2…'* "

He points at a specific answer at the screen and reads it aloud.

"*... that is calculating the mean of the middle 2, isn't it?*"

[14:34]

Teacher: "*Well, here...*" He points at another student answer and reads it aloud. "…'*There is g1* (text messaging spelling; pronounce 'geen', meaning 'no' in Dutch, jt) *middle…That must be SMS language…*"

Student C: "*But that's logical, sir, because then you are able to answer very quickly."*

Teacher: "*Now that is quite smart...*"

And he continues with reading aloud the student answer: [14:42]

"…'*You have no middle so there's nothing to divide if you've got 1 and 2…' Well, I think this is not true, you do have* ***two*** *middle observations, of which you can take the mean…*"

He clicks on the IWB and the next exercise appears on the screen.

Interpretation

Again the teacher starts with repeating the question as should have been answered by the students as homework. This is successful in the way that there is a lively discussion right from the start. The teacher then picks up a remark that sounds reasonable, but then needs some further elaboration. The teacher and class agree that for an ordered data set the median can be calculated by taking the mean of the two observations that together are the centres of the ordered data set.

Students react enthusiastically when seeing the display of a GC projected on the screen. The teacher picks out some student answers to discuss, which is possible due to CA. Again, there is student commitment to the classroom discourse as is shown by the fact that students are strongly engaged by giving their solutions to the exercises and comment on what they see happening on the IWB. The teacher now bases his feedback partly on the concrete students' solutions he gets presented by CA. This particularly interests the students, as is shown by the fact that a student comments on seeing the collected answers: "All of them are wrong!" This is contradicted by the teacher by picking out one of the first student answers appearing in the list on the IWB and by explaining why this answer is equivalent to the model answer.

Students seem quite well prepared for the discussion, presumably because the majority did their homework as is shown by CA. We presume that this has to do with the fact that they know that they are easily controlled by the teacher and that the CN gives them a real contribution to the classroom discourse. The intervention is too short to determine whether the CN has a sustainable influence on the students' homework behaviour. We presume that this is strongly dependent on the consequence of the teacher's use of the CN.

Use of classroom network (CN) for teacher feedback: CAF(2)

6.4.2 Feedback example with respect to ASS

Feedback on exercise 4.5-I

The first part of this exercise has an ASS character. The learning goal is constructing a boxplot for a data set represented by a list.

Classroom discourse

Lesson 5, [13:31]

Teacher:
"*I will just stop it* [CA, jt]*, we just look at a list, we look at the list of the 12-year-old boys and the 12-year-old boys, we're going to present them in a boxplot and that we'll do together here. You have the list in your calculator if everything's okay ... so eh ...* [teacher has started SV, key history is on] *STAT PLOT ... besides, I have to look first if I've got that list myself ... yes ... Shh, does everyone know, shh, everyone still knows how to get here?*"

The class grumbles.

Teacher: "*Well, it's possible like this...*" The teacher points at the key history representing of the keys he had pressed until then.

Teacher: "*...but this is too much, well, I do want to show this, but I have pressed many unnecessary buttons, but you can come here with 2ND Y =, then you will get this menu, we select then er the first, that we activate, so that's ENTER, STAT, ENTER. Now we choose the type of graph that we prefer. We want to create a boxplot and a boxplot, that picture looks like this, but we need the boxplot where you can also read off the median, so that is the one with 3 dashes, which is the second picture on your display ...*"

Student A: "*Yes, but you can't get there.*"

Teacher:
"*Yes you can, go with the arrow to the right, then I'm there, then I'm there, then just by going to the right...*"

Student B: "*Then you have to press ENTER?*"

Teacher: "*Yes. Accepting now: ENTER.*"

Student C: "*How can I get that list?*"

Teacher: "*Yes, we'll get there in a moment, wait a minute, uh, I first remove this list, for the X-list we select LJ12, so you can search for it with 2ND List, like this.*"

The teacher walks to student D, who asks: "*What does the one below mean?*"

Teacher: "*That's the frequency, we'll do that later, this one may be removed.*"

He goes back to the IWB and continues the plenary: "*Well, now I could, now I could create a boxplot, and then I get a Domain Error, yes, you see, I've forgotten, that* [data frequency, jt] *has to be 1, whoa...*" He clicks on the wrong attribute to fill in this 1.

Student E: "1?"

Teacher: "*Yes, all the observations occur once. With the green button you have to deactivate the alpha mode, and if it's okay then I see a boxplot. But I don't.*"

[17:05]
Students are trying very hard to get a boxplot of the list LJ12 on their screens. The teacher helps individual students troubleshooting, mainly with memory management. When helping a student who does get a boxplot on his display he mentions:

[17:35]
"*I don't understand why I don't get a boxplot, but this is all right...*"

Student F: "*Sir, you have got the wrong window settings. I've got 140...*"

The teacher goes to the whiteboard and reproduces these settings.

Student F: "*I've got 140, 200 there I've got 5, then 0.*"

The teacher does the same.

Teacher: "*Yes.*"

On the IWB the right boxplot appears.

[20:01]
Teacher: "*Shhhh. If you've got the window settings, if you've got the windows settings such that on the x-axis, ... A...*"

Student A: "*I've got him.*"

Teacher: "*Very good, if you've got the windows settings such that the heights of those 12-year-old boys fit into the window, one could say, then you get about this picture and if you use TRCE to retrieve where these characteristics are, then he* [the teacher means the cursor of the GC, jt] *moves to the median and he* [the teacher means the GC, jt] *says: 'That's 160'. That's the median. If I use the arrow to move to the right then he moves to the third quartile, then I've got 75% of all of the students, meaning that 75% of all of the students have got a height of 165 centimetres or less. 25%...*"

Student B: "154"

Teacher: "*Yes, have got a height of 154 centimetres or less, that's one quarter.*"

There is some consternation in the classroom.

The teacher uses GSS to retrieve the students' screens.

Teacher: "*Let's take a look whether all of you were able to...*"

Student C: "*I can't get those calculations performed.*"

Teacher, ignoring this last remark for this moment: *"... construct a boxplot, the one I see here, I don't know who that is, I will take a look* [Teacher activates 'Show students' names] *D, has an ...uh... another boxplot, that could probably be caused by the list that still contains a 0 height. Most of you have, uh...shhh, I don't think that's particularly funny, but take a look with us, most of you have a good boxplot...*"

[22:15]

Interpretation

First of all, CA is used to analyse the students' results. The teacher concludes that the students have mastered this skill. Thus, he gives a demonstration of how to construct a boxplot using a GC. This demo is supported by SV. The demonstration is very interactive, when looking at the students' contributions. In the end of his demonstration, he makes a classic mistake: not having inserted good window settings, resulting in a boxplot 'falling from the display'. Window settings are a lot more than just ASS. In fact, when using a GC the management of domains of the variables along the *x* and *y* axes can be considered as 'global function analysis'. Misunderstandings in the use of the GC have most frequently to do with failure with respect to this skill (Burrill, et al., 2002, p. 24). Student F mentions the incorrect display settings and suggests an improvement, the teacher repairs his settings and shares that with the whole class.

After this, he offers the students the opportunity to try again to construct a boxplot, using GSS to see whether they have been able to do it independently.

In our view, the fragment above illustrates a classroom discourse that would have been impossible without the support of a CN. Student F can correct the teacher because he can see the teacher's settings, and the teacher can see at one glance whether the students' results are satisfactory because he retrieves the students' screens with one click on a button. Besides this, there seems to be a classroom culture in which it is allowed that students make contributions like F (correcting the teacher). We presume that working with a classroom network strengthens this culture, although we acknowledge that there has to be a willingness amongst the students to use this opportunity.

The fact that the teacher can take a look at the students screens, all at the same time, possibly (he does not need to project them) in public, makes the students more accountable for their work.

Use of classroom network (CN) for teacher feedback: CA-SV-SSG-F(3)

6.5 Student questionnaire and interviews at the end of C2

6.5.1 Student questionnaire

After cycle C2 we used, just as after C1, a questionnaire to investigate the students' experiences. The results of this questionnaire were used as an input for open semi-structured interviews with three students (a good, average and weak student with respect to mathematical competence according to the teacher's opinion) We wanted to investigate whether the students were generally content with the mathematics lessons as experienced during the intervention. Therefore we used a pre- and post-questionnaire. The rise in appreciation of mathematics and the mathematics lessons was not significant ($\alpha = 0.05$).

When we analysed the comments students gave to expand on their opinions, we noticed, as before in the comments from the students that participated in C1, there was some polarisation: students were either 'ICT haters' or 'ICT lovers'. Positive substantive comments were along the lines of 'this looks like a future way of working'. Those having their doubts stressed the problems with the memory of the GC, the time during class it took troubleshooting and the problematic nature of typing on the GC.

The further focus of this questionnaire was on how the students had perceived working with a CN and especially on how they experienced the feedback they were provided on their work. We confine ourselves to those items on which students had an opinion that significantly differed from neutrality.

- The students considered it to be useful that the teacher used the IWB and CA to represent the exercises that had previously been made ($p = 0.0025$).
- The students considered teacher feedback on their work, as collected and presented by CA, to be useful ($p = 0.024$)
- The students considered the feedback that the teacher provided after he had collected their GC screens (with GSS) to be useful ($p = 0.0004$)
- The students were strongly convinced of the use of having a hard copy of the exercises ($p = 1.9 \cdot 10^{-6}$).

In conclusion, the results of the questionnaires basically support those of the observations. The students seem to appreciate the plenary feedback made possible by the CN. The finding that they appreciated the teacher feedback that followed on the collection of their screens (by GSS, typically feedback used in case of ASS activities) even more than they

appreciated the feedback that was supported by CA (mostly on DL activities) gives further evidence, supporting that from the observations, that feedback with respect to ASS was realised further according to our intentions than to DL activities.

6.5.2 Interviews

The results of the students' interviews, being input for a concluding interview with the teacher, and the results of this concluding interview itself, are described in this section.

Interview students

Three students (a weak one [A], a good one [B] and an average one [C] with respect to mathematical competence according to the teacher's opinion) have been interviewed with respect to four questions.

1. *Do you think mathematics lessons can be more fun through the use of the classroom network?*
   The students considered this to be true. In their opinion it is modern ('MSN like' according to interviewee A) and offers a possibility for variation during the lessons. Apparently, these students consider more variation to be more fun. All of the content delivered by the GC means that working everywhere is easier.
2. *Would you receive more feedback with a good functioning classroom network?*
   The students were really positive about the feedback on their work by the teacher, although one of them (C) considered this feedback as being given too fast and too little. Remarkable is that they mentioned that they prefer a non-anonymous representation of their (and others') work. They said that their own answer being more visible could engage them even more to participate in classroom discourse. They said they did not feel embarrassed by being so strongly visible and they thought their peers shared this opinion. They were positive about the possibility of the screen collection by the teacher. They suggested that a dynamical screen capture (thus as a movie [C]) or a capture with the student's key history included [B] would be even better. Opinions were divided about the use of the immediate feedback on the GC. They liked the idea, but were critical because there was feedback rather evidently leading to the right answer, even after a nonsense answer by the student. On the other hand the students did prefer this answer check, being easier and faster, above the usual answer booklets, that were not frequently used. Immediate feedback could have been more specific on given answers.
3. *Do you think your grades for mathematics would improve with a good functioning classroom network?*
   The students thought a classroom network could enable them to perform better in mathematics. Due to the lessons being more interesting, students felt more committed and there was more feedback, both by the teacher and by the GC. One student (C) stressed that some of the suggestions on improvement of the CN should be followed in order to be able to really improve her grades. She stressed that book and paper, besides the GC, were for her still important media.
4. *What were the main problems when working with the classroom network?*
   Switching between communication and calculation mode; it seemed that

having two GCs would be easier. For some students the Answer check was too tempting. In the opinion of these three students, some of their fellow students had a lot of memory problems on the GC. Further, some students did not take the lessons too seriously because there was only an assessment with a bonus test [meaning that only a result above their mean result so far counted for being graded, jt].

These results are consistent with earlier results. Students are positive about the CN with respect to the contribution to the classroom atmosphere, feedback and a possible raise of their performance. One student mentioned that she had expected more feedback, which is consistent with our observation that the feedback on especially DL activities was not optimally utilised.

We consider it remarkable that there were no comments during the interviews on the cumbersome typing on the GCs keyboard, because we observed these problems and they were mentioned in the questionnaire.

Teacher interview

The teacher considered the students to have a neutral position towards innovation and digital innovation. He had a positive attitude towards ICT in mathematics education and was strongly motivated. He considered the group of 24 students to be of an average cognitive level and attitude. He considered the relationship between the group and himself as average.

We interviewed the teacher with approximately the same questions as we used for the interview with the students.

1. *Would a student receive more feedback when the classroom network worked properly than without a classroom network?*
   The teacher distinguished two kinds of feedback: first, immediate feedback by the GC. He thought the students are receptive towards this because they like to know whether their answer is right or wrong. He considered it easier and more direct than an answer booklet. Second, he thought the delayed feedback he had given with CA was very important to open the discussion. He saw the correct and incorrect answers immediately. He thought it was more confronting, hence he chose a representation without the students' names, also selected because of inexperience. Getting the screens was very handy too, for example when explaining the role of parameters for quadratic functions. 'This is the graph of $y = x^2$, what is the formula for a parabola twice as wide?' Without CN it is impossible to check this for the whole class in an acceptable time, let alone almost simultaneously.
2. *Do you think the students would achieve better with a good functioning classroom network?*
   The teacher thought so, because the students would feel more committed, they would work harder and this would pay out. The teacher thought students liked it when their work was discussed seriously, and this discussion was an extra 'moment of control' that they perhaps needed, because normally they were not up to date with our planning. The teacher thought the improvement of feedback would for some students lead to better achievement.
   The teacher stated that a condition for a properly functioning CN is a big enough RAM of the GCs. In his opinion, there were too many memory errors. He thought the readability of the screen should be improved [he aimed at the

low resolution of the display; jt]. Reading long exercises was troublesome. He added that the same applies to the representation of graphs, but he mentioned: "I got used to that". Additionally he mentioned that he would prefer a wireless network, although the students did not seem to be bothered by the cables.

3. *What were your personal experiences during this intervention?* The teacher admitted he been feeling insecure during the weeks of the intervention. He mentioned that one has to be very aware of which applications are active and the order of using them. He met a number of technical constraints although he did not consider the system to be more complex than other ICT applications. He perceived the usability to be all right. A lot of things were working simultaneously, but each application itself was not hard to handle. He presumed his uncertainty had been caused by the fact that you do not know what is coming up. All these situations were new to him. He considered one day of training, as offered, as enough, but afterwards he would have liked to have one day at school using the system with concrete materials in real lessons.

Again we see consistency with the results we presented before: the teacher presumed an enhancement of student pleasure when working with a good functioning classroom network. The teacher thought a student would receive more feedback when working with a classroom network. He considered his possibilities for giving feedback as improved. This feedback will make the students feel more committed to the classroom discourse and improve their achievements.

The teacher agreed with the students about the shortcomings of the system they had worked with: the shortage in the amount of RAM on the GC and the resolution of the screen of the GC. He added real wirelessness as a condition for the CN to be really well functioning. Like the students, he did not mention the problem with the typing, probably because he had not done the typing himself. Unlike the students he did not mention the roundabout way of switching between the communication and calculation modes of the GC, presumably for the same reason as above: he had not been really confronted with this himself.

It is remarkable that the teacher seemed to stress in his answers, and the example of GSS he gave, the opportunities he saw for improving students' ASS. Classroom discussion on DL activities was not mentioned. This may be explained by the fact that we did not explicitly ask after the possibilities of the CN with respect to these two different sides of statistics education. However, we presume that the biggest advantages he saw and experienced were those on ASS activities. DL is not really prominent 'on his radar' (given the mean score of 1.38 on a scale from 0 to 3). The question is: was this caused by a lack of DL support by the CN or by his teaching style? In chapter 7 we will come back on this question, when evaluating the results of five teachers (with six groups) with the third prototype from the perspective of their teaching styles.

6.6 Conclusions from C1 and C2

We observed a lot of time spent struggling with technical problems. The classroom network itself operated smoothly, but the limited amount of RAM on the GCs caused memory problems. During one lesson the interactive whiteboard refused service. We conclude that working with new technology is still risky and probably always will be. When comparing the use of the CN as a support for teacher feedback we saw a difference with respect to learning goals. The support for feedback on DL exercises is basic, while

the support of ASS is satisfactory, with our HTT (see section 5.7) as a guideline. We observed that in classroom practice a broader range of tools can be utilised to support the feedback on ASS activities (CA, SV and GSS) than when it comes to supporting DL (just CA). Nevertheless, even when CA is only used as a presenting tool for the exercises, this tool can empower students' answers to these exercises as a background for a productive classroom discourse. Even when these answers are not checked during the presentation session, the discourse could still profit from the fact that more students have done their work and that the students realise that their work is just 'one button click' away.

We have to note that some of the 'no data events' could have probably been prevented with lesson planning including more space for repetition. Had we chosen the supply of clean GCs to the students instead of letting them use their own machines, less memory problems would have occurred. Perhaps more important is that we chose not to direct the teacher in when and how to give feedback, but instead to give him a general idea of the feedback possibilities a CN offers for the purpose of statistics education. This made it possible to observe in which situations the teacher instinctively used which features of the CN, but presumably prevented these features from being fully utilised. Roughly, we could state that the teacher was able to give feedback on DL activities using the CN, but was not able to develop a further classroom discourse with respect to these DL activities.

There was more student involvement during the classroom discourse with respect to ASS than to DL activities. Feedback on ASS seems to be more appreciated than feedback on DL. It is remarkable that the teacher seems to stress the opportunities he sees for improving students' ASS. Classroom discussion on DL activities is not mentioned. This could be due to the fact that ASS feedback:

- is better supported by the CN;
- thus receives more attention during the lesson;
- is more successful in terms of student engagement during classroom discourse;
- thus is better remembered by the teacher.

One student mentioned that she had expected more feedback, which is consistent with our observation that the feedback possibilities especially regarding DL activities were not optimally utilised. Thus suboptimal utilisation could be to do with the teacher's feeling of uncertainty as reported during the concluding teacher interview. He explained this uncertainty by the fact that managing the applications of the classroom network simultaneously was complex. This was precluding him from giving feedback on the DL activities, being a process that can be perceived as more vague than giving feedback on concrete ASS activities. We presume that a more explicitly worked out HTT with respect to DL, combined with an even stronger focus on DL in the teaching approach, could improve feedback on DL activities.

The students do not consider the *immediate* feedback by the GC to be particularly useful, but they might prefer it to the answer booklet as an alternative. They seemed to appreciate the *delayed* (teacher) feedback made possible by the CN. Students were in general positive about the CN with respect to the contribution to the classroom atmosphere, feedback and a possible improvement in their performance. The teacher presumed an enhancement of student pleasure when working with a well-functioning classroom network. The teacher thought the students would receive more feedback when working with a classroom network.

Finally, we mention that the use of a CN to support teacher feedback, as seen in our observations (see section 6.2), offers possibilities for a meaningful classroom discourse. The teacher considered his own possibilities for giving feedback as improved. This opinion is consistent with those of the teachers during C1 and the initial study. However, these possibilities have to be recognised and utilised by the teacher. In chapter 7 we will explore whether a further articulated DL character of the prototype and a better support for DL could further improve meaningful classroom discourse with respect to DL. Then, these possibilities could function as handles for further teacher questioning and probing, in order to develop students' higher cognitive thinking. It is thus the responsibility of well-articulated research and well-developed teacher skills to fully utilise this relatively new way of working in the classroom.

6.7 Adaption of the prototype after C2

We now formulate the results and conclusions from this chapter into concrete suggestions for revision of the prototype. In chapter 7 we will describe and analyse the results of this revised prototype being brought into practice. The revisions can be distinguished into three categories: with respect to *technical affairs*, with respect to *teacher preparation*, and with respect to *teaching means*. In chapter 7 we will discuss which of these suggestions were accepted for the redesign of the prototype and why, given the organisational constraints research always has to obey.

Technical

The main problems were on the side of the GCs: insufficient RAM, too low a resolution of the display and the tiresome typing of text on the keyboard. There are two options for improving this: hand out 'clean' TI84 plus machines and ensure the students do not install any non-disciplinary materials, or use a new generation GC.

Working with a notebook computer for the teacher did not diminish the chance that an interactive whiteboard (IWB) could break down, but took away the disadvantage of having to teach with one's back to the class because of the immobility of a desktop computer.

Teacher preparation

The preparation of the teacher should be more direct when it comes to the planning of the feedback. We stress that every feedback session should start with looking at the students' results and trying to give feedback on some striking answers. If possible, try to evoke discussions based on different students' answers. Besides that, we should make clearer during the preparation of each lesson on which exercises we expect the teacher to provide feedback. We presume this will diminish the teacher's feeling of insecurity. As a rough guideline we take all DL exercises and those ASS exercises in which a new skill is introduced. We will discuss the difficult nature of giving feedback on DL activities.

Research results point out the need to be cautious with non-anonymous feedback. For example, normative feedback (with reference to others) can result in lower learning outcomes for low-achieving students than self-referenced feedback (McColskey & Leary, 1985). However, students indicate that they appreciate non-anonymous contributions to the classroom discourse through the CN. This is not contradictory with the advice of McColskey and Leary as long as the teacher keeps his feedback non normative, which is a key guideline when implementing feedback in a setting of formative assessment. This

could mean that the teacher, by way of feedback, will ask a specific student for her or his argumentation for giving a specific answer. We recall here that the teacher during C1 noticed that he was informed more completely about student progress, especially of those who are more introvert. We will more explicitly promote our prototype as a chance to offer those students a voice in the classroom discourse.

We should further instruct the teacher explicitly *not to avoid* feedback leading to classroom discussion about DL, but, if possible, *to focus* on this important side of statistics education. We will indicate these chances explicitly in the redesign of the HTT.

Teaching materials

The DL part of the exercises should be further articulated. In order to decrease student typing, we tried to convert OR exercises to MC exercises without losing the challenge for the students to think themselves. Some new exercises with a DL character have been added in order to focus more strongly on DL. Frequent and consequent use of CA should confirm the students' feeling that their answers do matter.

We divide the teaching materials into parts that correspond to units (thematic) instead of sections (organisational). This could reduce the length of monolithic discussions on the students' results and will focus more on statistical or pedagogical issues.

The option 'Allow multiple answers' is to be disabled for each exercise. This means that students will not be able to 'cheat': fill in some nonsense answer, get feedback from the GC and then re-use that answer pretending it is one's own.

Chapter 7
Evaluation of the third prototype

In this chapter we describe the results of the pilot of the third prototype during six successive case studies. Five teachers from five different schools were involved. One of them taught two groups of students, so we have six case studies. For each case study, sources of data for evaluation of the pilot were: videotapes of the lessons, students' questionnaires, interviews with the students, teacher questionnaires, interviews with the teachers. We combine these sources of data in order to answer the following questions: How did the actual feedback during the implemented instruction compare with the intended feedback and did the subsequent classroom discourse match? We present the answers to these questions with respect to each specific C3 case study and to all of these simultaneously.

7.1 The classroom network revisited

The case studies we describe in this chapter took place in spring 2010. Since conducting C1 and C2 in 2006, the technology had significantly changed. Both handhelds and the network infrastructure underwent updates amongst others following the recommendations we made based on the results of the pilot rounds C1 and C2 and the previous preliminary study.

For now, we note changes relevant for our study:

1. The handhelds:
    a. had a higher resolution (320x240 pixels vs. 160×100 pixels for the old version), making more information, graphical as well as textual, visible on the display;
    b. had an operating system that supported a file system, thus making it possible to easily interchange documents between students and teacher;
    c. had a much bigger storage capacity, thus preventing handheld crashes like we experienced in the first and second pilots in C1 and C2.

Figure 7.1 The GC as used in C3

2. The network (see Figure 7.2)

Figure 7.2 The classroom network components as used in C3

a. utilised, on the students' side, send-receive functionality from a cradle (2), that is to be pushed on the calculator (1);

b. did not need the hub anymore as an intermediary between students' handhelds (1) and the access point (3) connected to the teachers' computer (4). This made students more free in the classroom and interaction more natural.

We encountered, nevertheless, problems because at the start of the first case study the technology was not officially available. At that time, we knew we took a considerable risk. In hindsight, it would have been safer to conduct this pilot using the former technology, being the same as we did in 2006. A big advantage of using the former technology would have been that the students could use their own graphing calculator, and there would have been no need, for the students nor for the teacher, to master a new device. The data thus gathered, would have been more comparable with those collected during the pilot runs in 2006.

However, we chose to work with the new handhelds and the new network. The new handhelds and the new network had been developed amongst others based on our recommendations. As we take our recommendations seriously, we believe that students' learning and our research interest were better served with the new equipment.

7.2 Adaption of the C2 prototype

We described in section 6.7 the suggested adaption of the second prototype with respect to three dimensions: technique, teaching activities, and teacher preparation.

We now briefly describe the way we changed the prototype with respect to these dimensions. The technical issues have been addressed in section 7.1 .

With respect to the teaching activities, a stronger focus was put on DL. That meant developing new exercises, reconsidering the previously developed exercises and sometimes skipping these.

With respect to the teacher preparation, a much stronger emphasis was put on sharing the developmental (and research) goals of the study. Besides the training we offered the teachers collectively, in the same way as we did in C1 and C2, we trained them individually. Each individual lesson, for instance, was discussed the day before it was to be given and evaluated shortly afterwards.

Our experiences during the six cases studies of C3, as described in this chapter, led to two substantial changes in the prototype.

First, we observed that during the first two case studies the students had substantial difficulties with an exercise that used a structured programming approach of a spreadsheet on the GC in order to calculate the standard deviation of a data set. The feedback on this single exercise took about twenty minutes. It resulted in many students just repeating what they saw being presented, albeit with the use of LPF via the classroom network. Apparently, they did not understand what they were doing. We then exchanged this exercise in 'raw data processing' with an exercise that makes use of the in-built possibility of the GC to perform this calculation.

The second change was perhaps even more fundamental for the intervention as a whole. We observed during the first two case studies and in the beginning of the third that teachers tended to discuss all exercises of the prototype with their students. This took an enormous amount of time and led to monotonous lessons, although they were interactive. In order to avoid the overload of the dominant teaching activity of our prototype (feedback sessions), half way the third case study, we split the prototype up into two types of sections:

- Sections consisting mainly of exercises aiming at developing algorithmic statistical skills (ASS). The students could correct these exercises themselves ('self-check'), with right/wrong feedback and the possibility to inspect the right answer. During the feedback sessions planned after the split, the teacher only had to pay attention to these exercises when something really remarkable, as shown by the students' results, happened.
- A section consisting mainly of exercises aiming at developing data literacy (DL). The 'self-check' in these sections was set to 'inactive'. Therefore, the students were dependent on the feedback sessions by the teacher in order to really understand how they had performed during these sections.

7.3 Further development of the feedback coding scheme

In section 6.1 we presented a scheme for the coding of teacher feedback. As we stated in section 4.4.4 we used the following basic format for each evaluation question:

> (How) does the teacher use the classroom network regarding exercise [ID] to give plenary feedback on [described per evaluation question]?
>
> (How) does this feedback prompt students to contribute to the classroom discourse around these issues?

We recall that, for this intervention, we aim for teacher feedback as a stepping stone to more interactive classroom discourse. Therefore, when coding teacher feedback, we judge

teacher feedback in the context of its subsequent classroom discourse. In other words, we consider the teacher reactions on the students' contributions during this classroom discourse as follow up feedback.

What about the nature of the students' contributions? Do they show 'emergence of mathematical meaning' (Cobb & Bauersfeld, 1995)? In our view, 'mathematical meaning' can emerge with respect to both data literacy (DL) and algorithmic statistical skills (ASS). In the case of the latter, this can, for example, be a matter of applying the algorithms in such a way that insight in goal and context appears. We have therefore only scored student contributions to the classroom discourse aimed at this mathematical meaning.

How did we judge the contribution of the classroom network (CN) to the classroom discourse? We wanted to know how, and how frequent, the CN facilitated teacher feedback and subsequent classroom discourse. Which parts of the feedback and the subsequent classroom discourse were made possible through what kinds of actions with the CN? The hypothetical as well as the observed use of the CN is limited to the characteristics described in the feedback classification scheme below.

Feedback coding scheme C3

StatisticalLearningGoal (SLG): A dichotomous variable indicating the learning goal, with values DL (data literacy) and ASS (algorithmic statistical skills)

ClassAnalysisFeedback (CAF): Feedback by the teacher based on information he gathered with the tool ClassAnalysis. To score as:(low(x), CH), (medium(x),CH), (high(x), CH) with:

low: #student-answers=1
medium: #student-answers=2
high: #student-answers ∈{3,…}
and CH (character), DL = data literacy, ASS = algorithmic statistical skills.

JustAnswerChecking (JAC): A specific case of CAF. The teacher casts an eye on the students' results, usually in the 'Slide show' representation. Scoring through the time needed, including the teacher conclusion:
0'00"-0'40": 7
0'40"-1'00": 6
1'00"-1'20": 5
1'20"-1'40": 4
1'40"-2'00": 3
2'00"-2'20": 2
2'20"-2'40": 1
> 2'40": 0

LivePresenterFeedback (LPF): A real-time capture of a specific GC screen, while commenting on the live input and output as shown on it.
To score as 'teacher' or 'student', depending on

	the respective actor and the character (DL = data literacy, ASS = algorithmic statistical skills)
GetStudents'Screens (GSS):	A capture of all of the screens of the students' GCs, for instance to check students' progress on tasks. To score as 'DL' (data literacy) or 'ASS' (algorithmic statistical skills)
QuickPoll (QP):	A fast, more or less improvised question in between. Was not part of the design, but QP, as part of the ICT environment, was discovered by teachers. Some of them felt attracted to its possibilities and could not help using it (incidentally).
StudentsInteractionInput (SII):	Represents three characteristics of the classroom discourse, usually after CAF: 1. The number of reacting students (low, medium, high) low: #students-reacting $\in \{0,1\}$ medium: # students-reacting $\in \{2, 3\}$ high: #students-reacting $\in \{4, 5, 6,\ldots\}$ 2. The number of students' reactions (low, medium, high) low: #student-reactions $\in \{0,1,2, 3\}$ medium: #student-reactions $\in \{4, 5,6\}$ high: #student-reactions $\in \{7,8,\ldots\}$ 3. The character of the reactions (DL = data literacy, ASS = algorithmic statistical skills)
TeacherConclusion (TC)(Scores, CH):	Scores: The quality of the teacher conclusion: low: short medium: clear and complete, referring to none or very few students' answers high: clear and complete, with substantial use of students' answers CH: The character of the conclusion (DL = data literacy, ASS = algorithmic statistical skills)

For each episode we have previously described an intended hypothetical trajectory of teaching (HTT). This HTT determines the character and the sequence of the elements of the teacher-student interaction of the *intended* curriculum. In this chapter we concentrate on the *implemented* curriculum and on the comparison between intended and implemented. For this comparison we use points to indicate the similarity: 2 points for a perfect similarity on CAF, 4 points for a perfect similarity on SII and 1 point for a perfect similarity on TC. We see initial feedback (usually CAF) as a stepping stone for an interactive classroom discourse in which the students are supposed to have an important input (SII), hence the maximum number of points reserved for SII (4) is twice as high as for CAF (2). The teacher conclusion (TC) is not the most important, but is nevertheless a substantial part of the feedback session. Usually, there is thus a 7 point scale for correspondence of the implemented curriculum with respect to the intended curriculum. When other elements in the feedback session are involved (GSS, LPF, QP), we scale the correspondence to 7 points, in order to make events like that with respect to similarity comparable with others (see, for instance, *JAC* in the feedback classification scheme). We stress that a high correspondence score only expresses a high degree of similarity between expectation and implementation and cannot be interpreted as a measure of quality.

Time needed to perform a feedback session was also scored. We consider efficiency to be a quality of a feedback session. However, we noticed that sometimes a teacher, when leading the SII part of the session, perceived a ‘flow of discourse’ and then extended the discussion. This usually added quality to the learning experience although in our scoring system this is not rewarded: SII(h, h, DL) stays SII(h, h, DL). Further, an efficient feedback session rewards itself by creating time for other learning activities. In the next section we will see, for instance, that the teacher who is very efficient in conducting a feedback session is the only one realising all of the feedback sessions as proposed in the HTT.
This led us not to include 'time used' as an explicit variable in our scoring model. We consider 'time used' as an imperfect measure of efficiency, but we used the exact number of students' inputs as a parameter. Thus, for instance: SII(h(3), h(5), DL) becomes SII(h(6), h(10), DL). We will use 'time used' in order to judge the JAC elements (see Feedback coding scheme earlier this section).

An independent researcher was asked to score an episode of implemented learning activities (part of C2). The principal researcher performed the same action. They compared their results and discussed the similarities and the differences in order to reach a consensus on how to use the scoring system. After this exercise, the principal researcher scored the complete implemented curriculum.

7.4 Experiences from teachers and students

In this section we present the results of the evaluation of the prototype. For this evaluation from the teacher’s perspective, we used a teacher questionnaire. In addition, we interviewed all the teachers individually directly after each case study and simultaneously during a group interview after the last case study. From the student perspective, we also used a questionnaire. After analysis of the results of this questionnaire for each case study, we interviewed three students of each participating group. We consider these results to support the interpretation of the results from the observations (sections 7.5 - 7.11).

7.4.1 Teacher questionnaire and supporting interviews

In this section we present the results of the teacher questionnaire, completed with remarks made during the subsequent teacher interview. Although the questionnaire and interview also investigated experiences with the teaching materials, support and other important contextual issues, we restrict ourselves now to report on the core issue: feedback.

Results of teachers' questionnaire and supporting interviews

Case	Item / Response
	Were the teaching materials suitable for developing students' DL (data literacy)?
S1	*I don't know.*
S2A	*Yes*
S3	*Very suitable, the students were really thinking about the contexts and the interpretation of data. That wouldn't have happened when we worked with chapter 7 from the textbook.*
S4A-B	*Very suitable, but a little too much depth.*
S2B	*Yes*
	Feedback process
	Were you more aware of your students' learning by using the wireless network?
S1	*Yes*
S2A	*No, then I would have had to retrieve more often the results per student. Now I just discussed per question the answers of all students together. In 'normal lessons' I always walk around the room and see what's in the students' notebooks (or what's not).*
S3	*Yes and no. I was surprised by the contribution of a number of students. Everyone was involved, but of 32 students it's impossible to keep a perfect record.*
S4A-B	*Yes, for a teacher it takes very little time to check the students' work. Thus you're able to check students' progress.* *During the lessons I noticed that students who were regularly working were far ahead of those who just worked in the lessons. This gap made the classroom discourse more difficult to manage.*
S2B	*Yes, they did visibly nothing.*
	Did the method of teaching in the experiment extend your ability to give feedback?

S1 *Yes*

S2A *Yes, especially because everyone could see where the feedback was given on.*

S3 *Yes, but that's what this project was set up for, wasn't it?*

S4A-B *Yes, and this can only get better if you get more experience. Probably you need to support teachers by offering training.*

S2B *In principle it did, in practice it didn't.*

Could you give *more* feedback to the students involved in educational discussion than during normal lessons?

S1 *Yes*

S2A *I have not really done so, but you can easily involve students in discussions.*

S3 *Yes and no. Here it went fine, because the chosen method requested so. In normal lessons exercises can also be included with this type of contexts that are adequate to involve many students. However, during the 'traditional teaching' this is usually not done.*

S4A-B *I think students were more engaged during the lessons, because the systems made it more difficult for them to hide.*

S2B *In principle it did, in practice it didn't.*

Could you give *more specific* feedback than during normal lessons?

S1 *Yes.*

S2A *Yes. You can see all that you are talking about and can provide feedback on the answers you had not anticipated or that you had not anticipated so frequently given.*

S3 *When it comes to the responses on the questions, no doubt. I see on the screen the exact response of the students.*

S4A-B *Yes, you get very quick and good insight into the students' problems.*

S2B *In principle it did, in practice it didn't.*

Could you utilise the feedback possibilities of the network to support the development of *ASS*?

S1 *Yes*

S2A *Yes, I used them just in the very last lesson. I would use these possibilities now **much** more often.*

S3 *Sometimes. Especially with skills: "Show how you construct a box plot". A student who is good at this can be the LifePresenter and show how this is to be done. The rest of the students can participate, using their own machines. Screen capture (of the whole group) is useful to see if anyone has the same figure.*

S4A-B *Yes, I could utilise them, but not for the full 100%. Timing is difficult and sometimes I just didn't think about it.*

S2B *Hardly*

Could you utilise the feedback possibilities of the network to support the development of *DL*?

S1 *Yes*

S2A *Yes*

S3 *Yes, that was the target.*

S4A-B *Yes, I could utilise them, but not enough through a lack of time.*

S2B *Hardly*

Did you consider the feedback process to be generally better than in 'normal' lessons?

S1 *Yes.*

S2A *Difficult question. The feedback sessions were certainly in the beginning far too forced. Later, having become more flexible, that improved a lot, but then there were not always data to base the feedback on.*

S3 *Yes!*

S4A-B *Much better. Students were more engaged.*

S2B *No*

Do you think students have learned more than usual by using this feedback?

S1 *I do not know.*

S2A *I do not know. But I have noticed that the students during the subsequent lessons were better motivated for a while. But that may also have to be addressed to the topic (probability). If there*

was another lesson cycle with the classroom network, I would expect a higher learning yield.

S3 *Yes, I think so, but I have no proof. At the final test, a couple of weeks after the last lesson, a lot of concepts seemed to be forgotten by a considerable number of students.*

S4A-B *They have learned ASS playfully and we have also discussed DL. Because we just to do just ASS, they have therefore learned more.*

S2B *No.*

Suppose that this teaching tool is part of mathematics teaching generally, what % of the time would you get this education activity for each class to spend?

S1 *50%*

S2A *50%*

S3 *33%*

S4A-B *50%*

S2B *I don't know*

We summarise these results as follows.

We consider S2B as a somewhat aberrant case. The teacher in this case study was more negative and unsure about the results of the intervention than the other teachers. She was that negative about the implementation of the curriculum in her specific case she did not seem to be able to judge freely about the possibilities of the system, according to the answer "In principle it could, in practice it didn't" she gave on several questions about the possibilities of this way of working. Strictly, she could very well be right because one needs a minimum of valid experience to be able to generalise this experience to a more than purely private opinion. In section 7.11 we will illustrate the implemented curriculum for case S2B with some examples of classroom discourse.

There is a considerable agreement among the teachers with respect to the potential of a classroom network in order to support feedback in statistics education. This may be remarkable, as we will see in sections 7.5 -7.11 that there were big differences with respect to the resulting classroom discourse.

The teachers considered the prototype to be suitable for developing DL. They felt better informed about the work of their students. In general, they gave more and more specific feedback. This counts both for ASS as for DL activities. The learning yields were perhaps bigger than in the case of more traditional instruction; on this the teachers were not that sure. They consider about 50% of the lesson time spent on feedback activities supported by the network as optimal.

7.4.2 Student questionnaire

There were 128 students (S1: 25, S2a:23, S3: 32, S4a: 15, S4b: 16, S2b: 17) who completed the questionnaire. The questionnaire consisted of 20 statements about statistics

education and feedback supplied by a classroom network. We selected the five most crucial statements for a statistical analysis. These five are the most important and we select just five in order not to pay too much for the Bonferroni correction. The five statements are:

1. Because the teacher gives feedback on my work in this way, I get *more* feedback.
2. Because the teacher gives feedback on my work in this way, I get *better* feedback.
3. Because the teacher gives feedback on my work in this way, I manage to master *algorithmic statistical skills* better.
4. Because the teacher gives feedback on my work in this way, I manage to master *reasoning about data and calculations* better.
5. I would prefer to contribute to the classroom discussion on the projection screen *anonymously.*

The statements were formulated on a four point Likert scale, referring to a difference with respect to education without feedback through a classroom network (1: I strongly disagree, 2: I disagree, 3: I agree, 4: I strongly agree).

Significance was considered at $\alpha = 0.05$.

Because we posed 5 questions simultaneously, we choose to compare the resulting *p*-values with $\alpha^B = \frac{0.05}{5} = 0.01$, with α^B the Bonferroni-corrected level of significance.

Significant results of this analysis, per case study and with all the students simultaneously (Total), are shown in Table 7.1.

Table 7.1 Results of statistical analysis students' questionnaire

Case	**Cronbach's α**	**Item**	**Response**	**p**
S1	0.68	I would prefer to contribute anonymously to the projected students' answers	Disagree	0.008
S2A	0.52	None		
S3	0.25	Because the teacher gives feedback on my work in this way, I get *more* feedback.	Agree	0.007
S4A	0.57	Because the teacher gives feedback on my work in this way, I get *more* feedback.	Agree	0.002
S4B	0.78	None		
S2B	0.60	Because the teacher gives feedback on my	Disagree	0.007

		work in this way, I get *better* feedback.		
Total (the responses of all of the students simultaneously analysed)	0.57	Because the teacher gives feedback on my work in this way, I get *more* feedback.	Agree	0.006
		I would prefer to contribute anonymously to the projected students' answers	Disagree	0.002

We conclude that, in general, the students perceived more feedback on their work. Differentiation between ASS and DL was not significant. Discussing their answers in public was not threatening. The results differ slightly per case study, as does the consistency of the students' responses.

7.4.3 Interview with selected students

In this section we present the results of the interviews with students. From each participating group, three students were selected: a good (1), an average (2) and a weak (3) student with respect to their mathematical competence, as appointed by their teacher. During these interviews, we took the results of the students' questionnaire in the specific case study as an input.

Students from case study S1

Student 1

Feedback is important, but the way it was done in these lessons wasn't clear enough. I would prefer that my answers would be shown as much as possible, because I like it to know whether I did it correctly. Besides that, I think it's good to see what other students' results on exercises you've done yourself. That my answers are shown as well, I don't consider to be scary. Although I myself prefer a textbook as a learning source, I experienced the advantage of the network in learning how to calculate statistical measures. Discussing about data this way was a little bit vague too. Sometimes there wasn't a real “yes” or “no” as an answer, but something like “both ways are possible”. That may be like the real world, but I prefer a clear “yes” or “no”. I know that it was possible for the teacher to check our homework more easily, but I don't really need that to do my homework. Some organisational issues should be improved, because I see that for some students this is a way of better learning. The system is new and nice, and I think it is interesting and challenging to learn working this way.

Student 2

I'm actually positive about this way of working, because the teacher could easily see what I answered on the exercises. I think I learn better this way than with just a textbook. For the teacher, this is just easier than looking into my notebook to see what I've done. With respect to the skills, I think it's handy because when the teacher is demonstrating how to calculate something, it's easy for me to do it on my own calculator. And that I can see the

work of other students is also quite instructive, because then you can complement each other and I think that's important. We can learn more from each other, because there's more discussion in the lessons. Reasoning about data is made more interesting. Other students perceived that there were a lot of keys on the calculator. But I think that this is just a matter of time before I can profit more from this way of learning. I don't think that my work in public is scary, everyone has his own answers and I don't care about what others think about my answers. I think this is a way of working that will be more common in the future, and I would like that, because new developments like the iPhone for me are motivating. I think the lessons were good and I think that I have learned a lot about statistics.

Student 3

Afterwards, I'm more positive than when I filled in the questionnaire. The calculator offers a lot of possibilities for calculations. But it took me quite some time to get used to it. Perhaps time was too short, but when I rethink it, this calculator was a pretty useful thing. It is impressive that a calculator contains for example a spreadsheet like I know from the computer. With respect to skills, this way of working helps. Although, I consider the calculator to be less clear with respect to the materials than a book, but the exercises on paper (as handed out) helped a lot. The feedback given was better than before, because this was clearer when we discussed it. It was more focussed on my work or on everyone's work actually. Feedback on skills did help me, since it offered me an example of how to perform the calculations. Feedback on exercises on reasoning about data also helped me, for instance by the discussion with other students. Seeing their work helps my own thinking. Seeing my own work on the screen isn't scary, but can be confronting. But I think it's good, I have this feeling that students will be more serious in making their homework. But I didn't talk about this point with others. This way of working makes me learn maths better than just with the textbook. I usually discuss maths exercises quite a lot, but I think that counts even stronger for working with the network, because you now can see it from everyone.

Conclusion on students' perception of S1

The interviewed students were positive about this way of working when it came to feedback, notwithstanding the fact that in this case study we experienced multiple technical problems. Both ASS as DL were supported better than during more traditional instruction. The feedback process was not perceived as threatening. The social aspect of this kind of learning was stressed as being positive.

Students from S2A

Student 1

I was more motivated to do my homework, because of the fact that the teacher checked it. When working normally with notebooks, this is not the case. You had the feeling that you *had* to do it. That's good, though not very pleasant. But I had a good result on the test. That stimulated me, although I consider completing homework to be one's own responsibility.

LivePresenter feedback with respect to performing calculations I consider to be useful. This improved my mastering of these skills, because it's more efficient.

Discussing exercises on reasoning about and with data was useful in order to understand those exercises better. And the students had completed their homework because it was

sent to the network and the discussions were discussions about mathematics and not just chatting with each other as usual. Although most of the answers of other students I could have formulated myself, you do learn from seeing them. So you learn looking at another way.

I didn't consider it to be scary to see my answers on the screen. But there were some more or 'funny intended' answers made. So, I would like to see who gave those, thus not anonymously.

Ideally, I would say that 50% of the time should be devoted to this learning activity. I'm not sure that this would improve directly the students' results, but it would certainly be more fun, to look at each other's answers. I would learn more about reasoning, if we worked this way. So a better collective learning.

Looking back, it was more fun working this way, the homework was performed better; the results on the test were good. But the feedback sessions were very long. That's my only critique.

Student 2

Discussing reasoning with and about data would be useful. We didn't really do it in our group. In other lessons, this was part of the routine. When working this way, students were forced to complete their homework, what normally is not the case.

LPF with respect to ASS is useful, but for me had not a specific added value because I usually ask peer students about these skills and this usually works. For lazy students this could nevertheless be useful.

It wasn't scary that my responses were public; I wouldn't prefer anonymous contributions for myself.

Learning to work with this specific handheld could be useful in order to work with this network for the sake of feedback. For the future, this way of working could be useful for correcting homework.

Student 3

Because it was possible to check the students work, you are more forced to do your homework. I have perceived that as an incentive to do so. I feel taken seriously. The teacher is better informed about me and that is positive.

The transition to the Nspire wasn't very pleasant for me. The handheld was difficult to master. I missed the walking through the class of the teacher, supporting me. Feedback on how to perform the calculations should be personalised.

Feedback on reasoning with and about data was difficult. It partly helped me in mastering this reasoning, but everything went faster than in normal lessons. The discussions themselves were valuable. The classroom discourse was livelier. This could contribute to better learning of mathematics.

The public presentation of my results isn't really annoying, but other students could have a laugh on those who are wrong. Although this didn't work out that way. I wouldn't prefer my contribution to be anonymous and I like it to see what others have answered.

I see a future for this way of working because it's faster. In the longer run you'll master the device more easily.

Conclusions on students' perceptions of S2A

Students were more active than during more traditional instruction. They did not consider this way of working as more threatening. Feedback sessions were long, which made some students miss the teacher walking through the class, individually supporting the students. Feedback, delivered this way, in general, was perceived positively. The interviewed students foresee a positive future for this learning activity.

Students from S3

Student 1

The Nspire handheld was useful because you could make your homework everywhere, but sometimes it took some time to complete the send and receive process of the files using the network.

Splitting up the sections: one should have the discipline to check the answers oneself. But I spent more time on my homework, partly because you can just do it on the handheld: no textbook, no notebook, no pen, no pencil, and no answer book, just the handheld.

Feedback was improved because answers were shown from other students and incorporated in the discussion. Statistical skills I have thus better learned. Reasoning with and about data could have been stressed even further. Not all of the students' answers were always included in the discussion. In the old way of working, it's always the usual suspects discussing the exercises. That was now better. My own share in the discussion was bigger. There was more discussion. That contributed to the learning.

A class is a closed community, so I don't fear my name on the screen. It's safe enough. The same counts for my classmates.

In the future I would like to see this more incorporated. Completing homework is easier and the group is committed more to the learning process. There were still some technical issues, but the teacher managed the process fluently.

Sometimes it was hard to pose questions to the teacher, so make sure that students are able to pose a question, maybe even before collecting the files.

Student 2

In the beginning there were some organisational problems. But I perceived more feedback than usual. In the classroom it was sometimes chaos, which negatively influenced the quality of the feedback.

When using the textbook, you can easily read what has to be done and on the handheld that was a little difficult. But the hand-out on paper was useful. The fact that the calculations on the handheld were easily shown on the screen was an advantage.

Reasoning with and about data is improved by the feedback and the discussion, when working the usual way, this doesn't work that well, because some students are less interested and don't say that much. Discussing this way helps me better with learning this statistical reasoning. Students were more committed, more incorporated into the learning.

In the beginning, this was a little bit scary, I thought I had given stupid answers and was afraid that they were shown on the screen. My answers weren't that stupid, but the exercises were quite different from what I was used to. I would still prefer to contribute anonymously. Answering teacher questions when working as usual is somehow a little less scary.

The handheld itself was difficult in the beginning, but it was very useful that you could use it both as a computer as well as a calculator. He was a little bit slow sometimes.

In the future, when the performance of the calculator was improved, I certainly see a role for working this way in the usual classroom. This would help me to master the statistical calculations. Reasoning with and about data would also improve. 60% of the time could be dedicated to this learning activity, and then 40% of the time should be dedicated to self-supported learning. Like we did in the second part of the project. Then I would learn more mathematics from my classmates.

Student 3

I spent more time on my homework, because everybody could see what I've done. The teacher is now in control and that helps. I got more feedback, because by discussing the exercises in the group, I just better understood the mathematics.

The self-control sections were easier, the handheld feedback was usually enough, although this was just ‘right/wrong’ feedback. Some small explanation on the handheld would be useful.

The feedback helped me in mastering the statistical skills, for example by the on screen instruction of how to use the calculator for specific calculations. For the reasoning with and about data, the feedback and the discussion also helped me a lot. This discussion was broader than usually, when always the same students answer the teacher's question and others just copy the answers from the blackboard.

I felt safe enough to contribute to the discussion. It was not scary that others could see my name with my answers. And I consider it to be useful that I can see who answered what.

In the future I consider it to be useful that this way of working was used more frequently in mathematics education. And in other types of education. The keyboard (AS1D) was rather difficult to master, so I would strongly prefer a QWERTY-keyboard. So I would recommend implementing it.

Splitting up the sections was useful because you could then ask the teacher a question. Before this splitting up, the sessions were that long that you think “Yeah right, whatever”.

Conclusions on students' perceptions of S3

Students perceived more and better feedback, both on ASS as on DL. They felt safe in contributing to the classroom discourse. Splitting up the sections in 'self-control' and 'teacher feedback' improved the lessons. They advise dedicating about half of the time to this teaching activity.

Students S4A

Student 1

Homework attitude has improved, I spent more time on it, the teacher checked our work much more often, which made me spend more time. I would have had the same feeling when the teacher would have taken in my notebook after each lesson.

I have got more feedback on my work, because it is so easy for the teacher to give plenary explanations, and I also profit from the feedback he gives to others. The teacher is really able to see what I need, so the specificity of the feedback was also better.

ASS: GSS and LPF helped me personally not very much with learning the skills, because I was able to figure it all out by myself, but when I did something wrong, it was easy for me to see what went wrong.

DL: discussing on reasoning behind the calculations, that helped me somewhat, but most of it, I knew already. Although there were some exercises that were pretty tough. But this way of working is really new. My contribution wasn't bigger than usually, I always have an opinion. The fact that the responses were projected could encourage the participation of the shyer students, but I didn't need that. In our group, the usual suspects were dictating the classroom discourse. Although I must admit that the teacher tried to broaden the discussion. That made the students more active. Students were more focused on mathematics.

I don't consider it to be scary, but I can imagine that for someone shyer or more insecure about his or her answers that it is a little threatening. With the names of the students on the screen I don't have any problems. Students do feel more responsible for their work.

In the future I think working with a classroom network should be embedded in the lessons. About one third of the time, I think, should be dedicated to this teaching activity. Then we would learn more from each other.

Student 2

I have spent more time on my homework, because I liked to complete my homework on the handheld. The only disadvantage of the handheld is that it was sometimes hard to find something back. Therefore I appreciated the booklet. Discussing the homework with the group didn't influence my behaviour.

I have got more feedback through the classroom discussion which gave me more information about my performance. Not only informed by the teacher, but also by the other students.

DL: discussing reasoning behind calculations helped me in my learning, because you can see what others have answered.

ASS: LPF and GSS were helpful in learning how to perform the calculations, so I learned better and faster how to do things like that. My own contribution to the classroom discourse was bigger than usual. This helped my learning. I was more active.

I don't consider it to be scary; I don't care about the projected answers, because everyone makes mistakes.

In the future, for the mathematics lesson, this way of working would be a useful addition. I think that about 25% of the time could be used for this teaching activity.

Student 3

I didn't spend more time on my homework. It didn't motivate me; working with the handheld was initially difficult for me.

I didn't perceive more feedback on my work. I missed the contact with the teacher; he himself didn't understand it all. Even at the end of the chapter he wasn't good enough in the technology.

I doubt whether even in the long term this would work for me. I prefer normally working on the blackboard.

Discussing our answers group wise helped me in understanding DL activities. GSS and LPF also helped me in mastering ASS skills, that was handy, I must say.

My contribution wasn't bigger than otherwise, I didn't like to be engaged in the discussion, the teacher tried to, but I prefer to be a listener. I wasn't more active, though some others were. Sometimes I had the feeling that the attention of the teacher was more with the computer and less with the students. The discussion was now more focused on mathematics, I must admit.

I consider it to be scary that my answers were projected, I would prefer anonymity. I am indifferent with respect to whether others are anonymous or not.

In the future, I see that a classroom network could be useful in mathematics education. I would learn especially from the discussions, which means from other students. If it's utilised each lesson, I would say use it no longer than 50% of the time.

Conclusions on students' perceptions of S4a

Students were enthusiastic about the improved feedback, both on ASS as on DL activities. They stressed the additional learning effect of collaborative work. There was one interviewed student – with low mathematical skills – who considered her public contribution to the discussion to be threatening. Although others considered this to be fine, this shows this way of working *can* be threatening. On average, they thought that about one third of the time during the lessons should be spent on this teaching activity.

Students S4b

Student 1

I haven't spent more time on my homework, perhaps even less, but that time was a busy period, from the time I did spend on homework, I spent most time on mathematics. So perhaps, it did make me spend more time on my homework. The fact that everything is shown on the screen motivated me to some extent.

I didn't experience more feedback compared to working in the traditional way. But we have discussed more exercises in the classroom, although we perhaps had less homework than before.

I believe, by the way, that this way of working will gain terrain, because it is really handy to have so much functionality in one machine. We will keep on using books, but less frequently.

With respect to DL, discussing about reasoning with data helped me to some extent, because we have done this more frequently than usual. Although I had expected it even more.
With respect to ASS, with LPF and GSS, these possibilities helped me in mastering the skills on the handheld, but understanding the ideas behind these skills, that I didn't learn better than in the usual way.

I didn't contribute more to the classroom discourse than otherwise. Although I must admit that I very frequently had the tendency to actively engage. So in my head I was actively occupied with mathematics. This could have been caused by the transparency of the discussion.

I don't consider it to be scary that my contributions were projected on the screen. To see my fellow students' names, to me it wasn't really useful. But I can imagine that this is useful for the teacher.

In the future, this way of working should be embedded in mathematics education, but it should be well prepared for a longer period. But I think this is inevitable in the future that some kind of network is introduced into the daily classroom. Tailored support by the teacher is then more within reach. I believe 75% of the time could be dedicated to this teaching activity. For me, both my learning with respect to ASS and DL will be improved, particularly because class wide cooperation is made possible.

Student 2

I didn't spend more time on my homework. It wasn't really much, so it was quickly done. I have quite some experience with Excel, which helped me perhaps more than others. I don't need to be extra motivated. If it isn't checked I would have still completed my homework.

I got more feedback than in lessons without a network, I knew better where I was in the learning process, this is even better than with an answer booklet. The feedback was even of better quality.

With respect to DL, I mastered this better, through the classroom discussion. I wasn't very active in these discussions, but I listened well, which was also productive. With respect to ASS, LPF and GSS helped me in mastering the skills. My contribution wasn't bigger.

It wasn't threatening at all that the names of the students were projected, and I think that the names of the students led to a productive climate. For me anonymity is not an issue.

It would be an improvement for mathematics education if this way of working was structurally embedded. I believe 75% of the lesson time could be dedicated to this way of working.

Student 3

The time I spent on my homework was bigger than before, because of the checking by the teacher. But also that my fellow students can see what I've done. I don't care that they can see my mistakes.

I think I got more feedback than usual, because the teacher was better informed, and he questioned those things that were not really understood. And this feedback was also more specific. Self-check sections were more difficult, right/wrong feedback for me wasn't enough.

With respect to DL, reasoning with and about data, this was far more stressed than in usual lessons, and this I consider to be meaningful and also to be a nice activity. I think this will be useful in professional practice.

With respect to ASS, LPF and GSS did help me in mastering the mentioned skills, I need explanation 10 times, but this was a step by step explanation, which is useful.

My contribution during these lessons greater? It depends on your own attitude. The teacher didn't address his questions personally. Usually, it was S1 who responded very fast, which hinders me in thinking. The teacher could manage this better, although he tried to. That means that it was the usual suspects reacting to the open teacher questions. The discussion pattern looked too much like the pattern in usual lessons.

I don't consider it to be scary that my name is on the screen, it can be handy to see who answered what, but perhaps most for the teacher.

I think this learning activity should be embedded in future mathematics education. I think that it is useful to spend about 50% of the time to this activity. I would be supported in learning better reasoning, because of the classroom discussions.

Conclusions on students' perceptions of S4b

Students agreed on the fact that there was more feedback on their work, both with respect to ASS as to DL, which they considered to be effective. They did not perceive this way of working as more threatening. On average, they advise spending about two thirds of the time during lessons on this type of teaching activity.

Students S2b

Student 1

More time on homework? I think so, yes, for me, because I like this way of working and then it's not hard to pay the homework more attention. But I'm an exception in this group. Other students didn't pay enough attention, and they didn't want to like the project. This attitude is quite common in this group, as I have experienced from grade 7 onwards.

I think I've got more feedback than during traditional lessons. The teacher will know more about the whole class and is able to give explanations on the problems. ASS: here there is more feedback possible through LP and GSS. That's handy. From a book, that's more difficult. Following these explanations with your own calculator, that helps in learning how to do it. DL: the discussions in the classroom perhaps helped me in learning how to reason. But I don't consider this to be a part of mathematics. I prefer exercises with concrete calculations. But perhaps my contribution was bigger than in traditional lessons. There were more students participating, normally there are two now there were six students engaged in the discussion.

I don't consider it to be scary. If you do your work properly, there's nothing to fear. And even if others would laugh at my answers, that wouldn't bother me much. For my part, anonymity is not needed.

But for a structural implementation of a classroom network in mathematics education, there will be support needed. The Nspire is a more complex machine than the TI84, so that will take training time. But it was worth it for me. I would like it therefore to be structurally embedded. Each lesson, 50% of the time could be dedicated to this teaching activity. This will help students in learning statistics. And discussing with each other will help. I myself was more active. Therefore it was easier to learn mathematics.

Student 2

I had to get used to the calculator and the questions were sometimes a little bit weird. Sometimes terms were used that we were not completely aware of. But after getting used to it, it went fluently.

Normally, I would have spent more time on my homework. But because this was the end of the year, I didn't. I wasn't motivated anymore. My grades were already determined.

I have got more feedback, through the classroom discussions. That helped my learning. It was more effective than in traditional lessons. ASS: GSS and LPF were helpful in learning how to perform calculations. You have more input. DL: discussing in the

classroom gave me examples of reasoning from which I could learn. My contribution wasn't bigger than normal. I wasn't motivated enough and didn't show much engagement. Normally, I would contribute more. Then I am one of the most dominant students.

I don't consider this way of working to be scary. There is not much difference with input through talking. I consider it to be motivating that the students' names are shown.

When structurally embedded, both ASS and DL will be enhanced, because you have a more intense discussion. I see the teacher as the conductor of this all. On my behalf, the whole lesson may be spent on this teaching activity.

The student engagement was greater than usual, which may sound odd.

Student 3

The calculator made it more confusing; I felt the need for some kind of notebook. Discussing these matters with the group didn't really make a difference, because other students don't complete their homework either. The teacher couldn't explain things right; she made mistakes of her own, which didn't motivate me. She just wasn't a good role model; she lacked control over the group.

In principle, this way of working gives more possibilities for feedback, because she can check the students' work more easily. This also makes better feedback possible. But this wasn't realised in our group. With is partly due to the teacher.

ASS: LPF and GSS offered a possibility to learn the necessary skills better, because you can not only learn from the teacher, but also from each other.
DL: discussing and reasoning, even about opinions, could possibly support learning. For instance, someone who thinks that it was warm discusses with someone who doesn't think so. But in our group, I didn't really hear a good discussion. I pity that, because I would like to have classroom discussions, because then we wouldn't have to work at that time. Although discussion perhaps belongs to mathematics, we are not used to that.

I don't consider it to be scary; if my faults are projected on the screen that's not a big deal.

In the future, if this is structurally embedded in mathematics education, I would really like a notebook. But the network has an advantage, to see what others have done, I would appreciate that. Discussing in the classroom would be instructive. I think that half of the teaching time would be appropriate.

Conclusions on students' perceptions of S2b

The students were quite negative about the lessons during the intervention, but they were positive about the potential of this way of receiving feedback, both with respect to ASS as to DL. One student specifically attributed the low quality of the lessons to the teacher. One other student blamed the group for a bad attitude. In section 7.11 we describe some classroom discourse examples that illustrate the problems between the teacher and group. The students had no problems with their contributions being made visible in front of the teacher and their peers. On average, they advised to dedicate two thirds of the lesson time to this teaching activity, suggesting it was not the way of working they disliked.

7.4.4 Conclusions on students' perception

Students in the six participating groups were positive about the improvement of the feedback. Usually, they perceived greater, and more specific, feedback, both on ASS as on DL activities, when compared with more traditional lessons. There was not much

difference in appreciation of the feedback between the groups, except for the fact that the students in case study S2b were negative about the lessons themselves. Neither do we notice a structural difference in the experiences of the students as seen from their capacities in mathematics. The students stressed the contribution of collaborative work to their learning. They did not experience this way of working as more threatening when compared to lessons without a classroom network, except for one interviewed student (out of 18) with weak mathematics skills.

7.5 Overview of the implemented feedback

We now present an overview of the implemented feedback, classified by the feedback coding scheme as described in section 7.3 , as expressed in correspondence scores. This score can vary from 0 to 7, indicating the similarity between the intended curriculum (HTT) and the implemented curriculum (as shown in the video-taped lessons). In Table 7.2 we present an overview of the mean correspondence scores (with standard deviation and percentage of missing feedback sessions) of the intended and implemented feedback in the six case studies. In sections 7.6 -7.11 we will illustrate the lessons behind these scores.

Table 7.2 Correspondence score characteristics of the cases in C3

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL-M_ASS	%-Missing
S1	5.14	1.70	4.89	5.60	-0.71	68.09
S2A	4.40	2.29	3.66	5.71	-2.05	14.89
S3	5.38	1.85	5.34	5.46	-0.12	0.00
S4A	3.60	2.13	3.65	3.44	0.21	31.91
S4B	3.89	2.09	4.11	3.33	0.78	34.04
S2B	2.04	1.70	1.95	2.25	-0.30	59.57

The columns respectively present: name of the case study, the mean total correspondence score, the standard deviation of the total correspondence, the mean correspondence with respect to data literacy exercises, the mean correspondence with respect to algorithmic statistical skill exercises, the subtraction of these means and the percentage of missing feedback sessions as compared to the HTT.

What are the most occurring characteristics of the case studies, as represented by Table 7.2?

First, we would like to stress the very high percentages of missing feedback sessions, as compared to the HTT in the first (S1) and the last (S2b) case study. In the other four case studies this percentage was 34% or lower, so 1 on 3 missing, or less. These high percentages occurred as a result of two completely different reasons: in case S1 we encountered severe technical problems (as we describe in section 7.6), but when the technology was up and running, the teacher managed to implement the curriculum rather close to the intentions. In case S2b the relationship between the group and teacher was problematic. This was discouraging to the teacher to the extent that she could not really

implement the intended curriculum. In section 7.11 some illustrative examples of classroom discourse will be shown.

In the following sections (7.6 -7.11 we present some examples of classroom discourse as implemented during the use of the prototype in six successive case studies. How did we sample these examples from the abundance of classroom discourse we collected and analysed? First of all, the selected example had to be *substantial*: that is, it had to contain one or more events that are interesting from the perspective of our research question. This criterion for sampling makes it likely that the correspondence score of the selected events is higher than the mean correspondence score of the case study.

Further, we selected examples concerning mainly DL. As mentioned, we consider ASS to be very important but the instruction of DL is perhaps even more tedious. Improving that is a main concern of this study.

Then, the selected example had to be 'somehow representative' for the case study as a whole. That is, the classroom discourse had to contain elements that were more or less typical for the specific case study. Hence, we decided to present two examples of classroom discourse; one example would be too arbitrary, and whilst three examples per case study would have been even better, this chapter would have collapsed by its size.

In order to make a comparison of the case studies more valid, we tried to find an exercise on DL that gave rise to substantial classroom discourse in all of the six case studies. Due to the considerable problems in the first and the last case studies, this was not realisable. The classroom discourse that sparked from the feedback on exercise 8.8 came, over six case studies, most close to this criterion. With this exercise, asking the students to draw a conclusion about the computer behaviour of boys and girls, we have a point of comparison for the first five case studies, which were the five best out of six with respect to the mean correspondence between the HTT and the implemented curriculum. Besides this, this exercise was an element of the last part of the prototype, where the first part of it had happened to be represented by other examples of classroom discourse.

7.6 Feedback in the first case study S1

7.6.1 Starting point and overview S1

Starting point as perceived by the teacher

Below we present the starting point with respect to this first case study (called S1), as perceived by the teacher and reported in the first part of a questionnaire.

The teacher was male, 40 years old, and had 15 years of experience. He described the relationship with the group as 'moderate' (2 on a 4 point scale). He stated he did not really bond with the group, as if he spoke a different language to them. The level of the class he rated as 'adequate' (3 on a 4 point scale). The level was not high, but that is the level he expected for senior secondary education grade 10 mathematics A nowadays. His own ICT competence he described as 'good' (4 on a 4 point scale) because he was an experienced user of ICT in multiple ways.

Correspondence between HTT and implemented feedback

In Table 7.3 below we present the overall correspondence data for this case; for an explanation of the items, see section 7.5 .

Table 7.3 Correspondence score characteristics of case S1

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL-M_ASS	%-Missing
S1	5,14	1,70	4,89	5,60	-0,71	70,21

Two features of the first case study require attention: a high mean correspondence score and a very high percentage of missing feedback sessions. To start with the latter, this was the result of this case study being conducted with an extremely new version of the technology. The hardware and software used were the first available in Europe (as actually was the case in the 2004 preliminary study with the same teacher) and the first lessons during the first case study were observed while the technology still was to be presented during a conference in the U.S.A.. However, the lessons were planned, so we did not have much choice. There were a lot of technical problems, most of which we were able to tackle during the case study, but unfortunately few of them in a timely manner. Some of the problems are still open. For example, why could we not establish any wireless connection in the classroom in which the lessons were originally planned (which resulted in a first lesson without network support)?

Another reason for the extremely high percentage 'Missing' was caused by the teacher sometimes deciding not to use the network, because he wanted to cover his learning goals and experienced delay when working in the 'new mode'.

The teacher had to be very flexible, because he was supposed to act in a new teaching environment in a more or less new way – he had experience in 2004, with a system similar in concept, but different in technology, which probably reinforced for him that the technology was not yet reliable. During the concluding interview he described this as "*a feeling of having to learn teaching again*".

Besides these problems, there were some problems in the relationship between the teacher and class, as we presented in the starting point. The teacher himself described this relationship as "*moderate, as if I speak a different language to them."* He characterised the capacities of the students as "*rather low on average, but normal for this type of education*".

The correspondence gap between DL and ASS feedback sessions was not really big: -0.71, meaning that ASS-sessions went somewhat more as intended than DL-sessions. This could be caused by the fact that for conducting the classroom discourse, a teacher has to feel 'comfortable and safe' and needs a certain extraversion. The latter is not likely to be the problem, as is illustrated by the selected classroom discourse examples.

When we abstract from the technical problems – a big step, as mentioned, because this was a dominant factor during this case study – how is it possible that the feedback sessions that were established scored on average 5.14 on a 7 point scale, with respect to our HTT?

We now present some examples of feedback typical for this case study that can possibly suggest an answer to this question.

7.6.2 Classroom discourse example S1-1

Exercise 4.6 (DL, MC)

On the next page there are shown two sets of data: data_set1 and data_set2.

1: What data collection has the largest range?

2: What data collection has the largest spread?

HTT Ex 4.6

After having inspected the students' responses using ClassAnalysis, the teacher has to answer this question: Are the majority of the students convinced that a data set with a bigger absolute range can have smaller variation? We expect not, thus teacher feedback is needed.

Coded HTT Ex4.6 S1

Ex 4.6 *SLG*=DL → *CAF* (m, DL) → *SII* (m, m, DL) →*TC* (m, DL)

With SLG= student learning goal, DL=data literacy, CAF=class analysis feedback, SII= students' interaction input, TC= teacher's conclusion.

Implemented feedback Ex4.6 S1

[0:00)]

T: “Now there were given two data sets, data set 1, data set 2. Which of these data sets has the largest range? And second question: Which data set has the largest variation? Answers you gave, oh, there is more variation.”

Students: “Ouch!” The slide shows students' answers: 7 correct, 7 wrong, 6 no reaction.

T: “All of a sudden, people didn't dare to choose, or were too lazy to choose. Someone of those who did it right... ”

Teacher switches to the student view of ClassAnalysis. He questions two students, but both of them admit that they just guessed this multiple choice item.

T: “S1, you answered this question correctly?”

S1: “Yes, well I just took a look at the umm variation.”

T: “Yes.”

S1:“At the first, it was 21, but in the middle there was very often 11, and at the other there was from 2 till 20 I think, but there were in between a lot of different numbers.”

T: “Okay. So, these two data sets, have the same range, is that what you say?”

S1: “No, the first one has a larger range. And the second just normal.”

T: “What is the minimum of the first data set?”

S1: “1”

T: “And the largest value?”

S1 and several others: “21”

T: “So what is the range?”

S2: “20”

T: “20. The second data set, the smallest value is...”

S1: “2”

T: “2. The largest value is...”

S3: “22”

T: “22. So that has a range of…”

S3: “Also 20.”

S1: “No, the biggest is 20.”

T: “The biggest is 20 and the smallest is...”

S1: “2”

S4: “18”.

T: “So, the second has a smaller range, is that right?”

Students discuss this.

T: “The range, S5 is saying it now, is simply the difference between the largest and the smallest number. But, the second question is: Which data set has the biggest variation? What do you think, where do you find there is more spread, where are the numbers more spread out?”

S3: “The second.”

T: “The second. That's what S1 answered. And that's right. Because the first data set, it's difficult to show them very quickly...”

He opens the teacher software on his computer to show the original exercise.

T: “What S1 states, is right, I think, that the second data set, that there the numbers are more spread out. Now, hurry up.”

He waits for the software to start up.

T: “You see you need a lot of patience, which fortunately you have. S6. And so have I. [It takes another couple of seconds.] Shit! [Not very loud, but because of his wireless microphone loud enough to be recorded] Well, if S1 is right, and the computer told us she was, then the variation in the second data set is larger than in the first.”

He opens the file with the exercises.

T: “There they are, the two data sets, take a look on the screen, S7, the left column has got a minimum of 1, they are already ordered by size, then there are a lot of elevens, I walk down the column, and then it suddenly is 21. So the range is 20, 21 minus 1. The second set starts with 2 and the maximum is 20, range is smaller is 18. But do you really think that the variation of this one (points at the second with smaller range) is smaller than of this one (points at the first with larger range). Do you think so?”

S8: “Yes.”

Teacher looks very surprised.

T: “Wait a moment, fingers, who says 'yes, the variation here (first data set) is larger than here (second set)'? And who says: 'the variation here (second set) is larger than here (first set)'? These are far more different values? Here there are almost just elevens, except for an outlier on the upper side 1, and an outlier on the down side, what was it? 20, 21? Two outliers, but all of the rest has the same value. So, the measure of spread, the range, isn't that good as a measure. Because here, the range was larger here [he points at the first set than it was here [he points at the second set], while we agree on the fact that the variation

here [he points at the second set] was really larger. The numbers are far more scattered. So, remember that: the range is a very, very fable measure of spread."

[07:37]

Coded implemented feedback Ex4.6 S1

Ex 4.6 *SLG*=DL → *CAF* (l(1), DL) → *SII* (h(4), h(10), DL) →*TC* (h, DL)

Correspondence score: 6.

Interpretation classroom discourse Ex4.6 S1

Which specific teacher behaviours do we notice here with respect to feedback and the classroom network? He first uses CA in order to inspect the results in a general way. Apparently, the spread in the students' results strikes both the teacher as well as the students. Then he traces these results back to individual students [1]. But he is unfortunate in that the first two students he asks to explain their – correct – answer admit that they had been guessing at this multiple choice item [2]. The teacher, being surprised by this possibility, chooses a third student, and asks her to explain her correct answer [3]. She does this very briefly and her answer is not completely understandable. The teacher then challenges her by paraphrasing her wrongly. She tries to make herself clearer, but her formulation "the second just normal" does not really clarify her answer. The teacher then decides to take a more directive approach with the algorithm for determining the range (maximum – minimum) in mind [4]. But the dialogue between teacher and S1 gets interrupted by S2 and S3, trying to contribute. This interruption is not taken positively [5]. S1, nevertheless, keeps cool and answers the questions well. The teacher can then conclude that data set 1 has a larger range than data set 2.

The teacher then switches to the second question that has more to do with DL than the first part. The teacher addresses the question "Which data set has more variation?" to the whole group [6]. He gets an answer to this question without further motivation. He then decides to show both data sets on the projection screen, which takes some time [7]. He gets irritated, probably reinforced by quite some previous technical troubles.

He nevertheless succeeds in making a joke again, this time about his impatience. When he has both data sets on the screen, he describes the difference in variation and rephrases the question to: do you think that the data set with the largest range has the largest variation too [8]? A student answers "Yes", apparently to his surprise. He does not ask "Why?" but he tries to start a poll by raising fingers. He has used Quick Poll for this sake before, but now seems to have reasons to do it the old fashioned way [9].

In the procedural specification we stressed that each feedback session should be ended with a conclusion. We therefore agree that there is a point at which the teacher has to take over the discussion to make a statement [10]

Remarks on the classroom discourse Ex4.6 S1

[1] Good start of the feedback session by using CA and track some responses down to the students who were responsible for these responses.

[2] Correct student guessing can be a negative consequence of the use of multiple choice questions, hence use of this type of activity should be carefully considered.

[3] The teacher does not get distracted, but persists in his strategy of letting students explain their answer.

[4] In an interaction with S1, the teacher is directing her to an answer, with the definition of range as a guide. Instead of asking her "What do you mean with 'the other just normal'?"

[5] The teacher could at this point have been more specifically directive towards other students than the one with whom he is interacting. In the evaluation of the third case study we will see that the responsible teacher is more directive while conducting the classroom discourse.

[6] Again whole group addressing.

[7] The fact that it takes some time to show both data sets on the screen has not got to do with the teacher's ICT skills (he rates them as 'good', out of 'low', 'moderate' 'adequate' and 'good' and this is exactly what our observations show). It could have something to do with the fact that he has not got enough specific experience with this system yet (in the concluding interview he estimates that he would need about six more lessons in order to master the system really well), but it in our view certainly has to do with the usability of the software, in this case: the big effort it takes to switch between exercises, results and different representations of these results.

[8] Representing both data sets on the screen (using LivePresenter) has a focussing effect. And his rephrasing of the question "Do you think that the data set with the largest range has the largest variation too?" has this effect too.

[9] It is perhaps a pity, most likely to be explained by the combination of his lack of experience, the cognitive load he suffers at that moment (he rated the sum of his tasks during a typical feedback session at one and the same moment as 'extremely stressful') and his irritation about the technology letting him down far too often. But the fact that he shows a 'poll reflex' is from our research point of view interesting. He thus creates a situation in which he is able to continue the feedback session meaningfully. Unfortunately, he formulates the poll ambiguously which is too complicated. Perhaps he sees these weaknesses too, because he does not execute the poll. Instead, he starts lecturing (which Cobb (1991) advised to avoid).

[10] We consider that the teacher in this example switches too fast to the 'statement mode needed for a conclusion'. He states "*The numbers* [in the second data set] *are far more scattered*". But a key question here is: scattered from what? Apparently, there is some kind of reference. What would that reference be? Posing this question to a student would enhance chances to come even closer to the core of understanding what variation really is. It would be a great bridge to the concept of standard deviation, which is the leading concept in the next section. The standard deviation considers variation as some kind of 'mean distance' that elements of a set have with respect to the mean of that set. The mean of a set is thus a fertile reference for considering variation. The teacher chooses not to use this possibility to deepen the discourse this way, although there was a chance.

7.6.3 Classroom discourse example S1-2

Exercise 8.8(DL, OR)

What do you conclude, using these data [mean, S, IQR and range of the boys and of the girls], about the computer behaviour of 12-year-old girls and 12-year-old boys in 2000?

HTT Ex8.8

Suggested response: “Boys use the computer on average more than girls. The variation is also bigger with respect to IQR, SD. The range, on the other hand, is bigger for girls. When inspecting the box and whisker plot again, you see two typical outliers amongst the 400 girls presented, responsible for the inconsistency in the measure of spread.”

In sections 1 and 2, the students had to explain differences just based on a measure of centre (mean). Now they also have three measures of variation (inter quartile range, range and SD) at hand.

They should include this in their comparison of the behaviour in computer use of boys and girls. The teacher collects and checks the students' responses. We expect that most students will not include the variation of both data sets in their comparison. The teacher then is to give feedback on the enhancement of the comparison when including measures of variation, if possible with selected students' responses as a starting point.

Coded HTT Ex8.8

Ex 8.8 *SLG*=DL → *CAF* (m, DL) → *SII*(h, h, DL) → *TC* (h, DL)

Implemented feedback Ex8.8 S1

[10:14]

T: “If you're gonna switch between the data of the boys and the girls, and then you look at the mean and the standard deviation, then you have to choose between the following answers (he shows the ClassAnalysis file and then sees a list of students' answers) oh no, you had to formulate the answer yourselves, let's have a look 'that the booys, ha, typo, that the boys use the computer more, was answered by one of you, can I see quickly who that was? By clicking on it, or something, no, that's not possible, I think, no. 'That the girls on average work less with the computer than boys' that says this person too 'Boys more often', that's the same as 'Girls on average less', er 'That boys play more with computers', that's still a question of course, perhaps they produce a lot of theses, or do they intensively use spreadsheets, 'Girls less frequently use it', 'Boys use the computer more than girls, 'Boys use the computer more than girls', the last one is better than the second last, with respect to grammar, 'Boys on average longer with the computer', 'Boys work longer with the computer and girls less'.

[11:20]

T: “The most remarkable is that...”

S1: “Everyone [inaudible]”

T: “...you had to calculate several measures, and everybody answers about just one of these.”

S2: “Yes”

Teacher is silent for a couple of seconds. He walks to the whiteboard and points at a certain phrase on it.

T: “Is there no difference in the variation between boys and girls?”

S3: “Well...yes”

T: “What was the standard deviation of the girls, that was 4.01, that's a measure of spread, that says something about how it varies, and of the boys, it was 8.36. Yes. And no one

mentioned that. If you had to do that still, could you say something about the difference in variation between the computer use of boys and girls?"

[12:30]

S4: "The variation among the boys is much bigger than among the girls."

T: "Well, this was the SD from the boys, 8.36."

S4: "But that's bigger, isn't it?"

T: "Is 8.36 bigger than 11?"

S4: "Than 4, wasn't it?"

S5: "4 it was"

T: "Was it 4?"

S4: "4.01"

T: "Of, I'm sorry, you're right, sorry. Yes. And what does this mean, that the variation of the boys is a lot bigger than of the girls? What does that mean, in plain language?"

S6: "That the mean is also different."

T: "The mean is different, on that we all agreed. On average the boys work more often, longer with the computer than girls. But what does it mean that the variation amongst boys is bigger?"

S7: "Bigger numbers."

T: "No, not that the numbers are bigger."

S4: "That the differences among boys are bigger."

T: "Yes! That the difference among boys is bigger, so that you for instance have boys that are gaming all night long, while others haven't touched a computer in their life. At the same time, with girls, the differences are a lot smaller, all the girls use MSN one hour a day, for instance. As an example. Or two hours. While among boys there are the really addicted and boys that hardly use the computer. That could be an explanation for the difference in range [he probably means 'variation']. That spread, that standard deviation was much bigger amongst boys than amongst girls. And with the interquartile ranges, you saw that too, that the differences were bigger."

[13:57]

T: "So what's most striking is that you, those who answered this question did very well by inspecting differences with respect to the mean. But if you'd done it very carefully, then you would have mentioned something like 'the differences amongst boys are also bigger than amongst girls. The mean doesn't always tell everything. If the mean test score in this group on the next test is a7 and in the other group it will be a 7 too, that doesn't necessarily mean that the test is made exactly the same. It is possible that in this group there will be 10's and 1's and in the other group just 7's as grades.

Then the variation in this group will be much bigger than in the other group. So, if you really want to look at boys and girls, don't just look at differences in mean, but also at differences in variation."

[14:49]

Coded implemented feedback Ex8.8 S1

Ex 8.8 *SLG*=DL → *CAF* (h(9), DL) → *SII*(m(3), m(5), DL) → *TC* (m, DL)

Correspondence score: 5.5.

Interpretation classroom discourse Ex8.8 S1

Between [10:14] and [11:20], the teacher quickly brings up nine students' responses and gives some very brief comments [1].

[11:20-12:30] He then asks an open question: what is most remarkable? He addresses this question to the whole class [2]. Apparently, he is not convinced that the students will come with a fruitful suggestion, so he starts hinting about the bias towards just using a measure of central tendency, without mentioning a measure of spread. He leaves a pause [3]. But again, he does not seem to have the confidence in his students' responses and hints even more strongly by rephrasing his question while using the term 'variation' [4].

[12:30]-[13:57]

This evokes a student to give the right answer. The teacher is then mistaken about the value of the standard deviation. But he quickly recovers, picks out the correct student contribution and translates it into a good prompt, to make 'variation' concrete in plain language, again addressed to the whole group [5]. This evokes two interesting but wrong student answers, both of which the teacher rejects without elaborating the why behind them [6]. The third student answer he is enthusiastic about and illustrates this with an example [7]. He further mentions IQR as another measure of spread that has a bigger value for boys than it has for girls [8].

[13:57]-[14:49]

After this, he draws a well formulated conclusion himself, illustrated by a well-known example [9].

Remarks on the classroom discourse Ex8.8 S1

[1] Productive use of ClassAnalysis. Mentioning these answers makes the problem perfectly clear.

[2] It is wise to translate and compress the students' responses to a question. It may be addressed to the whole group, before a thinking pause. But a considerable number of times a specific student should be invited to answer this question. This is not done here.

[3] He hints at the missing concept in the students' responses. And offers students some thinking time.

[4] Apparently, he does not expect too much from the input of the students, or he is just impatient, because he rather quickly mentions the key word here: variation. He could have posed other questions before this one ("What is the statistical measure that is used in all of these responses? To which class of measures does this belong? Is there another class of measures? What is indicated by this class of measures? Do both data sets differ with respect to this?"). Posing 'smaller' intermediary questions makes the identification of the students' Zone of Proximal development (Vygotsky, 1978) more subtle. This is in our opinion a very important skill of a teacher that wants to conduct a productive classroom discourse.

[5] Again a good translation of student input into a question that could be productive for the classroom discourse: how to translate 'variation' in the given context into 'plain language'. Again addressed to the whole group.

[6] Two student inputs are substantially set aside:
  6.1. “That the mean is also different.” The teacher rejects this idea. But given the fact that almost all of the students' answers were focussed on the mean, perhaps a conceptual discussion about the difference and the relationship between measures of central tendency and measures of spread would have been useful here.
  6.2. Perhaps even more intriguing is the student input from S7: “Bigger numbers.” There could be a conceptual misunderstanding about the presumed positive relationship between the size of the numbers and the size of the standard deviation. The teacher does not discuss this.

[7] Input from S4: “That the differences among boys are bigger.” S4 thus translates 'variation' into 'differences'. The teacher is immediately enthusiastic. He does not ask for comments from other students, but he illustrates this with some scenarios that could explain the observed differences among boys and the differences among girls.

[8] The teacher points at the fact that both standard deviation and the interquartile range are bigger for the boys. And concludes that the variation is bigger among boys. He forgets to take the range into account. When representing these data sets simultaneously as box and whisker plots with outliers, one immediately sees that there are two girls that cause the big range among girls. This, of course, is a concept for the next investigation: what does this say about the robustness of SD, IQR and the range?

[9] There were no student responses using a measure of spread when comparing the data sets for boys and girls. This is a good reason for not trying to link the conclusion of this exercise to one or more student answers, but perhaps the teacher could refer to the good answer S4 gave in the discussion.

7.6.4 Classroom discourse example S1-3

In this small example, 1:21 minutes of duration, unintended in our HTT, we see how the teacher fluently uses the network in order to reinforce a mathematics focused classroom culture.

Implemented classroom discourse example S1-3

[0:00]

The teacher inspects the ClassAnalysis file he just created by collecting the students' responses on section 8.

T: “I see here, some remarkable things. For instance, S1, who scored incorrect to all of these questions. S1? Any idea how that's possible?”

S1: “No”

T: “Did you complete the exercises, S1?”

S1: “Yes.”

T: “Yes?”

S1: “Yes.”

The teacher opens the student view of CA.

T: “We're gonna take a look for S1.”

Students are looking at S1, smiling, somewhat expectantly, somewhat nervously.

T: “This is section 8. About computing. Hmm. Well. Oh, this is S2. I thought so. S1, there you are, let's have a look. Score student zero, zero, zero, zero, response student, ah, I see why you have zero points.”

He points at the column of empty answers S1 gave. Then he looks at S1 and waits for a second. Silence in the classroom.

T: “You didn't do it.”

Students immediately laugh.

S1: “No, I haven't got 8.”

T: “You didn't complete section 8.”

S1: “That's right.”

T: “Why not?"

S1: “I don't know.”

Silence in the classroom. Teacher looks at Nick, but doesn't say anything.

S1: “I will make [inaudible].”

T: “Hmm. That, I will remember.”

Students are laughing again.

T: “On the other hand I see that S3…”

[1:21]

Interpretation classroom discourse example S1-3

There is no real mathematics education in here but we see a lack of the fundamental principles required to establish successful mathematics education: trust and student commitment. The teacher is performing a routine check on the students' work. His eye falls on the zero score of S1 and he perhaps has, besides this minimum result, reasons to doubt whether S1 speaks the truth. He personally asks this student whether he has completed the exercises. The student confirms that he did. The tone of the question the teacher then poses becomes inquiring. The student sticks to his 'yes'. After this, the teacher decides to check this answer in public. There is tension in the classroom. Students seem to realise that the trick S1 probably plays could possibly not be effective in the new situation of a classroom network. The teacher at first looks at the wrong student – S2, a conscientious student – he seems to doubt whether his suspicion against S1 was right, but then realises his mistake and finds the data he is looking for. He confronts S1 with that, in a straight, factual way. S1 immediately changes strategy and confesses. In a way, he seems to apologise for his behaviour, by making some sort of promise. The teacher reacts with a little scepticism, though includes humour before switching focus to other student responses.

What would be essential in this teacher behaviour in order to make the offered opportunities a success? First of all, the teacher has to be able to understand the support a classroom network can offer. Then he has to deploy this support with respect to the situation he is actually in. After that, he has to be able to personalise the feedback, even if this feedback draws out the performance of a student. A teacher has to be very careful here, because it all happens in public. Feedback design warned not to address feedback to the students' self, but here, it seems just and effective. It takes confidence to act like this:

being straight but fair. The teacher's fairness is reflected in the fact that he ends the dialogue with a joke and switches quickly and smoothly to another feedback issue.

In short, the teacher has to be a leader in the classroom and he must personally address students with respect to their learning performances. The teacher is able to do so with the support of the classroom network, but still he must act himself. This takes what we call *functional extraversion*: the confidence to dive into the data and to confront the students with his interpretation of these data. Besides this, it takes conversational skills to make this all productive in a classroom discussion.

7.7 Feedback in the second case study S2a

7.7.1 Starting point and overview S2a

Starting point as perceived by the teacher

Below we present the starting point as perceived by the teacher and reported in the first part of a questionnaire.

The teacher was male, 55 years old, 27 years of experience. He described the relationship with the group as moderate (2 on 4 point scale). The behaviour of a number of dominant students led to the teacher spending too much time keeping order. He considered individual attention to be important and felt that he was not able to pay this attention as much as he would like. He described the level of the group as moderate (2 on a 4 point scale). In the group there were some students who lacked potential but no more than usual. There were many students who should be able to meet the required level, but were completely unmotivated and therefore performed badly in the tests. He described his own ICT competence as adequate (3 on a 4 point scale). He did not consider himself as an early adopter of ICT in the classroom, but he nevertheless supported students through his own website. After some practice, he felt that he should be able to master the technology.

Correspondence between HTT and implemented feedback

With respect to the teacher's self-reported ICT skills, we noted that the teacher emailed with his students, used a web based learning environment and was programming his own website that requires knowledge of JavaScript. He coded these scripts in a plain text editor, which certainly defines him as an expert ICT user to our standards. But he was too modest to state this himself. This modesty was probably an important personal characteristic as we see later in this section.

Table 7.4 Correspondence score characteristics of case S2a

Case	Mean	SD	Mean-DL	Mean-ASS	M_DL-M_ASS	%-Missing
S2A	4.40	2.29	3.66	5.71	-2.05	14.89

When analysing the data of this second case study, two issues arise: the teacher was reasonably successful (4.41 on a 1-7 scale, SD= 2.27, percentage missing is 14.89) in realising feedback sessions as we had them in mind when developing the pilot. However, there are certainly some areas that could be improved from the perspective of our HTT. With some typical examples from the classroom discourse we try to illustrate what these aspects are. There is relatively quite a degree of variation in the scores (SD=2.27), meaning that there is some variation in the way the teacher succeeds in conducting the feedback sessions. There are some missing sessions, but not too many. Sometimes this was caused by the teacher choosing, for the reason of planning, to skip a session. Further,

a technical problem (a corrupt audio-layer in a videotape) withdraw us from following the implemented curriculum perfectly.

In the first case study we observed that in order to optimise a feedback session, the teacher needed (among others) to address feedback personally to the students. This was perhaps the main problem during this case study.

In contemporary mathematics education in the Netherlands, a substantial observation is that the teacher 'disappears in the classroom while supporting self-directed students'(This was mentioned in 2001). In the pedagogical architecture of this intervention the teacher has the means to become a conductor of the collective learning process again, while avoiding the pitfall of lecturing. At the same time, we utilise teaching support that can become an obstacle behind which the teacher can hide: the technology. The relatively big gap between the mean correspondence score on DL and this score on ASS could indicate that the teacher was able to conduct the classroom discourse with respect to the concrete ASS activities more easily than on the more vague DL activities. This was seen in C2 as well. The correspondence between the teacher in C2 and the teacher in this case study is their modesty and their poorly articulated "functional extraversion".

In general we observed that the feedback sessions took too long. There seemed to be not enough variation in learning activity for the students. Sometimes the feedback session took almost all of the time of a lesson. The teacher sometimes could not resist lecturing ("I had this strong feeling of having to give feedback," as he declared in the interview), but even if the feedback sessions were as interactive as we intended them to be, a lesson should have more variation than most of the lessons this case study consisted of.

7.7.2 Classroom discourse example S2a-1

Exercise 1.2 (DL, OA)

What do you mean by the mean maximum temperature of a given day (eg, September 3) over the years?

HTT Ex1.2

The students have to formulate their own thoughts on a statistical procedure. There are probably some problematic aspects for them. First of all, what is meant with "over the years"? How long should such a period be? Secondly, do they understand that they just have to take a fixed day into account (for instance, September 3)? And thirdly, the biggest conceptual problem is that the order of the operations ('mean' and 'maximum') does matter with respect to the result.

Coded HTT Ex1.2

Exercise 1.2 *SLG*=DL → *CAF* (h, DL) → *SII* (h, h, DL) →*TC* (h, DL)

Implemented feedback Ex1.2 S2a

[0:00]

T: "Well, we might like to look up someone to see his answer..."

S1: "S2!"

T: "Well, fine with me, but then I have to look up..."

The teacher inspects the Caf file with student responses that are projected on the screen.

T: “Uh, S2 ... it's alphabetical ...here her answer is: 'The highest temperature on a certain day ... which is averaged over several years.' I know someone who is very pleased with this. But is everyone pleased? Is this a good description ... we'll see what another has said ... It is very difficult to formulate. Wait, what did you just say, S3... "

Teacher scrolls through the answers.

S2: “Same!”

T: “Same? Oh, you did that task together, huh? Is there anyone who really had something else? S4? I hear S5. There he goes, S4 what have you answered? 'I consider the mean maximum temperature to be the warmest average maximum daily period of September through the years '... yes"

S2: “Fault!”

T: “Fault? Why that?"

S3: “Why should it be the warmest? That mean does not always have to be the same.”

T: “Say that once again, because I was distracted by S9.”

S3: “The mean temperature is not always the same at the same time of the day.”

T: “Does S4's answer state so?”

S3: “Yes!”

T: “The warmest part of the day .... that may not be the same part every day ... yes? Er... I just wanted to get to S6.”

S7: “He won't have it.”

T: “... Because S6 last time had a question about this. And if I remember correctly, S6, I have not answered you ... 'That is the highest temperature of the various September 3rds' ... Well, that's very clear and concise, nice and short, yet clear, the highest temperature of the various September 3rds, if you compare those to each other, is that what it is?”

[1:47]

S3: “But that's really a personal question, what do you mean, that really means ‘what do you think it is?’”

T: “Yeah, but ...”

S3: “Yes, there is a good answer, but I understood it quite differently ...”

T: “... I think there is a lot of chatting with neighbours ... and I would like to keep some focus because I think you are saying very sensible things, do you want to repeat it one more time?”

S3: “Yes, I think if you say 'what do you mean with', this can be something totally different than what someone else thinks, one can understand with 'mean' something else than another person ... there is a good answer, but ... ”

T: “Yes. Yes. You understand it differently than someone else. But ... that's actually the point where we want to be, with such a question, it is convenient that you understand the same by the same mean as I do. That question is tricky, however. We will indeed philosophise about this. What does this mean exactly?”

[4:01]

S2: “From each year, you just take the maximum temperature of September.”

T: “September 3.”

S2: “Each year, so between that, of all those years September 3, add them together and calculate the mean.”

T: “Yes, that seems correct.”

S3: “Yes, but that's what I meant too.”

T: “Yes. Yes, that was clear. S6! S6, can you remember, was this the question you asked me yesterday something about?”

S6: “I think so yes.”

T: “Yes. You had your finger on the right place. Uh, because we're talking about the mean temperature, but you actually said something to me yesterday, not heard by the others, not about the average of the maximum, temperature, but about the maximum average temperature. Are those the same?”

S2: “The mean is always adding everything together and then divide by something, right?”

T: “Yeah. But is the average maximum temperature the same as the maximum of the average temperature?”

S2: “No. No. The maximum of the average temperature is the highest of the average temperatures.”

T: “Yes. ... Yes, I'm sorry, I didn't follow you, I just wanted to ask S7, what do you think, is that the same?”

S7: “No.”

T: “No. What is the difference, could you indicate that?”

S7: “Uh ...”

T: “Yes it is difficult, so a little spontaneous, it's a bit of a word game, it seems, but if those words turn around, then it has a different meaning. If you look at the maximum temperature, the highest temperature of a day, and you calculate the mean, then you actually look at the mean of those very high temperatures, of the outliers, uh, when you do it the opposite way, look at the mean temperature, which will now always be much lower than the maximum, naturally, and you take the maximum of the average temperatures of each day, yes, then that temperature will be very different, much lower, very low. Or low, very low, but in any case lower than uh ... well let's continue.”

S5: “Is this also good, I had the maximum temperature of September3 for the last five years and the average of this.”

T: “Uh yes, can someone respond to this? Is that good? [...] You say, for the last five years.”

S5: “Yes, you may, you can take several years.”

T: “Well, for my sake, if you do statistics, hey there, I would like to know the mean of your age, well, for my convenience I just take the mean of you two and then I ‘know’ it [ironical intonation], you do have to include more numbers, take more September 3rds, to get a right answer.”

S2: “From every year. Every year, every September 3, add them together and divide that by the number of years that you have.”

T: “Yes.”

S2: “Because you want the maximum mean of 3 September. Thus, all September thirds, you add together.”

T: “Yes.”

S2: “The highest temperature of September 3.”

S2: “It says: for example.”

T: “Yes, yes, we can also take September 6.”

S3: “Yes, it does not matter.”

T: “For all September 3, add up the maximum temperatures, divide by the number of years you've looked at ...”

S3: “That's what it says in the exercise.”

T: “... Then you're done. We continue, folks, we shouldn't get stuck on just one exercise.”

[7:04]

Coded implemented feedback Ex1.2 S2a

Ex1.2 *SLG*=DL → *CAF* (h(3), DL) →*SII* (h(5), h(11), DL) → *TC* (h, DL)

Correspondence score: 7.

Interpretation classroom discourse Ex1.2 S2a

Checking the students' responses seems to trigger the students at least socially, as is witnessed by S1: “S2!”: a student would like to see the answer of a specific other student. The teacher, possibly realising that a good classroom discussion can be set up with more than one starting point, accepts this [1]. Collaboration (of S2 and S3) is easily seen (and admitted and accepted) because all the answers are simultaneously shown [2]. The teacher is pointing personally very well (“S4, what do you have?”) [3]. S4's answer, as an input into the discourse, makes S3 react very well, by stating that “The mean temperature is not always the same at the same time of the day.” Probably, she means “the *maximum* temperature isn't always reached at the same time of the day" instead of “mean temperature". The teacher then critically evaluates this input (“Does S4's answer state so?”) [4]. After this, the teacher switches to a crucial aspect of this exercise: the difference between 'mean maximum' and 'maximum mean'. He does this subtly by referring to an earlier discussion he had with S6. He is interrupted by S3, who, together with S2, were quite dominant in the classroom discourse according to our observation. This time, it is about an important issue: discussing the statistical opinions of their peers. After this, the teacher tries to get back to S6, but again he is more or less cut off by S2 again. He stays polite and the tries to involve S7 into the discussion. S7 hesitates [5] and then the teacher appears to surrender: he lectures about the difference between 'mean maximum' and 'maximum mean'. He does this quite well, but without student input [6]. S5 then intervenes, asking whether his approach is good. The teacher beautifully rebounds this to the whole class (“Can someone respond to this? Is that good?”). But he doesn't wait for students' responses and hints at the weak part in the students reasoning (“You say, for the last five years"), possibly again because he feels the time pressure [7]. S5 does not really pick this hint up either. Thus the teacher gives an illustration of the consequence of this reasoning: “I would like to know the mean of your age, well, for my convenience I just take the mean of you two and then I know it, so you do have to include

more numbers, take more September 3rds, to get a right answer." [8] S2 picks this up again well: "From every year. Every year, every September 3..." [9] The teacher agrees with this, repeats the good answer once again and continues with the next exercise.

Remarks on the classroom discourse Ex1.2 S2a

[1] Good use of CA to start the feedback session.
[2] The overview of students' responses activates the students.
[3] The teacher personally questions a student (S4), whose reaction evokes S3 spontaneously to a good remark.
[4] The teacher acts sharply to the good remark of S3 by asking whether S4's response satisfies S3's point. He could perhaps have asked this to another student than S3, in order to give the discourse an even broader support, showing that he is the one who conducts classroom discourse. The exchange of 'mean' and 'maximum' as statistical operations (S3's fault) will be problematised later on. Perhaps that is why the teacher decides not to do that here, although there was a nice occasion for it.
[5] We see that the teacher has a hard job in being the conductor of classroom discourse, with interruption for example from S3 and, especially, S2. The teacher and the students state in the interviews after the intervention that there were indeed many interruptions but that they were almost always mathematics related. This seemed to be different to what they were used to.
[6] There are several reasons why a teacher cannot always elaborate all of the students' suggestions. Possibly this occurs here because the teacher feels time pressure, as is witnessed by his remarks "Well, let's continue" and later on: "We continue, folks, we shouldn't get stuck on just one exercise". The art is to continue at the right moment. Here, the risen issue is fundamental: what's the difference between 'mean maximum' and 'maximum mean'? Possibly a little numerical example would have clarified this subtle issue of the order of the mathematical operations.
[7] See [6].
[8] A little personally addressed counter question like "Do you consider a mean over a period of five years to be reliable enough?" could have deepened the classroom discourse more than this good, but completely self-contained example. He could perhaps have tried to obtain, as a suggestion from one of the students, the dependence of the "sample size" on reliability. It is possible that his haste acts against this.
[9] S2 then spontaneously completes his answer by stating the use of "every September third".

7.7.3 Classroom discourse example S2a-2

Exercise 8.8(DL, OR)

What do you conclude, using these data [mean, SD, IQR and range of the boys and of the girls], about the computer behaviour of 12-year-old girls and 12-year-old boys in 2000?

HTT Ex8.8

Suggested response: "Boys use the computer on average more than girls. The variation is also bigger with respect to IQR, SD. The range, on the other hand, is bigger for girls. When inspecting the box and whisker plot again, you see two typical outliers among the 400 girls presented, responsible for the inconsistency in the measure of spread."

In sections 1 and 2, the students had to explain differences just based on a measure of centre (mean). Now they also have three measures of variation (inter quartile range, range and SD) at hand.

They should include this in their comparison of the behaviour in computer use of boys and girls. The teacher collects and checks the students' responses. We expect that most students will not include the variation of both data sets into their comparison. The teacher then is to give feedback on the enhancement of the comparison when including measures of variation, if possible with selected students' responses as a starting point.

Coded HTT Ex8.8

Ex 8.8 *SLG*=DL → *CAF* (m, DL) → *SII*(h, h, DL) → *TC* (h, DL)

Implemented feedback Ex8.8 S2a

[39:58]
T: “This is an important question. What do you conclude, based on these data, about the differences between boys and girls? Or are there perhaps no differences between boys and girls behind the computer? Yes? Well. Let's have a look at some of the answers.”

[40:19]

T: “You can conclude here that, as has been mentioned, that boys are working more frequently with the computer than girls. Is that true? Shall I act really severely? Why isn't this true? Perhaps that's a little dull, you mean probably 'longer on average'? Because there are people, who check their mail very fast, half a minute, then they wait five minutes, then they look again. These people use the computer very frequently, but not very long. So I think that boys spend longer working with a computer than girls. That's also with the second answer 'boys on average work longer with a computer'. Is that a correct conclusion?”

S1:“Yes.”

T: “Yes, it is, isn't it? The mean was bigger. Yeah, that's right. The mean was almost twice as big, 8.44 versus 4.01, did I say almost, more than twice as long the boys use the computer. But! There is also mentioned that 'With boys there's more variation, but with girls the range is bigger'.”

[41:39]

S1: “That's a real nerd.”

T: “That's a nerd? Well, it is in any case someone who is right. Someone who has thought really critically.” Some students laugh, probably because of this misunderstanding.

S2: “Who was that?”

T: “Oh, if I want to know that, I can take a look. But, that is of course, and I hope you see that, an excellent answer, because what happens, you see that the variation with boys may be bigger, how do you see that, well, the most handy measure of spread is possibly the standard deviation, which you calculated, or maybe the interquartile range, which you also calculated, both of them are bigger among boys than among girls, but, among girls there are, and that is a nuance of the statement that the variation amongst boys is bigger than among girls, that there are one or two girls that have been working with the computer extremely long. So the range is bigger among girls. But the range, that was the most unreliable measure, because it is determined by outliers, so the standard deviation or

the inter quartile range is a somewhat prettier measure. Let's take another look for some good answers… 'All data are pretty close' Hmm. 'The boys used the computer much longer'… If I take a look, I can't oversee it perfectly, of course, then I have the impression, what I had suspected, that a lot of you say 'Yes, but those boys they work on average longer with a computer'.

And now try to remember that that is correct, but that you have to mention something about the variation. Whether if they all work equally long, whether that varies more than amongst girls, from now on, if you inspect a data set, look both at the measures of central tendency as well as at the measures of spread."

S1 sings "Boys, boys, boys" (Lady Gaga, 2007)

T: "Aha. Is that a song of longing or how do I have to read it?"

[43:25]

Coded implemented feedback Ex8.8 S2a

Ex 8.8 *SLG*=DL → *CAF* (h(5), DL) → *TC* (h, DL)

Correspondence score: 3.

Interpretation classroom discourse Ex8.8 S2a

A remarkable feature of this 3'30" feedback session is that the teacher just restricts himself to the ClassAnalysis results. He picks out 5 student responses and gives feedback on them [1]. This is almost like lecturing, albeit with the students' responses as an input. This lecturing, the teacher does very well. He reads aloud the students' responses and is commenting on them in a style we would like to call 'internal dialogical aloud' [2]. Then he reads aloud the second answer 'boys on average work longer with a computer'. And asks: "Is that a right conclusion?" S1 answers "Yes". Then the teacher agrees and explains [3]. After that he finds an almost perfect student response: 'With boys there's more variation, but with girls the range is wider'. Then a very funny incident takes place. S1 comments 'That's a real nerd'. The teacher, possibly being used to students making negative comments about others, sees this as a comment to the student who formulated this correct answer. S1 was referring to the girl that caused the range of weekly computer hours among girls to be that big. The teacher firmly defends the student behind the last answer. Some students pick up this misunderstanding and snigger [4]. Then the teacher continues on his internal dialogical aloud with: "…you see that the variation with boys may be bigger, how do you see that, well, the most handy measure of spread is possibly the standard deviation, which you calculated, or maybe the interquartile range, which you also calculated, both of them are bigger among boys than among girls,…" [5] After this the teacher notes that most students were focused on just the mean and warns against forgetting to mention the variation of a data set [6]. He has not committed students personally to the classroom discourse [7].

Remarks on the classroom discourse Ex8.8 S2a

[1] Good start of a feedback session: check and comment the students' responses.

[2] An example of an internal dialogue aloud: "You can conclude here that, is being mentioned, that boys are working more frequent with the computer than girls. Is that true? Shall I act really severely? Why isn't this true? Perhaps that's a little dull, you mean probably 'longer on average'?" You see that the teacher in this short fragment poses four questions. You could expect that the students follow him and try to answer

these questions themselves. Substantially, he points at the gap between the data, given in 'number of hours per week' and this student answer mentioning 'frequency'. Strictly speaking, these are not one-on-one connected to each other. We ask ourselves: would this not be a good issue for students' input?

[3] Why does he explain the answer given by S1? He could ask her to do so. From the observations we have seen in this case study this appears to be a pattern in the way the teacher interacts with his students.

[4] Teacher and students agreed afterwards that the classroom discourse was more focussed on mathematics than during traditional classes. This is illustrated in a funny way here: the student was commenting on another students' answer substantially. But the teacher interpreted her comment as being an offense for the one who gave this answer.

[5] Here again the teacher could have paused briefly (Rowe, 1974) after: "…you see that the variation with boys may be bigger, how do you see that?" It is most effective when the teacher waits and is able to let all students think for a couple of seconds. The students should also have the discipline to wait with their answer until the teacher points someone to answer. In the next case study this discipline seems to be more normal than in this one.

[6] The teacher gives a good summary of the previous work but then an important aspect of the specific variation is not well indicated: what exactly makes the SD and the IQR for boys bigger, but the range for girls bigger? We question why the teacher did not use the network to present both data sets as box and whisker plots, making it immediately obvious that there are two girls with extreme computer behaviour causing this effect. Determining this, discovering the power of a visualisation and reasoning about what it shows certainly belongs to statistics education as we developed with the prototype.

[7] We observed that the teacher does not personally commit students to the classroom discourse, for instance by addressing questions individually, much more often in this case study.

7.8 Feedback in the third case study S3

7.8.1 Starting point and overview S3

Starting point as perceived by the teacher

Below we present the starting point, as perceived by the teacher and reported in the first part of a questionnaire.

The teacher was male, 28 years old, with 7 years of experience. His relationship with the group he rates as adequate (3 on 4 point scale). It was good with respect to mathematics, but he feels he has given this group of students too little personal attention. The level of the class he considered to be moderate (2 at 4 point scale). It was in his opinion a weak group, almost 40% had an average test score of below 6 when entering this intervention. It was a diverse group, with many students from vocational education, students from other senior secondary education schools in Haren, Beilen and Assen and students from school own SSE 3 classes. The teacher describes his own ICT competence as good (4 on 4 point scale) as he has always been a serious user of ICT in a variety of ways.

Correspondence between HTT and implemented feedback

This third case study can, from the perspective of approximation of the HTT, be considered as the best. The results are shown in table 7.5.

Table 7.5 Correspondence score characteristics of case S3

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL-M_ASS	%-Missing
S3	5,38	1,85	5,34	5,46	-0,12	0,00

The mean correspondence score was 5.38 (SD=1.85) and literally all of the planned feedback sessions were realised. Feedback to both ASS as to DL was realised close to the intentions as described in the HTT.

So the question is now: what are the characteristics of teacher behaviour that led to this case study coming closest to our expectations? Are there specific group characteristics playing a role?

In the first two case studies we noted during classroom observation that:

1. Giving feedback on the students' work in the case of an Open Answer exercise (see section 5.3.2 for a description of exercise types) in real time during the lesson was very demanding for the teacher. The system delivered a list of students' answers on the exercise. These answers, being given to an open question, cannot be evaluated by the system, so the teacher has to do it himself live.
2. Teachers had the tendency to give feedback on all of the exercises that were completed by the students. This led to very full lessons, and to too little variation in the way students were working during the lessons, as was noted by the observer, the teacher and the students.

Through splitting up the teaching activities in 'self-sections' and 'sections', we tackled both classroom problems. Teachers could collect students' answers on the sections through the network, analyse them and make a simple 'feedback scheme', while the students are working on a next 'self-section', in which they get rudimentary feedback on their GC and can ask for the right answer. We aim at feedback sessions that take 50% of the instruction time.

7.8.2 Classroom discourse example S3-1

Exercise 6.6 (DL, OA)

In exercise 5 we saw that the data set ‘pocket money’ was more capricious than the data set “height". How would you explain that?

HTT Ex6.6

A sequel to 6.5 with a comparable objective: thinking about the nature of the data and about the intrinsic difference between the presented data sets.

This time, no MC but OA, thus the students will have to formulate their thoughts themselves.

Answer checking and teacher feedback will lead to the students’ thoughts on the intrinsic differences between the amount of pocket money and the height, seen from the perspective of the data set: phenomena based on nature (more predictable) and those

based on human behaviour (more capricious). Height is height, but pocket money is not very well defined: what do the boys have to pay for with it?

Besides that, we see that the amount of pocket money will often be rounded off to a “nice number” (whole or half euros, for example).

Coded HTT Ex6.6

Ex 6.6 *SLG*=DL → *CAF* (h, DL) → *SII*(h, h, DL) → *TC* (h, DL)

Implemented feedback Ex6.6 S3

[08:33]

T: “Good. This is the last exercise of section 6. 'In exercise 5 we saw that the data set ‘pocket money’ is more capricious than the data set “height”. How would you explain this?' And in advance I've checked the answers. S1, what have you answered there?”

S1:“Which one is this?”

T: “Exercise 6 section 6. So: exercise 6, section 6. There is this question [he points at the projection] umm, and we've seen that pocket money is more capricious than height. S1, why would that be, do you think?”

S1: “First I'll have to look at the right exercise 6. With me it's exercise 7.”

T: “In the booklet, yes.”

S1: “Money you can influence yourself"

T: “Yes. So S1 says: You can influence money yourself. S2, what do you pose against that?”

S2: “Umm, I was just looking.”

T: “So that's exercise 7. In the booklet, that's exercise 7 of section 6. We are now comparing a couple of things, S1 says that you can influence money yourself...”

S2: “Heights, there are less outliers than with pocket money.”

T: “Yes, ‘Heights have less outliers than pocket money’. And how can that be explained, S3?”

S3: “...”

T: “What did you answer?”

S3: “I'm not there yet.”

T: “Well, perhaps you remember that S1 says 'Money, you can influence yourself', 'Height has less outliers' because...”

S3: “That depends on your parents.”

T: “Yes. That depends on your parents. Ssssh. Umm. Boys. A good remark is made here, S4 could you please explain it?”

S4: “Umm, well, that the parents both have influence on both factors.”

T: “So, the parents have influence on height and on money. How come?”

S4: “Do I have to explain it [inaudible]?”

T: “No no, but perhaps you could give a nuance?”

S5: “Income!”

S4: “The income of your parents, indeed, they can, if they give you more or less pocket money. And, normally, the height of your parents has also determines also your own height.”

T: “Yes, so the heights of your parents has influence on your own height, S2 says rightly, there are less outliers. But what gives us those outliers with pocket money? You say, that has to do with income. Does anyone have another,S6?”

S6: “I thought culture.”

T: “Culture. Is there another argument? S7?”

S7: “One earns just more than others.”

T: “Yes, that's the same as income. S1?”

S1: “Avarice of parents.”

T: “That's also possible. Avarice, yes. But...”

Students then talk through each other.

“S8, what do you for instance have to pay from your pocket money?”

S8: “Umm, clothes.”

T: “Clothes. Who doesn't have to pay clothes from his pocket money?”

[11:39]

S9: “I get separated clothing money.”

S8: “Ah, from my pocket money?”

T: “Yes, from your pocket money.”

S8: “Oh, I thought clothing money. Umm, I don't know, things I want to do. Going out.”

T: “Going out. Which one of you does have to pay his clothes from his pocket money?”

S9: “From my salary.”

T: “From your salary. Yes. Umm. Where am I pointing at, with this question? S10?”

S10: “That, I don't know.”

T: “For example?”

S5: “One gets more because one has to pay his own clothes.”

T: “Yes, there are different things for what you have to use your pocket money for.”

[12:14]

T: “So that was important, where human influence, about which we were discussing, in different wordings, is important, with pocket money and with heights, that's something more like a natural phenomenon, although S4 is of course right that heigth partially is determined by your parents.”

Students still discuss a little more about this.

[12:34]

Coded implemented feedback Ex6.6 S3

Ex 6.6 SLG =DL → CAF (prep) → SII(h(7), h(9), DL) → TC (h, DL)

Correspondence score: 7.

Interpretation classroom discourse Ex6.6 S3

The teacher has inspected the ClassAnalysis file while the students were working on the exercises of sections 7 and 8. He made up a feedback scheme. Then he shows the exercise on the screen and starts the feedback session that will take 4:01 [1]. He asks S1 to give her answer. That is a very good answer, focusing on the impressionability of both entities 'pocket money' and 'height'. The teacher asks S2 for his answer, that formulates it rather differently ('With height, there are less outliers'). The teacher starts searching for the reason behind this. S3 suggests that this is due to someone's parents. But then the objection is made by S4 that parents influence both height and pocket money [2]. The teacher specifically looks for an explanation for the outliers in pocket money [3]. When there is no convergence towards his idea (the functionally vague definition of what pocket money is) he poses a somewhat more suggestive question "What do you have to pay from your pocket money?" [4].This induces a lively discussion that is conducted very well by the teacher. When he has the definition issue covered in a satisfactory way he switches to the conclusion [5].

Remarks on the classroom discourse Ex6.6 S3

[1] The teacher has this habit for the exercises (usually of DL character) he wants to give feedback on and discuss with the group. He does not show the ClassAnalysis file on the projection screen. This maybe a disadvantage, because the students do not see where the teacher's information comes from. This could make it less intense.

[2] It is characteristic of this case study that within a few moves there is a spontaneous substantial discussion, but the teacher stays in control of the discourse. The interesting remark that parents influence both pocket money and height could have been elaborated by a next question: "How do they influence exactly?" Then the difference in influence would perhaps show up: on height they depend on nature and pocket money they can directly influence.

[3] It is of course good to search for the reason behind the fact that the data set pocket money showed more whimsicality than did height, but this whimsicality shows up in more than just outliers. The teacher should draw attention to the fact that 'round numbers' are particular popular when it comes to the amount of pocket money. This will induce peaks in the histogram.

[4] The teacher wants to discuss the vague definition of what pocket money is, in particular, what does someone have to pay with it? It is quite concrete and still subtle to ask a specific student what she has to pay. There is a considerable chance that the problem with the definition will arise.

[5] When formulating the conclusion, which he does carefully, the teacher uses 'human influence'. The fact that the influence mentioned is human is not really discussed. What could this effect be when compared to nature (if it is possible to oppose humans to nature)? This discussion could be very interesting.

7.8.3 Classroom discourse example S3-2

Exercise 8.8 (DL, OR)

What do you conclude, using these data [mean, S, IQR and range of the boys and of the girls], about the computer behaviour of 12-year-old girls and 12-year-old boys in 2000?

HTT Ex8.8

Suggested response: “Boys use the computer on average more than girls. The variation is also bigger with respect to IQR and SD. The range, on the other hand, is bigger for girls. When inspecting the box and whisker plot again, you see two typical outliers amongst the 400 girls presented, responsible for the inconsistency in the measure of spread.”

In sections 1 and 2, the students had to explain differences just based on a measure of centre (mean). Now they also have three measures of variation (inter quartile range, range and SD) at hand.

They should include this in their comparison of the behaviour in computer use of boys and girls. The teacher collects and checks the students' responses. We expect that most students will not include the variation of both data sets into their comparison. The teacher then is to give feedback on the enhancement of the comparison when including measures of variation, if possible with selected students' responses as a starting point.

Coded HTT Ex8.8

Ex 8.8 *SLG*=DL → *CAF* (m, DL) → *SII*(h, h, DL) → *TC* (h, DL)

Implemented feedback Ex8.8 S3

[14:29] At the whiteboard the results of section 8a are presented: mean, SD, IQR, range for both boys and girls.

T: “We take a look at what we've got, then I have to inspect the right note so we know where we are going. Eh, S1, we've heard him already today. Eh S2, we haven't heard yet. What's your conclusion there?”

S2: “Boys work a lot longer with the computer than girls"

T: “And from what did you conclude that?”

S2: “Eh, the mean.”

T: “So you compared the mean of both boys and girls. But S3 compared something different. Didn't you?”

S3: “I guess so. Which question do you mean exactly?”

T: “The exercise is the first one of 8b. What do you conclude with respect to the data from 8a about computer behaviour?”

S3: “In the next question it was mentioned that the variation was bigger.”

T: “Ah, you've looked forward. But where do you base that on? S4?”

S4: “That one is bigger compared to the other?”

T: “Which one? Yes, it's good you point at that, but which one is bigger in relation to each other?”

S4: “Girls, right?”

T: “Among girls the range is bigger, indeed, you're right. But, almost everyone looks at the mean, don't they?”

[15:56]

S5: “I don't!”

T: “S5?”

S5: “Well, I took everything together"

T: “You took everything together, very well.”

S5: “And then you see still that boys work longer with the computer.”

T: “How do you see that?”

S5: “The mean is a lot higher.”

T: “The mean is higher, very good.”

S5: “The standard deviation is bigger.”

T: “The standard deviation is bigger, yes, good.”

S6: “The median!”

S5: “The interquartile range...well, I don't know whether that has something to do with it, but it's bigger.”

T: “Yes, because the interquartile range is about...the...S7, what was that, the interquartile range?”

S7:“Median?”

T: “Yes, but?”

S5:“First quartile and third quartile.”

S4:“Ah yes, then you get a box and whisker plot, or something.”

S8: “I didn't understand that.”

T: “Umm….”

[16:32]

T: “We will in a moment further explain the box and whisker plot, but do the outliers belong to the interquartile range?”

S5: “No.”

T: “No. Very good. So you see that this one [IQR] was indeed bigger, so S5 draws a very fine conclusion, because she didn't just look at the mean. She compared also the standard deviation, which turned out to be bigger, the spread, therefore the spread of the central half is bigger, because the quartile distance is 7 instead of 4. What do we then know about some of the girls?”

S7: “That some girls work really very long with the computer.”

S6: “But that's just one girl.”

T: “Umm, I hear a couple of very good answers, umm, S8, what do we know about the girls?”

S8: “That some spend a lot of time on working with the computer.”

T: “Yes, but how many? S5?”

S5: “Umm...”

S5: “You just gave the correct answer.”

S5: “60 hours?”

T: “Yes, but how many girls?”

S5: “Yes, just one.”

T: “Yes! One or two. They're outliers. The range just takes the maximum, the biggest, and the minimum, the smallest, into account. It's very good that you realise that, then you really understand what the range is. Perfect.”

[17:59]

Coded implemented feedback Ex8.8 S3

Ex 8.8 *SLG*=DL → *CAF* (preparation) → *SII*(h(7), h(13), DL) → *TC* (h, DL)

Correspondence score: 7.

Interpretation classroom discourse Ex8.8 S3

In just 3'30" the teacher conducts a very interactive classroom discourse that starts with the feedback scheme he sketched in a couple of minutes while inspecting the CA file. An example of such a scheme is shown in figure 7.3.

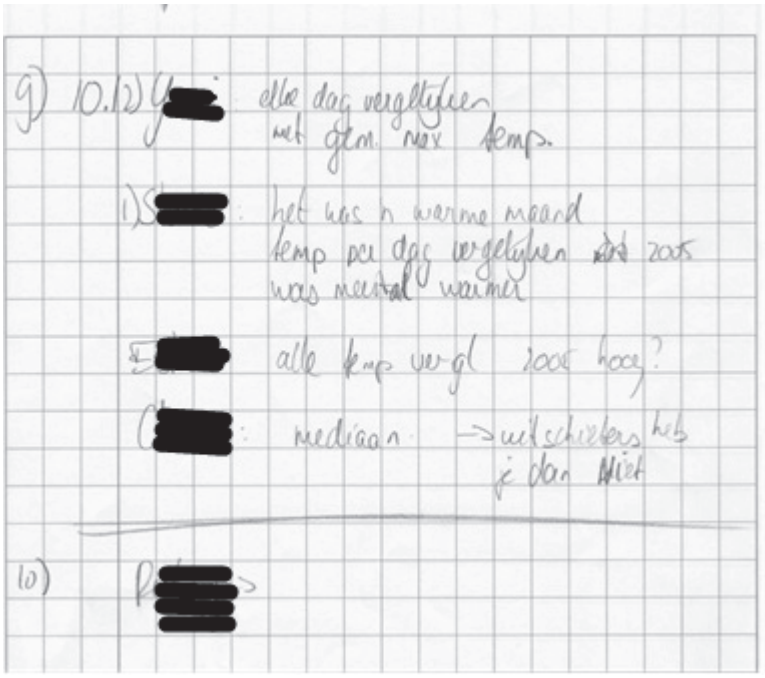


Figure 7.3 A feedback scheme based on the CA file for an OR item

It contains of a couple of remarkable students' responses (in the case of the shown exercise 1.9 there were four). These exercises together represent reasonably the ‘response space’ of the whole group. With these responses the students should be able to discuss the exercise, while being conducted by the teacher in his role of ‘substantive discussion leader’.

What do we see happening more specifically?

The teacher starts questioning the group with the feedback as a starting point. He rephrases his question, or refines it when he is not really satisfied by the answer. The input from S2 (‘bigger mean’) is too restricted but an obvious student response as we saw in the case studies S1 and S2A [1]. Because of his feedback scheme, the teacher is prepared for that and he poses the response from S3 as a counterweight [2], but S3 admits that she has looked forward in the exercises and that was how she came to the difference in variation [3]. The teacher takes the concept of variation now as given, and continues questioning while personally addressing [4]. S4 then compares the variation of both data

sets, although it is not very explicit with which measure. The teacher does not point this out, but translates it directly to the range (must be, because the other two measures of spread are smaller for girls than for boys) [5]. He points at the fact that almost everyone just used the mean to compare the data sets. Then S5 reacts, because she did not do so. The teacher lets S5 report about her inspection of the measures as gathered from section 8a. She stops at the IQR. He is about to tell her what the IQR stands for, but stops himself and rebounds the question to S7 [6]. In discussion with some students he succeeds in defining the IQR [7].Then he hears a student remark and reacts on this by noting that this they will come back to. He stays focussed on the question of how it is possible that SD_boys > SD_girls and IQR_boys > IQR_girls but that range_boys < range_girls [8]. The teacher asks what this means for some of the girls. Several students give good answers. He picks a student out of these. This student gives the first part of the good answer. Then he asks another student who gives the second part of the good answer [9].

Remarks on classroom discourse Ex8.8 S3

[1] Use of the feedback scheme as a starting point of the discussion. S2 just used the mean to compare both data sets.
[2] Because of his feedback scheme, the teacher knows that S3 answered differently.
[3] It is unfortunate that S3 has no substantial argument for her answer. The core of the answer was 'variation', and this the teacher takes as a starting point for further discussion.
[4] The teacher continues his questioning, while addressing his questions personally. This is difficult. In practice, you often see that personal addressing, if it appears at all, fades out after one or two prompts.
[5] Here the teacher wastes an opportunity. He could have asked further: “Which measure did you use for that conclusion? Are there other measures? Would those have yielded the same result?” He now gives away the use of the range.
[6] The fact that the teacher stops and passes the students’ questions to their peers illustrates how focussed he is on his task.
[7] Continuous questioning of the students
[8] The question: “How is it possible that SD_boys > SD_girls and IQR_boys > IQR_girls, but that range_boys < range_girls ?” is leading, but perhaps not stated explicitly enough.
[9] Until the end he keeps addressing refined questions personally. Perhaps showing both boxplots again with the outliers included, showing the correctness of this reasoning, could have made this feedback session perfect.

7.9 Feedback in the fourth case study S4a

7.9.1 Starting point and overview S4a

Starting point as perceived by the teacher

Below we present the starting point, as perceived by the teacher and reported in the first part of a questionnaire. We mention here that this group followed senior secondary education in a school for adult education. This meant for this group that the students were almost all 18-19 years old, instead of 16-17 which is normal for the other groups.

The teacher was male, 55 years old, and had 22 years of experience. He rated the relationship with the group as adequate (3 on 4 point scale).The level of the group he considered to be moderate (2 on 4 point scale). Although there were some good students

in this group, the majority have some or more problems with mathematics and their homework discipline is very bad with almost nobody doing anything outside the classroom. His own ICT competence he considered to be good (4 on a 4 point scale), as he has always been using ICT and pays considerable time to applying it in his lessons.

Correspondence between HTT and implemented feedback

From table 7.6 we conclude that there is a reasonable mean correspondence score between HTT and the implemented feedback, but there is still substantial improvement possible.

Table 7.6 Correspondence score characteristics of case S4a

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL_M_ASS	%-Missing
S4A	3.60	2.13	3.65	3.44	0.21	34.04

We also see that a substantial part of the planned feedback sessions are not conducted. During our observations, we noticed that the teacher was sometimes slow in his acting. In other situations, he seemed to trust his usual techniques of conducting the classroom discourse. Sometimes he seemed to forget his extended feedback possibilities. He reported this himself in the teacher questionnaire. When compared with the teachers in the case studies before, it could be that he needs more time in order to internalise this way of teaching.

Perhaps the biggest problem with the pace in the lessons was caused by the fact that about half of the students did not do their homework. How much this may seem, both teacher and students state that homework behaviour was substantially better than before.

7.9.2 Classroom discourse example S4a-1

Exercise 2.11 (DL, True/False poll)

Is it pure coincidence that in the situation of the temperature in September 2005 the median and mean are closer to each other than in the situation of the number of mp3 files on the computers of SSE-10 students?

HTT Ex2.11

Sequel of 2.11: the answers on the True-False poll are not as interesting as is the previous reasoning about why to choose this answer. Discussion should lead to the difference in the nature of the data: temperature as a natural phenomenon and the number of mp3 files as a result of human behaviour. This should be related to the difference in nature between mean and median, especially with respect to the role of outliers.

Coded HTT Ex2.11

Ex 2.11 *SLG*=DL → *CAF* (l, DL) → *SII* (m, m, DL) →*TC* (h, DL)

Implemented feedback Ex2.11 S4a

[20:00]

T: “It is pure coincidence that in the situation of the temperature in September 2005 the median and mean are closer to each other than in the situation of the number of mp3 files on the computers of SSE-10 students. True or False? Is that pure coincidence? Well, let's take a look at the results.”

Students: “Ah!”

Teacher shows a bar chart with the results: 2 answered ‘True’ (incorrect answer), 7 answered ‘False’ (correct answer), 3 no response. “Very good! Seven. Who had a correct answer? S1? Can you explain why you had chosen ‘False’ if you take a look at the question?”
He goes back in the slideshow to represent the question.

[20:48] “Why is this false? Why is this no coincidence?”

S1: “Well, because for the mp3 files there are outliers and in the temperature there are almost none.”

Teacher is silent for a few seconds.

T: “Yes. But is it a coincidence? That was the question.”

S1: “Yes. No, then it's not a coincidence.”

T: “And why isn't it coincidence.”

S1 laughs. Other students laugh too. Pause of few seconds.

S2: “In mathematics, coincidence doesn't exist.”

Students again laugh.

T: “Why, why is this no coincidence? Why isn't this true?”

S3: “The temperature differences are between 30 and...10. So there can't be an outlier of...50.”

[21:54]

T: “Maybe, one has to ask, should I ask: in which situations do outliers occur? Who's able to answer that?”

S3: “MP3 files.”

T: “At the mp3 files, there are outliers. But why? Why are there at the mp3 files outliers?”

S4: “Because it's possible.”

T: “Why is it possible? And why not with the temperature?”

S4: “Well, because if it's today zero degrees, it won't be thirty tomorrow.”

Students laugh.

S3: “That can't be the case.”

S4: “It's impossible.”

T: “So there are situations in which you know in advance: there won't be any outliers?”

S5: “If you look at students, then it's quite easy.”

T: “Yes.”

S4: “If you look at a looser, who has nothing to do all day, he will be downloading all day, so umm…”

Students again laugh.

T: “Okay, clear. So there are situations in which you can ask uh… if you going to investigate something: can there be outliers? Well, that's the next question...”

[23:06]

Coded implemented feedback Ex2.11 S4a

Ex 2.11 *SLG*=DL → *CAF* (l(l), DL) → *SII* (h(4), m(5), DL) →*TC* (h, DL)

Correspondence score: 7.

Interpretation classroom discourse Ex2.11 S4a

The teacher starts with ClassAnalysis. He reads aloud the exercise and shows the students’ results. Apparently this is still impressive for the students in some kind of way [1]. The results show that 7 out of 12 answered correctly. The teacher is enthusiastic about this result [2]. The teacher poses the question, but does not address it personally. S1 reacts spontaneously. He explains that this is a matter of difference in outliers. The teacher then asks for further explanation on this [3]. Eventually S1 laughs and other students laugh too [4]. This laughing challenges S2 to make a joke about mathematics [5]. The teacher accepts this relaxed classroom atmosphere by laughing himself, but keeps focused on the learning goal. Then S3 spontaneously states: S1: “The temperature differences are between 30 and...10. So there can't be an outlier of...50.” The teacher accepts this and abstracts it to the question in which situations outliers can be expected [6]. He continues determined, just as long as S4 contributes that in the concrete examples there's just a difference in the chance on outliers. Apparently the teacher is satisfied [7] and he fluently goes on with the next exercise in which a more concrete situation will be explored with respect to the chance of outliers.

Remarks on classroom discourse Ex2.11 S4a

[1] A good start to this session. There is something that surprises the students. Possibly the fact that with one mouse click their responses are evaluated and presented with a neat bar chart, although we think it is more likely that the length of the green bar (representing the number of good responses) causes their enthusiasm.
[2] The teacher really considers this to be a good result. One could question that, given the fact that there are 17 students in this group and that this item was ‘true/false’ so there was a gambling chance of 50%. Perhaps this characterises the school culture, which is not very demanding from the students.
[3] The teacher seems to be asking what the reason is for there being outliers in the mp3 data but not in the temperature data. He does this patiently, which offers S1 the full opportunity to show his deeper understanding.
[4] This laughing probably illustrates that S1 and his fellow students realise that he uses a circular argument. The laughing of his fellow students indicates that they have tried to understand his arguing, and come to the same conclusion as S1 does.
[5] S2's joke about mathematics illustrates that the discourse is focused on mathematics, even when it comes to joking.
[6] This is a good move from the teacher with respect to more abstract student thinking. The question: what distinguishes an outlier from a 'normal big value' of a certain phenomenon is not posed. Perhaps the teacher considered this to be distracting from the strict learning goal or it did not come into his mind as relevant for the discussion.
[7] For these concrete contexts, the question is answered, but the teacher does not extend the discussion to describe the difference *in nature* between both data sets.

7.9.3 Classroom discourse example S4a-2

Exercise 8.8(DL, OR)

What do you conclude, using these data [mean, SD, IQR and range of the boys and of the girls], about the computer behaviour of 12-year-old girls and 12-year-old boys in 2000?

HTT Ex8.8

Suggested response: “Boys use the computer on average more than girls. The variation is also bigger with respect to IQR, SD. The range, on the other hand, is bigger for girls. When inspecting the box and whisker plot again, you see two typical outliers amongst the 400 girls presented, responsible for the inconsistency in the measure of spread.”

In the sections 1 and 2, the students had to explain differences just based on a measure of centre (mean). Now they also have three measures of variation (inter quartile range, range and SD) at hand.

They should include this in their comparison of the behaviour in computer use of boys and girls. The teacher collects and checks the students' responses. We expect that most students will not include the variation of both data sets into their comparison. The teacher then is to give feedback on the enhancement of the comparison when including measures of variation, if possible with selected students' responses as a starting point.

Coded HTT Ex8.8

Ex 8.8 *SLG*=DL → *CAF* (m, DL) → *SII*(h, h, DL) → *TC* (h, DL)

Implemented feedback Ex8.8 S4a

On the whiteboard the results of section 8a are presented: mean, SD, IQR, range of the hours of computer use by 12-year-old boys and 12-year-old girls.

[59:21]

T: “What can you conclude with these data? Whom may I give the word? S1, can you say something about that? If you compare these numbers with each other? Can you tell something about the behaviour of the boys and something about the behaviour of the girls?”

S1:“Well, it's remarkable that boys spend more time on working with the computer than girls.”

T: “How do you see that?”

S1: “The mean.”

T: “Yes. The mean of the boys is...”

S1:“44 hours. And the mean of the girls is 4 point...”

T: “That's twice as much. Yes. And what else can you tell?”

S1: “Umm, I don't know what else I can say about this.”

[1:00:07]

T: “Who? Ah, that's S5. [Student enters the classroom] You were right, S2. Good morning. Take a sender and log on quickly. Good, you see from the mean that they spend twice as much time working with the computer. What can you tell about these three

numbers? [He points at the SD, IQR and the range]. How do we call these three numbers? What kind of numbers are they? That tells something about the...”

Teacher points at S3.

S3:“Spread.”

T: “About the spread. Standard deviation, interquartile range, and the range was...”

S1: “Biggest minus smallest.”

T: “Highest minus smallest.”

S4: “So among the girls there are some real computer nerds.”

T: “Ah! How do you see that, S4?”

S4: “Because the range is 60. So in the group there are the ‘normal’ [makes a gesture of quote with his fingers] and then you have the real ICT girls that really work a lot with the computer.”

T: “I think this is great. I think this is really great. You see here, looking at the range, that girls have here a bigger absolute spread than the boys. So there may be here [points at the girls] some outliers. Shall we inspect that visually?”

[1:01:44]

T: “Then we open the data set. I've forgotten in which section these data are, oh, of course in 8a. Now I'm not very handy by opening 8b when needing 8a. We're going to visualise this, but I think it's great, S4. That's the main conclusion. You can say, boys are working longer with the computer, but among girls, we see that by the range, there are a few outliers.”

[1:02:22] The bell rings. Students have been active until this very moment.

Coded implemented feedback Ex8.8 S4a

Ex 8.8 *SLG*=DL → *SII*(m(3), m(6), DL) → *TC* (h, DL)

Correspondence score: 4.

Interpretation classroom discourse Ex8.8 S4a

The teacher does not start with ClassAnalysis [1]. He formulates an open question (“Can you tell something about the behaviour of the boys and something about the behaviour of the girls?”) and points personally. In a discussion with S1 it appears that the mean of the boys is much higher than the mean of the girls, when it comes to the number of hours per week using the computer. The teacher then points at the other three measures on the whiteboard and asks what these measures indicate [2]. The first response from the group is good right away. Then the teacher asks for the definition of range, which comes pretty easily from the group too [3]. Then suddenly S4, while looking at the numbers, comes with this very sharp remark that there seem to be some girls that really use the computer an awful lot. The teacher is very pleasantly surprised, but fortunately keeps his professional reflexes when he asks for an explanation [4]. S4 explains his insight eloquently, but forgets to include the other measures of spread in his reasoning. The teacher is that enthusiastic about this student input that he forgets to give critical feedback on it. Further, he follows this way of reasoning in his recapitulation [5].

He then decides to use the network to visualise the data. He loses some time through a little clumsiness and then the bell rings [6].

Remarks on the classroom discourse Ex8.8 S4a

[1] He did that with the previous exercise, which resulted in the values of mean, SD, IQR and range for boys and girls on the whiteboard. Possibly, it is still not a standard act to use CA.
[2] In this group too, the students focus on measures of central tendency in order to characterise a data set, especially on the mean. The teacher therefore focuses on three measures of spread, as calculated in section 8a, and asks what those measures all have in common. This is a way of subtly directing towards the concept of variation.
[3] As we saw in the teacher and student evaluation of the intervention (see section 7.4), both teacher and students attributed the intervention to better learning, for instance because of the fact that the students were more actively involved with the exercises. Possibly this causes the ease with which the right answer comes from the group. Unfortunately, the teacher did not address this question personally.
[4] This is of course a key point in this intervention: always discussing the why.
[5] S4's reasoning is not complete. The conjecture that the data set for the girls contains at least one outlier is not only supported by the fact that the range for the girls is bigger than for the boys, but reinforced by the fact that the other two measures for spread used here (SD, IQR) indicate that the variation among boys is actually considerably bigger than the variation among girls (for instance, SD-boys=8.36, SD_girls=4.01). Thus there is really one very atypical girl in the dataset.
[6] The teacher forgets the time during the lesson; this happened often. He has the right intuition in looking at the two boxplots because the insight S4 showed by just looking at the numbers is not accessible for most students. However, seeing it in a graph will convince everyone that the girl that uses the computer most really uses it for a remarkably long time. Perhaps the possibility of the network for the sake of supplying feedback, which seemed to lack at [1], is nevertheless partly internalised for this teacher at this time.

7.10 Feedback in the fifth case study S4b

7.10.1 Starting point and overview S4b

Starting point as perceived by the teacher

The teacher reported that the starting point in group S4b was approximately the same as in group S4a (see section 7.9.1).

Correspondence between HTT and implemented feedback

In Table 7.7 below we present the overall correspondence data. We respectively present: the mean total correspondence, the standard deviation of the total correspondence, the mean correspondence with respect to data literacy exercises, the mean correspondence with respect to algorithmic statistical skill exercises and the subtraction of these means.

Table 7.7 Correspondence score characteristics of case S4b

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL-M_ASS	%-Missing
S4b	3.89	2.09	4.11	3.33	0.78	36.17

We see approximately the same picture as with the results case study S4a.

7.10.2 Classroom discourse example S4b-1

Exercise 2.7 (DL, OA)

Which measure of central tendency gives a good picture of the number of MP3 files on the computers of the students H4C? Why?

HTT Ex2.7

This is the key question of the first part of this episode.

Students are here forced to make up their minds about the influence of an exceptionally big value in the data set on the mean and the median. Which of both is more sensitive to outliers? Or – dually stated – which of both measures is more robust? Of course, the explanation is the most interesting part. Therefore this is an open answer exercise.

Coded HTT Ex2.7

Ex 2.7 *SLG*=DL → *CAF* (m, DL) → *SII* (m, m, DL) →*TC* (m, DL)

Implemented feedback Ex2.7 S4b

[07:49]

T: “These are the responses. Someone has answered ‘12’, thinking ‘I have to give some answer’, but the question was clearly: what is more appropriate here, the mean or the median?”

S1: “I was wrong, it seems.”

T: “Which one is yours?”

S1: “The second one.”

T: “That's quite a story.”

S1: “Yes.”

T: “The most frequent number bigger than zero gives the best picture of the mp3 files, because the mean is too sensitive for outliers and the median is not suitable, because that's the middle number and even if you sort them with respect to size, that doesn't have to be the most common number.”

S1: “That was just a little mistake, from me.”

T: “Why?”

S1: “I would probably, I would now say the median, because the mode is zero and you can hardly say the mode above zero [he possibly suggests a restricted domain for the mode] or something. That's what that says, basically.”

T: “Yes, there was a distinction between two measures of central tendency, the arithmetic mean and the median, and the mode was in this stage not introduced yet. And you introduced the mode. Why, umm, S2, why did you choose the median?”

S2: “I don't really remember.” Teacher waits.

S2: “Because that's the middle, so the mean, so also in general, is it that approximately, it is just the mean, but then the median, that is, let's say, the middle number...”

T: “Yes?”

S2: “..er yes, I don't know.” He laughs.

T: “Who else chose for the median? There were some more.” [Points at the ClassAnalysis bar chart]

[09:52]

T: “…'Median. Because, there are three measures of central tendency.' Anyway, the median was chosen here. Who else chose for the median?”

S3: “Those answers are not all of them, are they?”

T: “No, I can scroll down. There are for without a response, who didn't make this exercise yet. There was one who chose for the arithmetic mean. Why is median the right answer on this question?”

S1: “Because the arithmetic mean is too sensible for outliers and the mode hasn't been introduced so far. So just the median is left.”

T: “Exactly! What S1 says, the median doesn't suffer from...”

S1: “Outliers.”

T: “Outliers. If we present an example, to illustrate this better, imagine I have collected these observations...”

He writes down on the white board a data set ordered at size.

T: “...I just write down some numbers, what they represent is not that important, I have 6 data, what's the median?”

S1: “(10 + 12)/2 = 11. Middle two, normally the middle one, but there are six observations, so you have two in the middle.”

T: “Yes. The problem with the middle number here is,...,S4, what's the problem with the middle number?”

S4: “Yes, I can't see it, so I have no faint idea.” [S4 suffers from a sight problem.]

T: “Why don't you come and sit here?”

S4: “Yes, but that won't work in the end, so I will stay here.”

[0:11:46]

T: “Okay. S5, what's this problem? If you want to know the median of these observations?”

S5: “There is no middle number, there are six.”

T: “It's an even number of observations, so you can't point a middle number, so you take the middle two, and from those you calculate the mean, and that's the median. Well, imagine that there's an observation made that is very large, 2000, then you see that this two thousand is actually some kind of exception among the other observations here. What's the median now?”

S1: “12,”

T: “Now you can determine a middle number, first time, second time, the median is now 12. You see, that despite the outlier of two thousand, the median doesn't change, hardly changes, it stays 11 or 12.”

S6: “That just has to do with what number is in the middle?”

T: “Yes. It happens to be the middle number.”

S6: “If you have ordered them at size?”

T: “Ordered from small to large, yes. That is important.”

[0:13:10]

S7: “There is two times median mentioned.”

T: “Yes, the first time, I didn't include 2000, and the second time I've done an extra observation which is suddenly 2000. You see that his 2000 is much bigger than the other numbers and that doesn't influence the median. Scarcely.”

S7: “Ah yes.”

[0:13:39]

T: “So, the answer on this question, about those mp3 files, that someone enters with a lot of files, and that you want to have a reliable picture of the 'normal' SSE grade 10 student and that person that enters, let's say, if you look at the median, hardly has any influence. This on the contrary with...”

S1: “The mean.”

T: “The mean. Because in the first case you had to calculate the mean of these and that's a complete other number when this 2000 has to be included.

This, you could have calculated with those mp3 files. There is a clear difference between the median and the mean if, all of a sudden, such an outlier enters. So the answer on this question is: the median gives a more reliable picture than the mean.”

[0:14:45]

Coded implemented feedback Ex2.7 S4b

Ex 2.7 *SLG*=DL → *CAF* (m(2), DL) → *SII* (h(6), h(11), DL) →*TC* (h, DL)

Correspondence score: 7.

Interpretation classroom discourse Ex2.7 S4b

The classroom discourse starts with an overview of the students' responses generated by CA [1]. See fFigure 7.4.

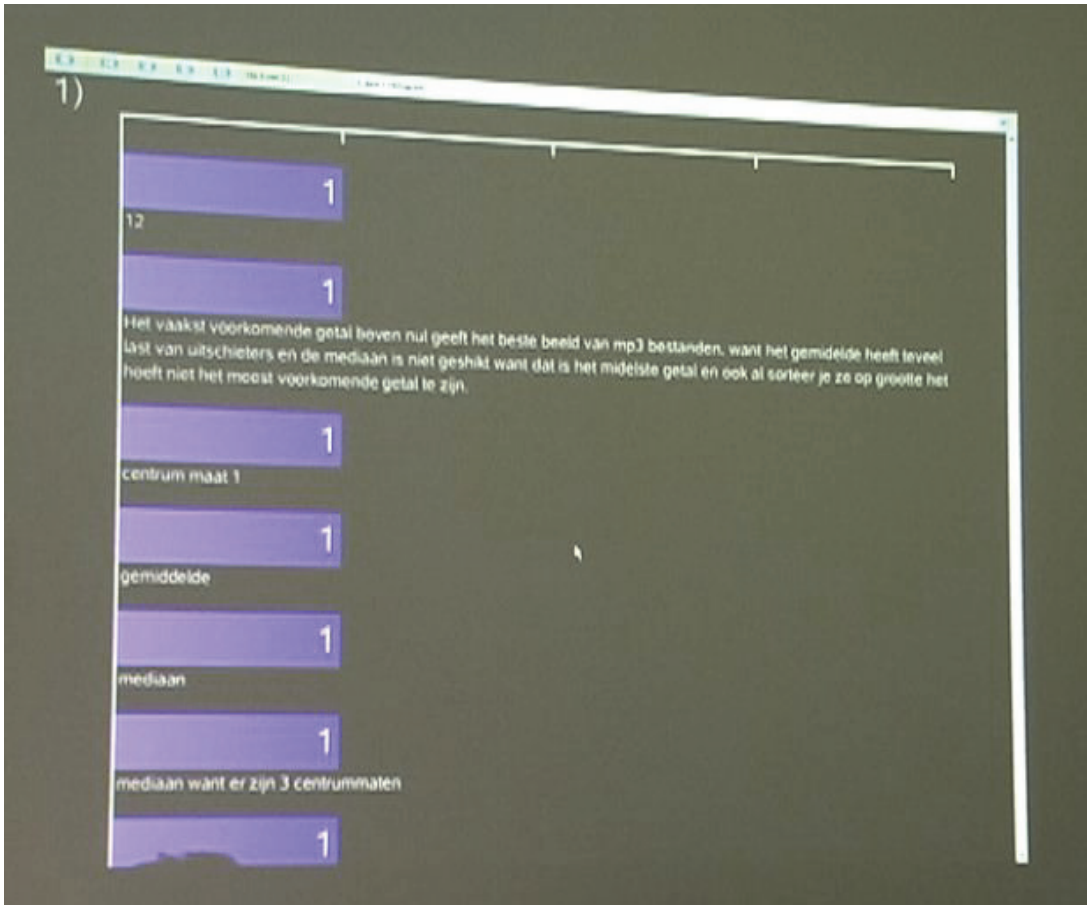


Figure 7.4 Students' responses on exercise 2.7 in case S4b, as represented by the CN

We see seven students' responses, usually just an answer without an explanation (although that was explicitly asked). The second is given by S1, who in the discussion is correcting this answer right away [2]. S1 gives a detailed exposé about the mistake he makes. The teacher, with rather minimal feedback, then switches to S2. Due to the teacher's persistence S2 tries to make a story of his answer but he is unable to make it convincing. His laugh betrays that he feels this himself too [3]. After this, the teacher comes back to S1. Dialogically, S1 almost gives a perfect answer: the median, being more insensible, for outliers [4]. After this, the teacher gives an improvised example of another data set. When he starts questioning this data set with respect to the mean and the median, again S1 takes the lead. The teacher lets this happen. He lets S1 calculate the median in two different situations. One with the 6 element data set, the other with this data set with a seventh element (2000) entered. The median shifts from 11 to 12 which is pretty invariable [5]. There is a nice fragment with S6 who contributes spontaneously to the discussion making some key points about the median [6].

Remarks on classroom discourse Ex2.7 S4b

[1] To start with, ClassAnalysis was not a routine for the teacher, surprisingly being used less in the last lessons than in the first lessons. This teacher was the only one participating with two groups, giving him some kind of 'experiential advantage' above the others. Making a short feedback scheme in advance, based on ClassAnalysis, was not realised once, although the success this had in the previous intervention was communicated with him.

[2] S1 is a very intelligent, fanatic, student with hearing problems and possibly with a slightly autistic disorder. He dominates the student input in this example. This was far

from an exception. In fact the teacher, being very polite, had problems with keeping the discussion broad because S1 always completed the most comprehensive thinking in advance of the lesson, which was unlike other students, as the response rate sometimes showed. He had, so to speak, the right to speak, but his speaking, that sometimes resembled lecturing, was not always a stimulus for his peers, who quite often had not done their homework, to think for themselves. In figure 7.2 you can see how his response to this exercise typically differs from the ones of his peers.

[3] The teacher shows patience here. S2 tries his best, but there is simply not more in it for him. This part shows how unfamiliar students can be with a measure of central tendency other than the mean.

[4] In our opinion, this should be accompanied by the notion that the context of numbers of mp3 files on the computers of 16-year-old students is a context in which outliers are to be expected.

[5] Both means actually had also to be calculated, to see the enormous difference in sensitivity for one outlier.

[6] S6 focuses on the fact that when trying to determine the median, the data have to be ordered at size. This is a very distinguishing characteristic with respect to the mean.

[7] The patience of the teacher also has a problematic side. This example took almost 7 minutes of classroom discourse. We cannot deny that the pace seemed a little low now and then during this case study. The teacher was patient, open, and willing to improvise. He really likes to explain mathematics, which is, of course an advantage in his profession. However, given the technological opportunities as given during this study, the teacher has to choose his moments of explaining to a greater extent and thus has to temper his tendencies.

7.10.3 Classroom discourse example S4b-2

Exercise 8.8(DL, OR)

What do you conclude, using these data [mean, SD, IQR and range of the boys and of the girls], about the computer behaviour of 12-year-old girls and 12-year-old boys in 2000?

HTT Ex8.8

Suggested response: “Boys use the computer on average more than girls. The variation is also bigger with respect to IQR, SD. The range, on the other hand, is bigger for girls. When inspecting the box and whisker plot again, you see two typical outliers amongst the 400 girls presented, responsible for the inconsistency in the measure of spread.”

In sections 1 and 2, the students had to explain differences just based on a measure of centre (mean). Now they also have three measures of variation (inter quartile range, range and SD) at hand.

They should include this in their comparison of the behaviour in computer use of boys and girls. The teacher collects and checks the students' responses. We expect that most students will not include the variation of both data sets into their comparison. The teacher then is to give feedback on the enhancement of the comparison when including measures of variation, if possible with selected students' responses as a starting point.

Coded HTT Ex8.8

Ex 8.8 *SLG*=DL → *CAF* (m, DL) → *SII*(h, h, DL) → *TC* (h, DL)

Implemented feedback Ex8.8 S4b

On the whiteboard, the results of the exercises of section 8a are presented. Mean, SD, IQR and range for both of the data sets (boys and girls).

[12:02]

T: “Why do you have to give a measure of spread when giving a measure of central tendency?”

S1: “Well, you need to...”

T: “I'm sorry, S1, I would now like to hear someone else. I think you're a good boy, that's not the problem, S2?”

S2: “Well, you have, centre indicates how much it is and the spread how often it occurs.”

T: “I think you mean it correctly. I think I have given an example once of two tests in mathematics. Of one test I calculated the mean score, that was a 6. And the second test also calculated the mean, and this was also a 6, then you think, ah well, that was the same test. But now I present the standard deviation, as an example. With this test there was a standard deviation of 1, and this one is 3.”

S3: “That means how far it is away from the mean, doesn't it?”

T: “Yes, deviation from the mean. Are you able to make a statement about these two tests?”

S3: “Yes!”

T: “Do so.”

S3: “The first is from 5 until 7, the second is from 3 until 9.”

T: “Yes, the data are more densely concentrated. Because the deviation from the mean [6-1, 6+1] is much smaller than this [6-3, 6+3] deviation.

This diversion, the standard deviation, you subtract from the mean and then you add with the mean. And then you get these two numbers. When we'll discuss the normal distribution, you'll exactly calculate which percentage that is. Then we can calculate that. For now, I say, that there are a lot of numbers between the 5 and the 7. In other words, this test had been made much more uniformly than this test. Here I have considerable outliers below, a lot of people did the test poorly, but happily also a lot of people did it very well. But they are much more scattered. Therefore it is important not only to use the measure of central tendency, but also the measure of spread.”

[14:06]

T: “Now we take a look again at the boys and girls. We have found one number, that's the goal of a measure of centre, now, what does the mean tell us?”

S1: “That doesn't say that much, in fact, because there are a very lot of outliers.”

T: “How do you conclude that, that there are a lot of outliers? If you just look at the mean?”

S3: “It does say a lot.”

T: “What does this say about the boys and girls?”

S3: “They have done it much better.”

T: “No, this is about using the computer for hours per week.”

S4: “With the boys you subtract a lot...”

T: “The boys use the computer almost twice as much as the girls.”

S3: “More than twice as much.”

T: “More than twice. So the boys are on average working 8 hours per week with the computer, and the girls just 4. When we look at the measures of spread, then these measures of spread do not tell that much, but have a look at the range. What does the range tell? S1?”

S1: “That girls more, that there is a girl that works more with the computer than the boys...”

S3: “Or the shortest, right?”

T: “Maybe more than one.”

S1: “Yes.”

T: “Maybe more than one. The range of the girls is bigger. That means, there are some girls, i don't know how many, we'll see that later, how many, who spend many times more time working with the computer than a boy.”

(0:16:00.4)

S3: “It can be shorter, can't it?”

T: “No, the spread is larger.” He points at both ranges on the whiteboard: 50 for boys, 60 for girls.

T: “The highest minus the lowest observation. And that is larger with girls.”

S3: “If you have 50-20, if you have them between20 and 50, then you get 30, and if you've got 10 and 50 then you get 40, then the variation is bigger, but then they work less hours with the computer.”

T: “Yes, but that story isn't completely true. And where that's exactly, it stays larger.”

S3: “Yes, but I ask if it's possible that they perhaps spend less time.”

T: “No, if i just look at this, i see that the range with girls is larger, either if i start at10 hours per week or a 0, that doesn't really matter.”

[17:02]

T: “In any case, the smallest observation and the largest observation is larger with the girls.”

S3: “Yes"

T: “And that's smaller than with the boys.”

Teacher looks at S3. “Do we talk past each other, S3?”

S3: “Yes. You just do not answer my question.”

T: “Yes.”

S3: “[inaudible] because if a boy at least spent 10 hours working, I just pick something, 10 hours of working...”

S2: “On average.”

S3: “On average. And the highest is 50, then you get a number of 40, don't you? Then you do 10 minus 50, umm 50 minus 10?”

T: “This is about all the boys. I draw a sample, you confuse two things, you talk about 10 hours on average, and what is then the highest number? You say 10 hours on average per week...”

S3: “No, not on average. The spread, okay, just another, if they have a spread between 20 and 50, then...”

T: “The smallest observation is 20 and the largest is 50?”

S3: “Yes, and then the range is then...”

T: “30"

S3: “30, yes, but if you've got, umm, 50 and 10 then you have a measure of spread (S3 means ‘range’) of 40. But then they could have worked less with the computer.”

[18:30]

S3: “If you just look at, if you have a range between 10 and 50, then one girls has worked 10 hours?”

T: “Yes.”

S3: “Or hasn't she?”

T: “Yes.”

S3: “And if the spread is between 20 and 50, then one boy has, the least, worked 20 hours?”

T: “Yes.”

S3: “So then the girls can have spent less time, and then the measure of spread (he means 'range') can still be larger.”

T: “Yes, that I see with this measure.”

The teacher points at the mean of the hours with girls, 4.01.

S3: “Yes, but I'm talking about that.” He points at the maximum and the minimum of the data set.

S3: “So it is a general question.”

T: “Yes, but this is not a measure of centre but a measure of...”

S3: “Measure of spread.”

T: “Yes. So that says something about the distribution of the numbers.”

S2 “You two talk completely past each other. I think.”

S3: “Just let it go. I know myself that I'm right.”

Students laugh.

S3: “Or I do not understand what you're saying, that's also possible.”

[19:28]

S2: “I agree that S3 is right, but the two of you are talking past each other. Because you [he means the teacher] are also right, but then at another point.”

S3: “It was just a question…”

T: “I look at these measures that I have here...”

S3: “Yes, but I talk in general.”

[20:00]

Another two minutes of repetition of arguments.

S2: “It's about the range. You said that that meant that there were girls that more frequently...spent a lot more time on computing.”

T: “Yes.”

S2: “And then he said: that can be shorter too. That, if someone has a lot more and another one a little less, that it together becomes the same again.”

T: “Yes, well, that's exactly what I'm trying to say. But that wasn't received apparently.”

[21:42]

Coded implemented feedback Ex8.8 S4b

Ex 8.8 *SLG*=DL → *SII*(h(8), h(14), DL) → *TC* (l, DL)

Correspondence score: 4.5.

Interpretation classroom discourse Ex8.8 S4b

Results of section 8a are presented on the whiteboard.

Again the teacher does not use ClassAnalysis in order to start the feedback session. Instead, he starts with asking “Why do you have to give a measure of spread when giving a measure of central tendency?” [1]. The teacher then stops S1 from again leading the discussion (see previous example about S1). The discussion comes to the standard deviation. A spontaneous remark from S3: “That means how far it is away from the mean, doesn't it?” The teacher agrees this is true: “Yes, deviation from the mean.” [2] He presents an example and invites S3 to comment with respect to the spread on this example. S3: “The first is from 5 until 7, the second is from 3 until 9.” This answer suffers from the same problem as the second last S3 gave [3]. The teacher again does not give this much attention, but starts explaining [4]. Then the teacher refocuses on the exercise and the context itself. There was a little misunderstanding between the teacher and S3, who is referring to the 'test context' while the teacher is back at the 'computer use context'. Then they discuss the computer use and the spread in it. At some point the teacher mentions: “When we look at the measures of spread, then these measures of spread do not tell that much, but have a look at the range. What does the range tell? S1?” He addresses this question personally, which is very good. But we do not understand why he mentions that the measures of spread do not tell that much. A SD twice as big, when it concerns data sets of more than 400 entries, is very convincing [5]. S1 gives a remarkably subtle answer: “That girls more, that there is a girl that works more with the computer than the boys...” [6] The teacher reacts with “Maybe more than one” on which S1agrees: “Yes" [7]. Then S3 comes into the discussion with: “It can be shorter, can't it?” with which he means that the larger range could be caused by an outlier on the minimum side of the range [8]. The teacher completely misunderstands this, because he thinks that S3 is concluding about this specific context, and they start discussing this. S3, as we have seen in two earlier contributions he made to the discourse, does not always formulate as carefully as this subtle matters require and the teacher is unable to take another

perspective. This is also hindered by the fact that S2 also contributes to the discussion by saying “On average”, too easily agreed upon by S3 “Yes, on average”, because this is not the case with the range. The range has nothing to do with the mean, only being concerned with the minimum and the maximum of a data set. This ‘easy agreeing by S3 with S2’ is fatal, because the teacher then focuses on this incorrectness, making him possibly more concerned of his own correctness. At 19:28, S3 surrenders: “Just let it go. I know myself that I'm right.” Students laugh. S3: “Or I do not understand what you're saying, that is also possible.” [9].After this, S2 plugs in into the discussion. He tries to show that both the teacher as well as S3 has a point from their own perspective. But these matters are that delicate that S2, although quite eloquent, does not succeed either [10].

Remarks on the classroom discourse Ex8.8 S4b

[1] This is, of course, a good question, but it does not correspond to the first learning goal of exercise 8.8, according to the HTT: let the students realise that only using a measure of spread as a representation of a data set causes a specific loss of information. The teacher gives this away by asking directly about the ‘why’ of a measure of spread.

[2] Here is a problem seen in a lot in mathematics education: a student brings in an intuitively good idea that lacks mathematical rigor. Defining the standard deviation as ‘how far it is away from the mean’ brings immediately the question in mind: what is ‘it’ that is away from the mean? One could compare the standard deviation with the mean (absolute) deviation from the mean. There is a difference between both measures, but for this target group we do not problematise this. The didactical dilemma is: you want to reward a good student input, but you also want no conceptual mathematical mistakes. Further, you want the discussion to stay focussed on your learning goals, and not to get distracted to mathematical details that could confuse the lesser gifted students. The teacher chooses here not to go deeper into the technical mathematics and to reward the student's input.

[3] S3: “The first is from 5 until 7, the second is from 3 until 9.” Again, we could ask: “The first ‘what’? The second ‘what’?” S3 suggests some kind of frequency distribution, but reasoning like this does not belong to the targets for this intervention.

[4] Explaining has the danger in it of monologueing. As this is a classical component of almost all teachers' repertoire, it is hard for some teachers not to fall back on.

[5] Apparently the teacher wants to focus on the range specifically. That is logical, because it is the only measure that could indicate that there are outliers but these outliers can only be outliers when the other measures of spread are relatively (with respect to the other data set) small.

[6] S1 starts his answer with “That girls do more…”, but he apparently realises that he cannot state this in general, so he changes this immediately into “That there are some girls that spend a lot of time using the computer.” The teacher for some reason is too busy with conducting the classroom discourse that he does not reward this correction as much as we think would be just. The fundamental difference is that between SD and IQR on the one hand and the range on the other hand, is that the SD/IQR include all/50% of the data and the range just includes two.

[7] T: “Maybe more than one.” S1: “Yes.” The teacher is right that there could be more than one outlier. S1 is right to agree on it, but that is not guaranteed by the range which solely includes the very extremes.

[8] In principle, this is a very lucid remark. A larger range is not automatically caused by an outlier on the right hand side of the spectrum (maximum), but could just as well be caused by an outlier on the left hand side of the spectrum (minimum). The elaboration

of this remark will take the rest of the classroom discourse on this exercise (5:42), in which the teacher and S3 have an intense discussion. In this discussion there is an intermediating contribution of S2, who very sharply sees at which point the teacher and S3 do not understand each other.

[9] S3, who, from a general perspective is right with his opinion, apparently has the feeling that the teacher is not sensitive to his arguments. Remarkable, given the effort he laid in the discussion, is his nuance that it is possibly his own fault in 'not understanding the teacher's arguments'. Perhaps this nuance is fed by the fact that other students laugh when he mentions that he is convinced of his own correctness.

[10] All of this discussion would have been so much easier if the teacher had used the network to show the box and whisker plots of both data sets. Then you could see that S3 has a principal point in stating that a larger range could be caused by a smaller minimum in the data set and not just by a larger maximum. However, in this case, the minimum of both the boys' data set as well as the girls' data set is the same: zero hours of computer use per week, see figure 7.3: left the computer behaviour of the boys, right that of the girls. That means that in this specific case S3 cannot be right so a difference in range has to be caused by a difference in the maximum.

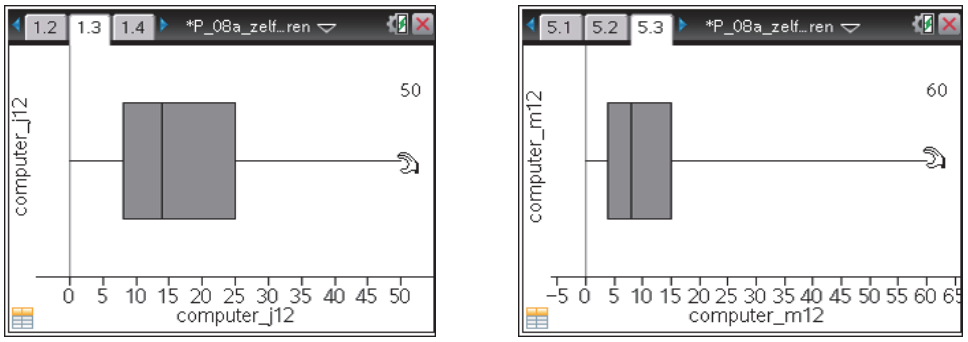


Figure 7.5 Boxplots that represent the use of the computer by 12-year-old boys (left) and 12-year-old girls (right)

[11] The discussion as described above, between the teacher and S3 and later S2 is extremely remarkable for this level (grade 10 senior secondary education). We suggest that making use of the network would probably have clarified the discussion for everyone. It is remarkable that the lesson from which this example is sampled was just a couple of hours after the same lesson in the other intervention group at this school. During the discussion of this exercise the teacher was intending to show the boxplots as presented above by using the network (but he could not do so because he ran out of time). Why does he not show this intention right now while he could really use it? We presume that this is again a matter of immature internalisation of the network as a didactical tool. Although the teacher is an experienced user of ICT, there seems to be more going on here. The possibilities of the network as well as the didactical approach following directly with these possibilities have to be internalised. For some teachers this takes more time than for others.

7.11 Feedback in the sixth case study S2b

7.11.1 Starting point and overview S2b

Starting point as perceived by the teacher

Below we present the starting point, as perceived by the teacher and reported in the first part of a questionnaire.

The teacher was female, 54 years old, and had 14 years of experience. The relationship with the group she described as moderate (2 on 4 point scale).The level of the group she rated as poor (1 on a 4 point scale). Her own ICT competence, she considered to be moderate (2 on a 4 point scale).

Correspondence between HTT and implemented feedback

In Table 7.8 below we present the overall correspondence data. We respectively present: the mean total correspondence, the standard deviation of the total correspondence, the mean correspondence with respect to data literacy exercises, the mean correspondence with respect to algorithmic statistical skill exercises and the subtraction of these means.

Table 7.8 Correspondence score characteristics of case S2b

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL-M_ASS	%-Missing
S2B	2.04	1.70	1.95	2.25	-0.30	59.57

We see a very low average corresponding score and a very high percentage missing. This may be surprising because this was the last case study we did so we brought a maximum of experience.

We observed a bad relationship between the teacher and students.

We will try to illustrate the gap between intended (as described in the HTT) and implemented feedback (as described in Table 7.8 and in the examples below).

7.11.2 Classroom discourse example S2b-1

Exercise 1.9 (DL, OA)

The page below shows again the average maximum temperature of the month of September during the years 1971 to 2000, sorted by size. What is the range of the average maximum temperature in September during those years? How much warmer was September 2005 than the warmest month of September in the period 1971-2000, as measured by the mean maximum temperature? Do you consider that to be a big difference?

HTT Ex1.9

The steps to be taken are not that difficult and are accurately described. However, the reasoning that has to be done and the conclusion to be drawn could be unusual for these students (“In 2005 September was 1.1 degree Celsius warmer than the maximum of the Septembers during 1971-2000, which ranged from 17.1-19.2 degrees Celsius. I consider that gap to be big enough to call September 2005 a warm month.”), hence feedback will most likely be needed.

Coded HTT Ex1.9

Exercise 1.9 *SLG*=DL → *CAF* (h, DL) → *SII* (h, h, DL) → *TC* (h, DL)

Implemented feedback Ex1.9 S2b

[23:44]

T: “Of the September months during 1971-2000, so over thirty years, and they are ordered by size. Now it is a question about the range, but...oh wait a minute.”

Students laugh. The teacher seems to show the students' responses a little accidentally. ClassAnalysis shows 4 substantial reactions, 9 non-response (on a total of 20 students).

T: “There were not many reactions.”

S1: “Which section is this?”

T: “This is 1c. Section 1c. And there are very few people...”

S1: [inaudible].

T: “What do you say?”

S1: [inaudible]

T: “I take a quick look at the other questions.”

[24:23]

Some further ClassAnalysis with general teacher comments. The picture is the same: 4 students have responded substantially.

[25:14] Back to exercise 1.9.

T: “There is a question about the range of these data, about the absolute dispersion. Do you have an idea what that is?”

S2: “Spread of the numbers, absolute numbers.”

S3: “The first one was 17, the minimum 17, maximum 29, that is...”

T: “Listen you all to what S3 says.”

S2: “And to me!”

T: “Yes, no, but I, S3, I understood with him, if I translate it, you look at the largest?”

S3: “Yes. The largest is, umm, 17 is the minimum temperature...”

[26:10]

T: “Okay, the minimum, the difference between them.”

S3 nods affably.

T: “Well, that that would be a spread a measure of spread. The difference between the largest and the smallest. Umm, what was in September 2005, that you have calculated yourself, the maximum temperature, the mean temperature? Did you find a mean temperature over the years 1971 until 2000?”

S2: “18.5”

T: “Okay. It is to me that it was 18.25.”

S4: “That I calculated too.”

T: “18.25. If you had calculated that well. What did you find for September 2005? S3?”

S3: “Is that this question?”

T: “Does anyone remember?”

S5: “No.” [Very definitely]

T: “Does anyone know...because with that, you have to compare, the exercise says.”

[27:15]

T: “How much warmer was September 2005?”

S6: “That was 20.51.”

T: “Exactly, That was 20....”

Students discuss rather loud.

T: “Do you consider that to be much warmer?”

S4: “Yes.”

S7: “No.”

T: “You think so. And you don't. Why do you think this is true?”

S4: “Umm, yes, well, I don't know.”

T: “You've seen the complete list.”

S4: “Yes.”

T: “Was 20 present in it?”

S4: “Yes.”

S8: “That can be an outlier!”

T: “That can be an outlier, yes.”

S3: “It is about [inaudible] outlier.”

S4: “No, 20 didn't show up in that list.”

T: “20 didn't appear, no.”

S4: “19 at maximum.”

T: “Yes.”

S4: “19.3.”

S2:“1000 [inaudible].”

S3: “That is a pretty big difference.”

T: “You think it is a pretty big difference? First, you didn't think so.”

[28:13]

S3: “Now I do.”

T: “Now you do. Was September 2005 a warm month?”

Students talk through each other.

T: “Do you all consider September 2005 being warm, do you all consider this to be remarkable?”

S4: “Yes.”

T: “Because I see, there is someone who doesn't consider this remarkable.”

S3: “Yes, three quarters hasn't done their homework, so how can you consider...”

S9: “Those 2degrees you won't notice. No one will be bothered.”

T: “If you look at it that way.”

S10: “Madam, how do you make something [inaudible]?”

T: “That, I don't know. Shall we...”

She again shows the slide with the students' responses on this exercise.

T: “In the beginning, ‘no’ and also ‘no not really’ (student responses shown on the slide) but meanwhile, I have the idea that we changed our opinion, didn't we? S4, but S3 also.”

S3: “Because 20 doesn't appear.”

S2: “That is striking, though.”

T: “Remarkable difference, right.”

[29:25]

Coded implemented feedback Ex1.9 S2b

Ex 1.9 *SLG*=DL → *CAF* (0) → *SII* (h(7), h(12), DL) → *TC* (h, DL)

Correspondence score: 5.

Interpretation classroom discourse Ex1.9 S2b

Teacher starts with CA, reading aloud the question. When she clicks through to the students' responses, seemingly accidentally, students laugh [1]. The teacher mentions that these results are disappointing and clicks through to the results of subsequent questions. These show the same picture: just 4 students responded substantially. Back to exercise 1.9. The teacher poses a question (“What is the range?”) addressed generally [2]. S3 reacts quite well. He uses in his answer the concepts ‘minimum’ and ‘maximum’. The teacher tries to let S3 give a clear definition of range (absolute dispersion), but this does not come out easily. Then, at [26:10], she drops it herself [3]. S3 agrees with that, in an affable way [4]. The teacher was not always very concise in her wording, as is witnessed for example by “Umm, what was in September 2005, that you have calculated yourself, the maximum temperature, the mean temperature?” [5] And after this:

T: “…S3?”, S3: “Is that this question?”, T: “Does anyone remember?” , S5: “No.” [Very definitely] [6]

With the range of the maximum temperatures of September over 1971-2000 now operational in the classroom discourse, the teacher asks for a mean maximum temperature over this period [7]. Then she asks what the mean temperature of September 2005 was, and how September 2005 is to be compared with these. In this comparison the teacher comes to this statement: T: “You think so. And you don't. Why do you think this is true?” [8].This evokes some discussion. Then S3 states that there is a pretty big difference. The teacher mentions that this means he has switched his opinion [9]. She then asks if everybody is convinced that September 2005 was warm. Then S9 mentions: “Those 2

degrees you won't notice. No one will be bothered." The teacher reacts with "If you look at it that way" [10]. She reshows the initial students' responses and concludes that the group has now decided that September 2005 was a considerably warm September month.

Remarks on classroom discourse Ex1.9 S2b

[1] It is not very clear what they are laughing at. The teacher is not a very fluent ICT user (self-rated 2 on 4 point scale), which she does not hide ("…oh, wait a minute…"), but the disappointing results of the students' responses (group of twenty students, 13 logged on, 9 non-responses, 4 substantial answers) is the most likely reason for consternation.
[2] The teacher could have used the information from the network. There are just 4 substantial responses. But the classroom discourse can be fed with these responses.
[3] In the interaction between S3 and the teacher, neither formulate their responses very sharply. It is possible that the teacher considers the interaction will not lead to S3 giving a concise and understandable answer, so she gives it herself. She does not try to give the floor to another student.
[4] Later on during this intervention, we would call this 'a jaded way'. The way a lot of students behaved towards the teacher did not seem very respectful.
[5] The teacher was always very concise in her wording. She mentions here 'mean temperature' and 'maximum temperature', but for September 2005 the mean maximum temperature had been calculated.
[6] There is not a very productive cooperation between the teacher and group.
[7] It is wise to calculate a measure of central tendency for the maximum September temperatures during 1971-2000. However, the range [17.2, 19.3] is not to be forgotten, because the limits of this range tell something about the spread and thus about how likely mean maximum temperatures are that fall outside this range.
[8] With two students diametrically disagreeing you can hand over the responsibility for the discussion to those students. We have seen this in the first case study in which the teacher played it that way. This was communicated with the teacher in this case study. She nevertheless was not able to use this strategy here.
[9] The teacher concludes that S3 has switched his opinion: first he did not consider that there was a big difference between the long term data and 2005. Then he does. However, she does not ask him what exactly he considers to be this big difference, which is likely to be the comparison of the means. Further, she does not ask him what convinced him then. S3 is not saying it spontaneously.
[10] "Those 2 degrees you won't notice. No one will be bothered" S9 mentions. Now this is exactly the point. Will you really not notice 2.25 degrees? And will no one be really bothered? What kind of change in ecosystems could this induce? The teacher is laconic: "If you look at it that way." She does not appear to take this input seriously.

7.11.3 Classroom discourse example S2b-2

Exercise 1.12 (DL, TF)

In December 2009, in Copenhagen, the United Nations Climate Change Conference was held (http://en.cop15.dk/). During this conference, UN member countries discussed steps which should be taken against alleged global warming.

Poll (true/false): your calculations on the temperature in September 2005 are evidence for global warming.

HTT Ex1.12

At the end of section 1 the calculations of and reasoning by the students is drawn in a broader perspective. What conclusion is justified?

In order to force the students to sharply choose sides, this exercise was designed as a 'True/False' item. The teacher uses ClassAnalysis to show the results. After this election like outcome, a debate should be easily organised between those who believe that the calculations are evidence for global warming and those who do not. The latter are correct: the calculations just count for one specific month on one specific place (De Bilt, The Netherlands).

Coded HTT Ex1.12

Exercise 1.12 *SLG*=DL → *CAF* (m, DL) → *SII* (h, h, DL) → *TC* (h, DL)

Implemented feedback Ex1.12 S2b

[31:42]

The teacher uses ClassAnalysis slideshow. First she repeats aloud the question.

T: "Could it be that September 2005, if that, that is a warm month, we think, when compared with the months, with the years before. Is that a proof, do you think, for the statement that the earth is slowly warming up?"

Meanwhile she clicks through to the results bar chart.

S1: "But this is an opinion issue."

T: "Yes, this is an opinion issue."

S1: "So you can't be right or wrong. Coincidentally, I've done it right, of course. But..."

T: "Here are three people who think it isn't true."

S2: "That's because the measurement is too short."

T: "Okay. Therefore S2 says...could you repeat it?"

S2: "Therefore the measurement is too short. Because you can measure over five years, but that goes by peaks and valleys over this period."

[32:32]

T: "So September could by chance be a peak, I understand, that September 2005, that could be an outlier?"

S2: "Yes."

T: "That could be. Yes."

S2: "That doesn't necessarily mean that the earth is warming up."

T: "No, that is not…and where is it...it depends also...that's a good, good idea...does anyone have…can anybody still comment on that?"

Silence.

T: "Can anyone call something about this? Maybe? It can be a coincidence. It can be an outlier, September 2005, it was measured once...it was also measured on a certain place in the world, of course, but there's one person who thinks it is true though. Why?"

S4: "He probably just has clicked something."

T: “Oh. Yes, that's possible.”

S5: “That guy pushed the wrong button.”

T: “Well, I'm gonna stop with this.”

[33:38]

Coded implemented feedback Ex1.12 S2b

Ex 1.12 *SLG*=DL →*CAF* (m(2), DL) →*SII* (m(2), m(4), DL) →*TC* (l, DL)

Correspondence score: 4.5.

Interpretation classroom discourse Ex1.12 S2b

The teacher starts with ClassAnalysis and reading aloud the question. S1 is apparently a little surprised about this type of question in the mathematics lesson. The teacher takes his input literally and reacts thus [1]. Then she mentions that there are three people who think it is not true [2] (see Figure 7.6).

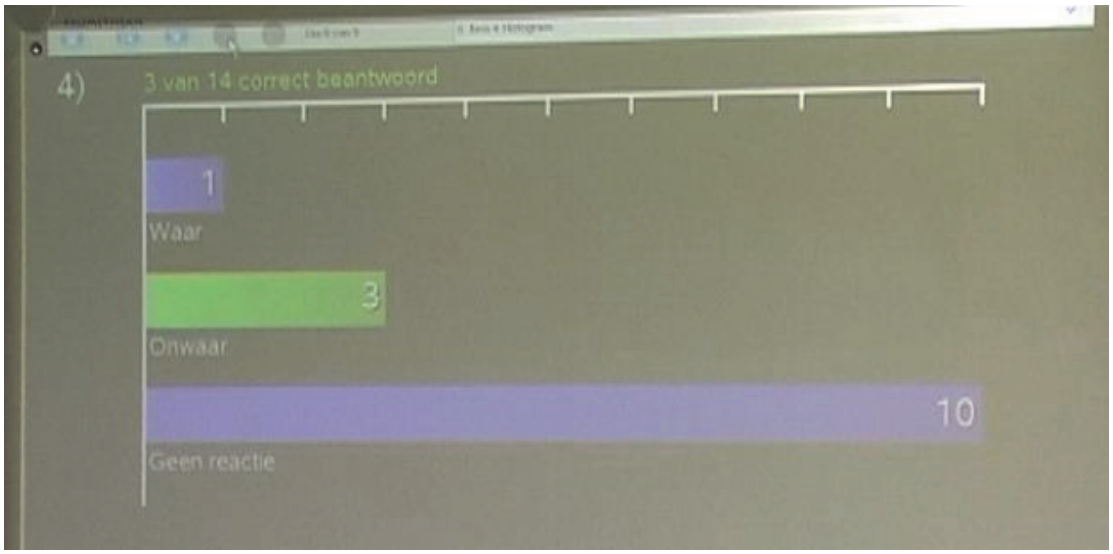


Figure 7.6 Students' results on exercise 1.12 in case S2b, as represented by the CN

She picks out S2, who starts with mentioning: “Therefore the measurement is too short" [3]. S2 has a good point, though. The teacher thinks so too and reacts with: “No, that is not...and where is it...it depends also...that's a good, good idea...does anyone have…can anybody still comment on that?” This does not evoke further questions [4]. She then tries it one more time, stating that the calculations had been performed on just one place on earth [5] and that there was someone who at least had thought that the calculations were evidence for global warming [6]. The students act as if this should have been a 'typo'. The teacher then cuts the discussion off [7].

Remarks on classroom discourse Ex1.12 S2b

[1] “This is an opinion issue," S1 mentions. This seems to surprise him. On this point there could be two interesting discussions possible: do opinions belong to reasoning? What is the value of non-evidenced opinion? Does the formulation of an opinion about something you have explored with statistical techniques belong to statistics? The teacher does not appear to believe so.

[2] With this, she uses the information from the network in a productive way for the classroom discourse.
[3] A very logical question would be: what kind of a period do you not consider to be too short? The teacher does not pose this question.
[4] Although S2 draws a very fine conclusion (based on arguments that could perhaps be broadened a little), the teacher reacts somewhat phlegmatically. Further, in the wording, there are perhaps too many hesitations to be convincing for this group.
[5] This relativisation of the calculations with respect to global warming could have been provoked from the group with just one intermediate question like: What does global warming exactly demand? Global in time, which our calculations are not, as S2 reasoned, and global in…?
[6] She does not switch to the student view of CA to look up who this person was.
[7] The discussion between teacher and group does not seem to be based on mutual trust. In C1, we have seen this too. This makes it very hard, if not impossible, to set up a fruitful classroom discourse. Good education is an intimate process and feedback and classroom discourse are perhaps the most vulnerable parts of it. Mutual trust is indispensable in establishing them.

7.12 Conclusions with respect to the pilot of the third prototype

7.12.1 Conclusion in general

Before we draw conclusions per case study, we present general conclusions pertaining to all six case studies conducted.

As a starting point we recall the correspondence scores (on a scale of 0 to 7) of the six case studies during C3, now presented in Table 7.9.

Table 7.9 Correspondence score characteristics of cases in C3

Case	Mean	SD	Mean_DL-Mean_ASS	%-Missing
S1	5.14	1.70	-0.71	68.09
S2a	4.40	2.29	-2.05	14.89
S3	5.38	1.85	-0.12	0.00
S4a	3.60	2.13	0.21	31.91
S4b	3.89	2.09	0.78	34.04
S2b	2.04	1.70	-0.30	59.57

When inspecting these results, we see that there is a considerable range in mean correspondence. For two case studies (S1, S3), there is a strong correspondence, for one case study (S2a) the correspondence is reasonable, for two other case studies (S4a, S4b) the correspondence is moderate, and for case study S2b the correspondence is weak. We think we may conclude that when three cases out of six show a correspondence varying from very reasonable to strong, according to teachers and students, it is possible to improve feedback. Further, we observe that the percentage missing feedback sessions also varies quite a lot: from very high (S1, S2b) through moderate (S4a, S4b) and low (S2a) to very low (S3).

With respect to the difference in correspondence scores between feedback on DL and feedback on ASS we notice that there is not an obvious difference in correspondence between feedback on DL activities and feedback on ASS activities. Case study S2a is the exception: the feedback on ASS here is substantially more corresponding to the HTT than the feedback on DL was.

We further note that the standard deviation does not differ dramatically across the case studies. The case study with the lowest mean score has the lowest standard deviation and the case study with highest mean has the second lowest standard deviation. This means that in the case study that is worst predicted by the HTT, these sessions have a relatively equal low level of predictability. In the case study where the feedback sessions were well predicted sessions have a relatively equal high level of predictability. One could say that this makes these case studies more 'easy' to interpret.

When realising that the chronology in C3 stretches from S1 to S2b, we see that we did not manage to reach 'successive approximation of the ideal intervention' (van den Akker, 1999) over the six conducted case studies. Given the fact that we used micro cycles of development during C3, offering us the opportunity to use the actual learning experiences to improve our prototype in-between the case studies, one question comes to mind: what hindered successive approximation? Apparently, the influence of 'external factors' was stronger than those we were trying to influence with our intervention. What were these 'external factors'? We try to answer this question by comparing the conclusions on the successive case studies.

7.12.2 Conclusion S1

For case study S1 the characteristics of the correspondence score are recalled in Table 7.10.

Table 7.10 Correspondence score characteristics of case S1

Case	Mean	SD	Mean_DL-Mean_ASS	%-Missing
S1	5.14	1.70	-0.71	70.21

On a scale from 0 to 7, this case study comes with a mean correspondence of 5.14, fairly close to our intentions as formulated in the HTT. We did not succeed in preparing the case study technically well. There was far too much technical failure, as is expressed in the very high percentage of missing sessions. We found practical solutions for these problems, but that strongly hindered the fluency of teaching activities in the classroom. The teacher experienced this way of teaching as "*extremely stressful*", but when the network was up, the teacher nevertheless succeeded in conducting a classroom discourse that came fairly close to what was intended. This was possibly due to the teacher's experience, because he participated in 2004 in the initial study, to his good ICT skills, or to his professional behaviour that includes *functional extraversion:* the capacity to be a leader with respect to the students' learning input in the classroom discourse. Acting this way allows him to hold the students personally accountable for their work and to discuss their work with them.

The teacher perceived that he had more "*feedback opportunity"*. The students perceived more feedback. Neither the teacher, nor the students experienced a big difference in the usefulness of the network for ASS or for DL development. Both teacher and students considered that this way of supplying feedback deserves a place in the mathematics

education of the near future. The different mathematical capacities of the students did not seem to cause a big difference in how they experienced this way of working.

7.12.3 Conclusion S2a

For case study S2a the correspondence score characteristics are recalled in Table 7.11.

Table 7.11 Correspondence score characteristics of case S2a

Case	Mean	SD	Mean_DL-Mean_ASS	%-Missing
S2a	4.40	2.29	-2.05	14.89

From the mean correspondence score of 4.40 we conclude that there was quite a similarity between HTT and practice. This was especially true for feedback sessions on ASS activities. For the DL activities, the teacher perhaps should have behaved more *functionally extravert* (like the teacher in the case study S1): use the network data with respect to the students more personally and rebound students' input in the classroom discourse as much as possible. Perhaps a better relationship between the teacher and group would have made this easier for him. The teacher made a very modest impression, as is witnessed for example by the self-assessment of his ICT skills. Modesty is perhaps not positively correlated with *functional extraversion*. We observed a similar situation during C2 (see chapter 6), where a teacher of the same sex and age with similar experience and comparable ICT skills and attitude did not reach the full potential especially during the DL feedback sessions. This teacher also was rather introvert and modest with a less strong classroom presence.

The students in this group were unmotivated in learning mathematics. Working with feedback sessions, in their opinion, did not increase their contributions to the classroom discourse. On the other hand, although this did not appear from the questionnaire, the students in the interviews thought that their homework discipline was better than before. They did not consider the new way of working to be threatening. The teacher and students agreed on the fact that the lessons were too monotonous. Almost all of the lesson time was dedicated to the feedback sessions. There is consensus that these feedback sessions should take about 50% of the lesson time. The teacher was rather positive about his lessons during this case study. He agreed with the students that they worked harder than before. The results on the summative test were better than he had expected.

We further note that there were some very good feedback sessions during the first lessons, proving that it is not necessary to train for months in order to able to lead a feedback session the way we intended it.

7.12.4 Conclusion S3

For case study S3 the correspondence score characteristics are recalled in Table 7.12.

Table 7.12 Correspondence score characteristics of case S3

Case	Mean	SD	Mean_DL-Mean_ASS	%-Missing
S3	5.38	1.85	-0.12	0.00

As seen from HTT perspective, this was the best case study out of the six that were part of this study. The mean correspondence score (5.38 on a scale from 0 to 7) is high. The

standard deviation is low, indicating a steady level of correspondence. The difference between DL and ASS sessions is negligible. The percentage missing (0) is impressive.

What were the main findings of this successful, and hence important, case study? We made an important decision to rearrange the prototype. We forced a split between sections that could be self-checked by the students, mainly on ASS, and sections with a focus on DL, where the teacher supplied feedback as the start of an interactive classroom discourse. The teacher and students agreed that this was an improvement.

The teacher showed almost exactly the behaviour he was intended to. Not only did he prepare the lessons very carefully – some other teachers involved in this study perhaps equalled him on this – but he was able at the same time to become fluent with the technology and translate this fluency into a valuable information stream which he used to set up a powerful classroom discourse. He was very straight in his classroom management: students who forgot their GC were sent away without discussion. The observed classroom discourse was highly interactive. It was strongly focused on the mathematical task. The students were engaged. The teacher addressed questions personally, using very fine tuned questions and rephrases of student input, as initially collected by the classroom network. The teacher used the network possibilities almost as a routine. Looking at the classroom discourse we note that our prototype might ideally be a setup to a feedback oriented interactive classroom discourse. However, fulfilling this potential requires quite a lot of the teacher's performance and perhaps even of his personality, because the acting desired in the classroom discourse seems to ask for a *professional extraversion*. If a teacher is not naturally extravert this is very hard to develop.

From interviews and questionnaires we obtained data supporting the above. Students and the teacher agreed on the fact that there was more and better feedback, both on ASS as on DL. Students felt safe in contributing to the classroom discourse. They advise dedicating about half of the time to this teaching activity. The teacher would like to restrict this to one third.

7.12.5 Conclusion S4a and S4b

For case studies S4A and S4B the correspondence scores are shown in Table 7.13.

Table 7.13 Correspondence score characteristics of cases S4a and S4b

Case	Mean	SD	Mean_DL-Mean_ASS	%-Missing
S4a	3.60	2.13	0.21	31.91
S4b	3.89	2.09	0.78	34.04

The results are quite similar. Apparently, given the fact that both groups differed at least somewhat (according to the teacher), the correspondence score is strongly determined by teacher characteristics. This is consistent with our prior observation that these teacher characteristics, which we will describe in section 8.2.1 , are important in determining the degree of success in a case study.

In both case studies there was a moderate correspondence (mean 3.60 and 3.89 respectively) between intended and realised curriculum. The teacher and students agreed the most important characteristics of the prototype. They experienced more feedback, on both ASS and DL. The students were more engaged with the mathematics and the

classroom discussion more focused on this. The fact that the lessons of S4b were always on the same day as those of S4a, but just a few hours later, could explain why the correspondence for S4b is slightly better (3.89 versus 3.60) especially on the DL sessions (4.11 versus 3.65). The teacher and the principal researcher always evaluated the lesson directly afterwards. This could have had an immediate learning effect.

The teacher sometimes showed that he was not completely adapted to the new way of teaching. For case study S4a the variation coefficient (the rate $\frac{SD}{Mean}$) was the largest of all case studies (except for S2b, which was an outlier, as we will see in the next section). During the observations we felt that the teacher behaved somewhat whimsically. In the concluding interview he mentioned that he recognised the feeling of having to learn teaching again, because he needed that much attention for the technology and the process of feedback itself, he sometime felt like he was forgetting his students. Some specific aspects, like addressing questions personally as much as possible, he did not master well enough during the course of the intervention. His time management was too optimistic on many occasions.

Teachers as well as students were enthusiastic about the improved feedback, both on ASS and on DL activities. They stress the additional learning effect of collaborative working. There is one student interviewed – with low mathematical skills – who considered her public contribution to the discussion to be threatening. The other five considered this to be unproblematic. On average, they thought that about half of the time during the lessons should be spent on this teaching activity. The teacher agreed with them on this too.

7.12.6 Conclusion S2b

For case study S2B the resulting correspondence scores are recalled in Table 7.14.

Table 7.14 Correspondence score characteristics of case S2b

Case	Mean	SD	Mean_DL-Mean_ASS	%-Missing
S2b	2.04	1.70	-0.30	59.57

During this case study, we obviously did not reach our goals. When compared to the HTT, not enough feedback has been given compared with the five previous case studies. The correspondence of the feedback and the thus initiated classroom discourse was very low. The teacher stated in a personal interview: "*I was very dissatisfied with the lessons. There was nothing good about them.*" This may be a too negative image of the case study, but in a way she has a point. Which factors caused this relative low correspondence with our intentions?

To start, and perhaps end, with: the relationship between the teacher and the group was not good. The teacher described it as 'moderate' (2 on a 4 point scale). In an interview, a student describes the attitude of his fellow students as very unmotivated. Another one of the interviewed students points at the teacher: "*She just wasn't a good role model; she lacked control over the group"*. In both of the examples of classroom discourse (describing interaction that came most close to the HTT as observed during this case study) students showed in our view disrespectful behaviour. The teacher sometimes appeared confused, which was perhaps a reason for the students to act as if they did not take the lessons and/or the teacher (it is often hard to draw a sharp line between these in education) seriously. The teacher in both examples sometimes acted as if she did not really care about the outcome of the interaction. From the personal discussions before and

after each lesson, we know that this is not true, but observed from a distance, it certainly could be interpreted this way. As we saw during the intervention in cycle C1 (see chapter 6), a bad relationship between the teacher and group is lethal for feedback and classroom discourse.

The teacher self-rated her ICT skills as 'moderate' (2 on a 4 point scale). This, in interaction with her sometimes tentative acting, contributed to what one interviewee called 'a lack of control over the group'. The questionnaire showed that students considered the given feedback as 'not better' than the usual feedback. The teacher agreed on this point. It may, from the perspective of the above, be surprising that students showed nevertheless confidence during the interviews in this way of working when looking to the future. Apparently, they could abstract from their concrete experiences. The teacher agreed with them, as she indicated in the questionnaire and interview that she saw the potential of the network to support feedback.

Chapter 8
Conclusion and discussion

In this chapter we first recapitulate our study. Then we answer the research questions, underpinning them with results from chapter 7. After that, we draw conclusions from this study. We follow this with a reflective section. Here we discuss the scientific and practical importance of the results. We look back at the design and research process and at the quality of the prototypes of the intervention. Also, we discuss the limitations of this study. We end with some recommendations: for the practice of mathematics education as well as for the design of classroom networks and similar technology. Further, we make recommendations for researchers who want to further investigate the use of classroom networks in mathematics education in order to enhance feedback.

8.1 Recapitulation of the study

8.1.1 Why this study?

In this study we designed, developed, implemented and evaluated three prototypes of an intervention aimed at the improvement of feedback in statistics education supported by a classroom network.

Many mathematics teachers in secondary education are of the opinion that teaching time falls short for attaining the usually ambitious learning goals. In The Netherlands, after the curriculum reform in upper secondary education in 1998 which included a reduction in contact time while learning goals did not reduce proportionally, this was probably felt more intensely. At the same time Dutch mathematics teachers seemed to “disappear in their lessons" because of the rise of self-supported learning by the students at the expense of classroom discussions (Onderwijsinspectie, 2001, 2011).

Besides this issue, there is a persistent problem of students' perceived difficulty of mathematics (Berch & Mazzocco, 2007; Geary, 2010; Hembree, 1990; Küchemann, 1981; Rosnick & Clement, 1980). Especially poor performing students in mathematics consider statistics to be difficult (Hong & Karstensson, 2002). This means that it takes time to master mathematics (and statistics), if possible in interaction with a teacher. However, this time seemed to be decreasing more and more.

How to solve this problem? Creating more hours in a day was not realistic. It was considered to be more feasible to concentrate on improving mathematics education within the constraints of the actual teaching time available. The basic idea of this study is to utilise teacher *feedback* on students' work as a starting point for more *teacher-student interaction* in the time spent in the mathematics classroom. That means concentrating on the quality of the teaching time, that is, to make mathematics education more efficient.

What would we need in order to improve interaction? We chose an environment in which we could utilise one-to-one computing: a classroom network with graphing calculators. In 2004, we conducted an intervention in statistics education utilising such a network, while asking: is there a possibility to improve the interaction between the teacher and students? This initial study resulted in a possible answer: *feedback*. This generated the main research question:

> What are the potentials of a classroom network in supporting teachers with providing feedback in statistics education?

In this research question we aim at feedback that is a stepping stone to a meaningful classroom discourse.

8.1.2 Research methodology

Our main research question starts with '*what*', implicating a *how*, that is: *exploring a mode*, eventually trying to find out *why* (an intervention succeeds or fails with respect to its goal). *Educational design research* (EDR) is supposed to be a logical paradigm for this type of research. Kelly (2007) states that design research is most appropriate for *open wicked* (Rittel & Webber, 1973) problems. Our problem could be *wicked*, because feedback, although a classical theme in educational engineering, is still not completely understood (Cohen, 1985; Shute, 2008) and not very well structurally implemented in classroom practice. Our problem was *open* since it is highly unlikely that there was just one way to just one answer. What is more, the technology we were exploring and its use were very new. They were that new we presumed we would need several iterations in order to create a teaching setting specific enough to yield data that could possibly lead to an answer to our research question.

First, we investigated the context of the problem: Dutch mathematics education and its curricular development and surroundings. In order to investigate the state of current knowledge we conducted a *literature study* with respect to feedback, statistics education and ICT. What does research tell us about these topics separately and, if possible, in combinations of two, or perhaps all three together? The results of the literature study served as an input for the formulation of sixteen *design principles*: seven on feedback and nine on statistics education. ICT principles were mixed in between. All together, these principles described function, form and content of the intervention to be developed.

All of these sixteen design principles were used to develop a first prototype, heavily drawing on exercises. The literature review with respect to statistics education gave rise to the essential distinction between *data literacy* (DL) on the one hand and *algorithmic statistical skills* (ASS) on the other. DL belongs to conceptual knowledge: know *why*. ASS partly belongs to declarative knowledge: know *what* and mostly belongs to procedural knowledge: know *how*. We distinguished on the feedback side *immediate* feedback (delivered by the students' GC) and *delayed* feedback (delivered by the teacher after an evaluation of the students' answers). We coupled both dichotomies into the feedback matrix as recalled below in Table 8.1. This matrix was the most important guideline for developing the teaching materials (exercises). From the research perspective we were particularly interested in what teachers would do or would fail to do, while using these materials and the feedback possibilities, in order to develop a meaningful classroom discourse.

Table 8.1 Feedback matrix for statistics education

Timing of feedback / Learning goal	Immediate	Delayed
Data literacy	Type I	Type II
Algorithmic statistical skills	Type III	Type IV

These types I-IV represent exercises on which a specific type of feedback is required. This matrix was used in order to develop exercises from all four types that together formed the teaching materials (in a digital format that is used on the GC).

What did the *basic teaching flow* of exercises through the classroom network look like? The exercises were sent from the teacher's computer to the students' GCs. The students completed the exercises and received immediate basic feedback on their work generated by the GC. The completed exercises were collected by the teacher's computer and analysed rudimentary. The teacher was able to inspect the results and to present (parts of) the results to the group. He could give feedback on the results and use this feedback as the start of a classroom discussion on the exercises.

As a topic and as a target group for the intervention we chose descriptive statistics for senior secondary education grade 10. Besides the teaching materials, most of which consisting of exercises on descriptive statistics utilising the GC and the classroom network (see Chapter 5), we developed with respect to the *intervention*:

- a student's manual meant to master the new handheld (used during C3) independently;
- a hard copy booklet containing all students' materials;
- teacher training material for the new GC and the classroom network hardware and software.
- a hypothetical teaching trajectory (HTT), mapping our teaching intentions on the students' exercises. This HTT also served as teacher training material, in order to prepare the teachers in an exact way.

For *research goals* we developed:

- observation forms;
- coding schemes and a correspondence metric for the observation data;
- the HTT now served as a reference with which the observation data were compared;
- questionnaires for teacher and students;
- interview protocols for teachers and students.

We piloted three prototypes of our intervention during three macro cycles (C1, C2 and C3). After each macro cycle, we evaluated the results and adapted the prototype according to the findings. During each macro cycle, we considered each lesson to be a micro cycle, in which the experiences during that lesson that possibly gave rise to changes were incorporated.

Macro cycles C1 and C2 were very much small-scale: one teacher with one group each. The adaptations to the prototype after these two cycles led to the prototype we employed during a macro cycle C3 in six subsequent case studies (with five teachers). These cases were that tightly scheduled, we were only able to make minor changes in between. Each case study in macro cycle C3 could therefore be interpreted as a 'meso cycle'.

The three prototypes of the intervention were evaluated with a couple of instruments. Evaluation of the (student / teacher) questionnaires were used as an input for the (student / teacher) interviews. The results of these interviews served, in turn, as an input for the interpretation of the observations.

The coded observations ('implemented curriculum') were compared to the HTT ('intended curriculum'). By using a correspondence metric we calculated a score for each case study. This score reflected, on a scale from 0 to 7, the degree of correspondence between the intended and implemented intervention. Besides these correspondence scores we sampled typical fragments from the classroom discourse in order to illustrate how the feedback process in each case study was realised.

8.1.3 Main findings

This intervention study was carried out because mathematics teachers experienced a curricular overload. Our main idea was to improve the quality of contact time through an improvement of teacher feedback. We therefore utilised the possibilities of a classroom network. Teachers and students participating in this study agreed on the fact that during the intervention more feedback was established than during the usual educational setting.

When piloting the prototype of the intervention in C1 and especially C2, we found that feedback on ASS was improved. Immediate feedback on the handhelds itself (type III feedback) was considered to be useful by the students and the classroom network could be utilised for delayed feedback (type IV feedback) in order to monitor the students' skills on the GC and to demonstrate these skills. This demonstration could be given by the teacher, but also by a student of whom the teacher had observed (made far easier by the network) that the skill was mastered.

However, in this study we were even more interested in improving feedback on DL. From all of the three macro cycles C1-C3 it appeared that immediate feedback on students' work on DL, delivered by the students' GC (type I feedback), was very limited. The technology was not sophisticated enough to deliver really useful feedback on the GC with respect to DL activities. However, from C3 it appeared that, in order to deliver delayed feedback on DL (type II feedback), the teacher was very well informed by the classroom network. Therefore, after evaluation of C3, we concluded that the classroom network offers good support for the supply of feedback for three out of four feedback types. Further, we concluded that the delayed feedback, both on ASS as on DL activities, can be a good starting point for productive student-centred classroom discourse on statistical topics. Of course, the improvements in feedback and classroom discourse were realised under certain conditions. In the next section we discuss these conditions in detail.

With respect to the methodology, we developed and used a systematic way to code the HTT (intended intervention), to compare it with the implemented feedback (implemented intervention) and to express the correspondence between these curricular representations in a number between 0 (no correspondence) and 7 (perfect correspondence).

When looking back at the approach of educational design research we conclude that it facilitated us very well in answering the research questions.

8.2 Discussion on main findings

8.2.1 Teachers' and students' characteristics

In order to answer our research question, we first stress that through the use of an iterative design and development of the prototypes, the data from the last and third prototype of the intervention implicitly contained important aspects of the results of prior cycles. Therefore, we base our answer to the research question here on the data as gathered during C3, and when we use data specifically from C1 or C2, we will mention that explicitly.

We now summarise the main results of C3, as presented in chapter 7, in two tables. Table 8.2 represents the *teachers'* characteristics while Table 8.3 represents the *students'* characteristics, both with respect to the implemented curriculum. The dimensions depicted are those that emerged from cycles C1-C3 to be the most important ones considering the goal of the intervention: improving feedback by the use of a classroom network.

For the dimensions 'Relation class', 'ICT skills' and 'Skills discourse' we use a four point scale (•: bad, ••: moderate, •••: good, ••••: very good). We map the gathered data on the presented dimensions. These data mostly consist of self-reports, sometimes including our own judgments, based on our observations.

Table 8.2 Teacher characteristics with respect to the implemented intervention

Teacher / **Dimension**	S1	S2a	S3	S4a	S4b	S2b
Relationship class	••	••	•••	••	••	•
ICT-skills	••••	••••	••••	•••	•••	••
Skills discourse	••••	••	••••	••	••	••
Correspondence Mean ASS	5.60	5.71	5.46	3.44	3.33	2.25
Correspondence Mean DL	4.89	3.66	5.34	3.65	4.11	1.95
%-Missing	68	15	0.00	34	36	60

Table 8.3 Students' characteristics with respect to the implemented intervention

Case / **Dimension**	S1	S2a	S3	S4a	S4b	S2b
Learning gains (according to teacher)	?	?	•••	•••	•••	•
Commitment	?	•••	•••	•••	•••	•

(according to teacher)						
More feedback perceived (teacher)	••••	•••	••••	••••	••••	•
More feedback perceived (students)	•••	•••	••••	•••	•••	•••
More time spent on homework	•	•••	•••	•••	•••	•
% CN (vs other teaching activities | teacher-students)	50%-?	50%-?	33%-60%	50%-65%	50%-50%	?-65%

8.2.2 Answering the first subquestion

In the light of our main research question

> *What are the potentials of a classroom network in supporting teachers with providing feedback in statistics education?*

we will now answer the first subquestion of this study:

Was the technological support by means of the CN adequate for the intended feedback in the lessons? (Conditional question)

After case S1, the technology was stable, both on the handheld and on the network side. The percentage missing was between 0 and 36 (when not including S2b). Case S3 proved that it was possible to conduct every intended feedback session in the classroom setting. In cases S4a and S4b, with a percentage missing of 35%, it was mainly time management by the teacher that obstructed utilising the CN more frequently. We nevertheless consider this percentage of 35% as reasonable for classroom practice as roughly 2 out of 3 sessions have then been realised. In case S2b, with a percentage missing of 60%, we were faced with such a low motivated group of students that the teacher felt that establishing more feedback sessions was not useful.

8.2.3 Answering the second subquestion

Has it been possible for a mathematics teacher to implement the prototype in accordance with the intentions? (Existence question)

We consider case S3 to be a convincing implementation of the intended intervention. This had a high mean correspondence score, both with respect to ASS and DL. Students and the teacher were equally enthusiastic about the improvement of feedback. All of the feedback sessions were carried out as intended, demonstrating that the technology served the intervention very well. Besides a convincing case study being a *proof of existence* for the goals of the intervention it is remarkable that in every single case study there were feedback sessions with a convincing correspondence score. This means that every teacher, under the right circumstances, has been able to conduct a feedback session as planned. This is even true for case S2b, with a poor correspondence score (2.10 on average), low learning gains and low student commitment, during which there were feedback sessions with a satisfactory correspondence score (5 and 4.5; see sections 7.11.2

and 7.11.3). The interviewed students in this group voiced a preference for an average use of the CN in order to start feedback sessions for 65% of the teaching time. Apparently even in this case, the essential power of feedback supported by a CN emerged. We consider these as *'micro proofs of existence*': the teacher succeeded in conducting at least one feedback session sufficiently according to the intentions while the students were convinced of the feedback potential. As this relates to case S2b, it suggests that for the other cases the evidence is much stronger.

8.2.4 Answering the third subquestion

Was the feedback support of the CN equal for ASS and DL? (Didactical question)

The support for ASS proved to be better, but, with a highly specified HTT, we managed to support the teachers in giving feedback on DL in a satisfactory way. The slightly better support for ASS is shown by the fact that the mean difference in correspondence score between DL and ASS was 0.37 (in the advantage of ASS). We consider this gap, with respect to a variable on a scale from 0 to 7, to be quite small. The 'built in' support of the CN for developing students' DL has to be completed by specific teaching methods and by more directing teacher preparation. Therefore we used one-to-one instruction before, and a similar evaluation after, each lesson. During C1 we were not able to collect valid data with respect to the difference in the support of ASS and DL, but during C2 we failed with respect to the support of DL (see chapter 6). This brought us to the changes in the teaching methods and teacher preparation.
The big difference in mean correspondence between feedback on DL and ASS in case study S2a during C3 deserves some attention. We noted that the teacher in this case study did not have a strong *functional extraversion*, that is, he is not very focussed on leading the students' learning input during the classroom discourse. This hinders him more considerably in the feedback sessions on DL than on ASS. We presume that this could be because, for a mathematics teacher, discussing 'hard' procedures is easier than 'soft' processes. Discussing DL could be perceived as more vulnerable ('Why nagging when having an answer?') and therefore would take more *functional extraversion* than it takes to discuss ASS. In mathematics education, there is a stronger tradition of focussing on procedures than on concepts and ideas.

8.2.5 Answering the fourth subquestion

Which teacher characteristics promoted/hindered the implementation of the CN as intended? (Identification question)

We concluded in chapter 7 that the data as collected during C3 proved that improving feedback in statistics education by the use of a classroom network was possible. But what was needed to make the step from a successful implementation during two case studies (out of six, like we did in C3) to a successful implementation in further case studies? First, we recall from section 7.12 that we did not reach 'successive approximation' (van den Akker, 1999) of our 'intended use of the intervention': case S3 (chronologically the third case) came closer to the intentions than the fourth, fifth and sixth case, despite the fact that we continuously used our experiences in order to prepare the teachers in a better way. Apparently, there was a teacher influence, a group influence, or an interaction between the teacher and group that was bigger than the influence on the intervention. The little difference in corresponding scores between S4a and S4b (with the same teacher for different groups) suggests that correspondence is more teacher dependent than group dependent. With respect to the teacher, this brings up an interesting question related to

our main research question: what are the strong teacher influences that cause this variation in correspondence score?

Using the results of C3, as presented in Table 8.2 and Table 8.3, we can conclude that there are at least *four conditions* that have to be met before a teacher, trained and supported as we did during C3, in a learning environment that is technically stable, can fully utilise the feedback potential of the classroom network in statistics education.

First of all, we observed during case S2B of C3 as well as in C1 there should be a relationship between teacher and group that is based on *sufficient mutual trust*. If this trust is lacking, all education is to fail, however well-resourced the learning environment potential may be. Good education is an intimate process. Feedback and classroom discussion are perhaps the most vulnerable parts of it. Mutual trust is indispensable for making these succeed.

Secondly, the teacher has to have deep *conversational skills*, including the attitude (or is it even 'personality'?) to apply them as productively as possible in the classroom discourse. This means that she or he has to be a 'conductor' (Drijvers, Doorman, Boon, Reed, & Gravemeijer, 2010) of the classroom discourse, which in this context should be interpreted as 'the spider in the web of the educational process'. In this educational process classroom discourse takes a prominent place and leading it means being able to take up responsibility, especially in a communicative way. A sufficient level of *functional extraversion* is needed in order to be able to take up this responsibility. As a supplier of constructive critical feedback on students' work and when acting as a conductor of classroom discourse, the teacher has a very prominent role in the classroom theatre. This role he not only has to deserve, he has to demand it. We would not go as far as stating that this concerns the immutable level of the teacher's personality. It is about functional behaviour, which can be acquainted, but this is a fairly severe requirement which can make it hard for a considerable part of the population of mathematics teachers to utilise the full potential of a classroom network. Very important was, as we observed during all the case studies, that the teacher was using students' names in order to spread the discourse among the whole group. Using students' names is more confronting, because it is more difficult for students to hide. The teacher therefore has to compensate by creating a safe environment, for example by showing some things of his own, without undermining his position as a leader. He has, as a real conductor, to make his musicians excel in their own way, without losing the collective goal. We especially point at the teacher's timing as a conversational skill. As when conducting musicians playing together, timing is essential for the proper performance of a piece of music, timing is also essential for a teacher in order to optimally implement an intended curriculum (in this case: an HTT). In general the teacher's repertoire on formative assessment (Black, et al., 2003; Black & Wiliam, 2009) and dialogic teaching (Alexander, 2008) has to be at a sufficient level.

Thirdly, besides these conversational skills, the teacher should have competence in *quickly interpreting students' answers* as he has a greater number of these to handle than without the use of a classroom network. He should be able to make ‘statistical sense’ of much more student input than before. Due to teacher feedback being needed in the case of new ASS or DL student activity, this input will very often have the form of open answers. In this case, making a rough 'feedback scheme' based on students' answers, as we observed in C3 during case S3, can be very useful in giving feedback the right direction, simultaneously doing justice to the students' input. This capacity in interpretation of students' input requires sufficient *subject matter (mathematical) knowledge and*

pedagogical content knowledge (PCK). The teacher should have a sufficient level of both knowledge types. This may sound trivial or perhaps even offensive. However, in our view there is no simple mathematics. It takes a deep understanding of concepts, ideas, procedures, strategies and links between different mathematical subdomains to be able to effectively process all of the students' responses on mathematical questions (Even, 1993) in real-time. Having much more information to process, like has been reported in this study, makes this job harder.

Fourthly, the teacher should have *skills with respect to ICT*. Using technology, both on the handheld side as well as on the network side, should only result in a low cognitive load so that the teacher is able to concentrate on giving feedback and directing the classroom discourse towards meaningful interaction with respect to statistics. If the technology requires too much attention to be handled successfully, this may lower the flow in a discussion. This technical condition may seem somewhat trivial, but in general the teacher acquisition of ICT skills is not to be underestimated (Hakkarainen, et al., 2001) and the integration of ICT skills for pedagogical use is especially difficult (Hughes, 2005). In recent research this aspect of teacher skills is more and more stressed (Mishra & Koehler, 2006). It takes a sustainable effort to maintain these skills in order to be able to smoothly switch to new tools or to new versions of familiar tools. PCK is nowadays supplemented by technological pedagogical content knowledge (TPCK) (Koehler & Mishra, 2009).

During C3 teachers as well as students considered that the supply of feedback using a classroom network, as perceived during this intervention, has such a high potential that they advise dedicating on average half of the lesson time to this teaching activity. It is remarkable that even in the case study which was far from a success the interviewed students mentioned percentages between 50 and 100. Students experienced more feedback than during education without a classroom network. Teachers reported improved feedback possibilities, with no difference between feedback on algorithmic statistical skills and on data literacy. Teachers reported that there was more discussion on data literacy than usual and that students in general were more engaged in the classroom discourse which in turn was more focused on mathematics. Teachers are a little careful in reporting positive learning gains. In three cases it is reported, in the other three cases the teacher said they did not know.

8.2.6 Implications of the findings

Teacher investment

We have formulated four conditions to be met in order to potentially profit from these possibilities. Although not insurmountable they are still quite strong. Adley (2006, p. 55) describes the simultaneity of the conditions of his model for professional development of teachers:

> "*A critical feature of this model is that each of its elements is necessary. That is, it is like a chain: if any one element is missing, the whole model collapses. There can be no compensation of weakness in one area by extra strength in another*."

We do not state that the feedback supply of a teacher lacking one (or more) conditions will completely collapse, but there will be a substantial loss of quality in the classroom discourse. Picking out a mathematics teacher from the membership base of, for instance, the Dutch Council of Mathematics Teachers and telling her: "Here is your equipment and that of your students, you are going to start teaching with the emphasis on feedback next

week", is not going to work. A profound, sustainable enhancement of teaching repertoire takes 2-3 years (Adey, 2006). This is supported by the fact that most of the teachers agreed in the questionnaire with the statement that their acting during this project felt like 'having to learn teaching again'. This also shows how deep the intervention effected daily lessons. In order to master this, a serious investment is demanded, which a teacher can only do once every two to three years at most. Nevertheless, we consider the yields of this investment in this case big enough to recommend it is done.

Organisational schedule

All of the teachers participating in this study considered participating in the project to be very hard work. Partly this was caused by the tight schedule. We had to plan all of the six C3 case studies within half a year. Observations had to be done by only one person. This caused a lot of stress on the organisation and strongly limited the flexibility of the teachers and the possibilities to prepare them well enough. From the interviews and questionnaire it seemed that they were content with the support in this project but that they did not feel well enough prepared. This contradiction should be explained by the fact that their actual support, consisting of training and coaching on the job, was perceived as good, but that learning the necessary ways of acting in classroom practice was difficult to the extent that it could not be accomplished without concrete experience in that actual classroom practice. 'It can only be learnt by doing it'.

More data literacy in statistics education

The majority of teachers reported that they were better able to stress data literacy in their lessons. However, we did not just add feedback possibilities to statistical instruction. We also redesigned the content of statistics education, while trying to focus this process more on data literacy. The bigger visibility of data literacy in the realised curriculum was thus not only due to the fact that the teacher's feedback possibilities were improved. We presume that the more prominent position of DL in the prototype put DL on the stage and that the improved feedback (and subsequent classroom discourse) put it in the spotlight.

Immediate feedback

In the feedback matrix on statistics education (see section 5.1.4) we distinguished two types of feedback (immediate versus delayed feedback) and two types of statistics education (algorithmic statistical skills and data literacy).The immediate feedback was designed as feedback supplied by the students' GC. However, we had no possibility to observe the use of immediate feedback (most important as mentioned for declarative and procedural knowledge, consisting mostly of ASS activities in our prototype) other than walking through the classroom, and we have not been able to structurally investigate them. For now, we did not go any further other than to note that teachers and students considered the immediate feedback 'handy', although there were also a few students who mentioned that this type of feedback was a bit superficial.

8.3 Reflection

8.3.1 Reflection on intervention

In order to reflect on the contribution of this study to the solution of the problem of a lack of time in mathematics education, we use the well-known SWOT analysis with respect to the intervention: what are the Strengths and Weaknesses (as we observed during this study), and the Opportunities and Threats (for future advancement)?

1. *Strengths*
   The most important result of the intervention is that it succeeded in improving feedback, as was intended. During C2, there was improvement with respect to feedback on algorithmic statistical skills (ASS). During C3, there was additionally an improvement in feedback with respect to data literacy (DL). A particular strength of the intervention was that these improvements were perceived both by the participating teachers as well as by their students, although not all the case studies were convincing. Four out of five teachers during C3 estimated how much of the teaching time should be dedicated to feedback related activities, as deployed during the intervention. On average this was about 45%. In four of the six case studies we asked the same question to their students. This led to an average of 60%. So participants of C3 suggested spending roughly half of the teaching time on feedback related activities, which we consider to be a substantial success.
2. *Weaknesses*
   The most important weakness of the intervention was its complexity (and, therefore, its cost to develop): participating teachers had to manage both a new pedagogy towards statistics (emphasis on feedback that initiates a student centred classroom discourse), with a shift in learning goals (more emphasis on DL) facilitated by a new tool (classroom network), that actually consisted of a number of new tools. During C3, both teachers and students had to get used to a new handheld device too. This led to one teacher mentioning that he felt like having to learn teaching again. The students participating during C3 usually succeeded quite quickly in mastering a new handheld device, but, as mentioned during the interview, it nevertheless took some effort. In each group there were students who handed in the handheld after the intervention reluctantly. Another weakness of the intervention was the limited intelligence of the immediate feedback on the handheld.
3. *Opportunities*
   In our opinion there is an important battle to be won regarding the intelligence of the immediate feedback as generated by the handhelds. When some kind of intelligent tutoring system could be implemented in order to generate immediate feedback, at least the quality of the feedback on ASS could strongly be improved. This would further improve the value of the intervention. When this could be combined with an expansion of the intervention from one chapter to a greater number of chapters and this intervention could be conducted in about a dozen case studies, we think this would resemble a very strong intervention. Beyond this, we did not find any specific 'learning domain dependent' aspects with respect to the intervention. It has been deployed in statistics education, but could be developed in a similar way for other subdomains of mathematics, for science education, economics education or other education aiming at developing both students' procedural skills and their conceptual skills.
4. *Threats*
   There are two major threats for the intervention. The first is the complexity as mentioned during weaknesses. It is possible that the short term usability of the intervention will appear to be too big an obstruction for the average mathematics teacher to adopt this way of working. The second threat is of a technological nature. It is not quite clear at this moment what the future is of dedicated devices like graphing calculators. The iPad and similar devices are rapidly gaining ground when it comes to one-to-one computing. Interactive whiteboards combined with

voting systems could also threaten the position of our specific classroom network. However, mathematics education does not primarily ask for general functionality, such as a word processor and an internet browser, but for support in solving mathematics problems and in thinking mathematically. So far, with respect to this, the graphing calculator is the only one-to-one tool that has succeeded. Nevertheless, although technological changes could mean a threat with respect to the implementation of the intervention, the design principles as described and evaluated will remain valid.

8.3.2 Reflection on scientific relevance

What did this study contribute to the body of scientific knowledge? We make the distinction here between the knowledge with respect to the function and characteristics of the intervention and the knowledge with respect to the process of design and development of the intervention.

Scientific relevance with respect to intervention characteristics

From a theoretical point of view, this study contributes to the theoretical underpinning of an effective intervention supporting teachers with respect to feedback in statistics education. The most concrete result with respect to this is the formulation of four conditions to be met in order to be able to implement an effective intervention aimed at better feedback in statistics education. Besides that, the combination of different knowledge types in statistics (ASS and DL) with two different feedback types (immediate and delayed) is theoretically interesting. The most important methodological contribution is the development of a metric that maps the relation between intended and implemented curricula to a correspondence score.

Scientific relevance with respect to the process of design and development

On the theoretical side an important opportunity is the scaling up of the intervention, both with respect to the number of case studies involved, as well as with respect to the size of the content. We are confident that this study hands the design principles, procedures, and attributes needed for both types of up-scaling.

On the theoretical side, there is a possible threat that up-scaling the intervention with respect to the number of case studies and/or the size of the content could result in data that are in some way contradictory with the criteria we have formulated for successful implementation. It could, for instance, be possible that an important success factor for the intervention is the age of the teachers, as an intermediary variable for their flexibility to a new didactical approach. In our C3 sample the youngest two teachers (28 and 40 years old) produce better correspondence rates than the other three (54, 55 and 55 years old). We nevertheless consider substantially different findings to ours to be unlikely and expect that up-scaling might possibly lead to one or two extra criteria for success and/or to a refinement of the ones we formulated. With respect to the correspondence metric, it is possible that up scaling would lead to an adjustment of the procedure. However, we consider that to be an advancement in science and therefore an opportunity rather than a threat.

Another threat could be the feeling of insecurity as reported by several teachers in different ways. The duration of the intervention was too short to determine whether this feeling would diminish when teaching longer in the developed environment. We recommend that this is investigated in future research.

8.3.3 Reflection on research methodology

Educational design research

During this research we encountered problems of multiple kinds: technological, organisational, didactical and even substantial with respect to statistics itself. In hindsight, we realise that these problems more often occur when conducting an educational research design (EDR) study. However, the research question we posed could only be answered when deploying EDR. The phased and iterative approach that characterises EDR and the specific research design we chose (ADDIE: Analysis, Design & Development, Implementation, Evaluation) was actually perfect in order to develop our intervention, adjust it iteratively, during both micro and macro cycles, and evaluate it. Therefore, we do not have any regrets when it comes to the choice of our research paradigm. In the sections below we will reflect in more detail on the specifics of our methodology.

Even with the problems we met and the stress we posed on the participating teachers and ourselves, we believe we made the correct choice in selecting the paradigm of educational design research. Seeing things work in practice, and even seeing them not work, is essential for a design researcher. The eclectic approach, for instance transcending the classical opposition between rationalistic and naturalistic, suited very well the approach to our initial problem: how to create more efficiency in mathematics education?

Further, our intervention proved to be far from teacher-proof, regarding the strongly varying correspondence scores. An important yield of this study is the formulation of the conditions to be met in order to create a higher level of 'teacher-proofness'. Absolute teacher-proofness is, of course, an unattainable ideal and possibly an undesirable ideal. EDR is, anyhow, the only way of leading an intervention that comes close to teacher-proofness.

Planning of activities in technology rich intervention studies

During the empirical stages of the whole study (initial study, C1, C2 and C3) we continuously had the feeling that we were always a step behind with respect to technology. An extreme example of this was the preparation of C3, case study S1, which started in February 2010. We used the first Nspire handhelds in Europe, which were presented three days before at a conference in Stockholm, Sweden. Of course we knew this was taking a considerable risk, but in hindsight we still think it was the right decision. The alternative was using the technology as we had used during C1 and C2 with which, as we reported in chapter 6, we had technological problems. We had immediately after C1 and C2 reported these problems to Texas Instruments, the supplier of the technology. Based on our recommendations TI developed a new generation of the technology. We felt and still feel that when conducting EDR the research team has to use the results of prior iterations. Here, this meant that we had to commit ourselves to the latest technology. Having six case studies, we decided to start with the first immediately while we had not tackled all of the technological problems, as these problems always occur, sooner or later, when working with technology. We decided that it was most wise to encounter the problems as soon as possible in order to be able to solve them as soon as possible. Some of these problems can be generated in a simulation, but others occur in real life hence we chose this approach instead of trying to solve the problems outside the classroom. Although it was risky to use the latest technology, we now feel that this was not the whole story behind our hurry and continuous lack of time. We concluded this after C2 and we had to conclude it again after C3, that apparently we were too optimistic in making estimations that are important in planning a design research study and it was continuously

difficult to plan more realistically. This also had to do with the fact that a lot of organisational affairs in the schools were beyond our influence. This may be a lesson for all design research drawing as heavily on new technology as we did: plan those things that can be planned in a way that may appear exaggerated. The ultimate realisation of this study would not have been possible without help from colleagues, who were at ungodly hours flexible and available for all kinds of support. For us, there is a firm lesson to be learned to schedule all activities more realistically within a reasonable margin in time. This type of research is not something that can be done on the edge of the night.

Designer / researcher role

While conducting educational design PhD research one has to be both a designer-developer as well as a researcher. It is tough but not impossible to incorporate both competences in one person. “Zwei Seelen wohnen, ach! in meiner Brust.” (Faust. Eine Tragödie. Johann Wolfgang von Goethe, 1808). The biggest problem is to timely switch between the two of them. In particular, we have found it difficult to release the designer role. While there was always a terrible hurry, as mentioned in the previous section, we still had a tendency to keep on tinkering, even on the tiniest details of the prototype. For instance, overnight we refined a certain section of the teaching materials. The following day the teacher had to resend this to the students, because he had already sent the previous version the day before. Of course, there were always one or two students absent. Those students were then happily working in the old version of this section. When the new file was collected by the teacher, the files of the formerly absent students were not included by the network, because their file had an older version extension in its name, and were thus not analysed. Things like this disturbed the teaching process, something that was difficult enough anyway. Fortunately, things like this remained accidents because we tried to stick to the ‘check, check, double check’ rule and the A.D. de Groot advice for research design “If you think you're ready, think it all over again".

Another example of an intertwining interest between designer and researcher emerged behind the video camera when we saw the teacher doing something in a way other than explicitly addressed in the preliminary discussion of the lesson. A designer then wants to intervene and yell: “Stop! We're gonna do this otherwise!” but the researcher is supposed to simply record what happens. Each time we had an impulse like this, we managed to remain silent, but we should mention that this was not always easy. It was nevertheless possible not to interfere with the lessons, because of sticking to a standard procedure: make a note and discuss the events directly after the lesson with the teacher in order to diminish chances of repetition.

A third mixed, not to mention conflicting, feeling every design researcher will recognise is the disappointment when live observations show no steady progression towards the intended ideal intervention. The design researcher then tends to intervene. However,, as a colleague reacted: “The disappointment of the designer is a chance for thinking for the researcher.” This actual thinking on the how and why behind the gap between a priori agreement and de facto acting has indeed been a valuable source for us in making failed expectation productive. And afterwards it is, of course, logical that in as complex an environment as the one we designed and researched the uncontrollable variability is bigger than the one we were trying to manipulate by design. Kelly (2007) could say: “If that's not the case, you wouldn't need educational design research to investigate it”.

In conclusion, we state that reflection is an integral part of educational design research. Very roughly this requires: the researcher to reflect on what the designer-developer has implemented. One could consider reflection as some kind of meta-evaluation. Therefore

we constantly posed ourselves questions like: what happens now, is this consistent with our theoretical framework, is this according to our plan, what does a possible aberration from the plan mean, does this influence our research question, should our plan perhaps be adapted, do the results of the evaluations match our expectations, if they do not, what could be going on? In this study, design-development and research activities have been carried out by one and the same person. We could imagine that for complex trajectories like the one described in this study, let alone for even more complex educational design research projects, a distribution of responsibilities among different members of the team would relief Faust and could improve results.

8.3.4 Limitations of this study

Limited sample size

The limited sample size is a weakness with respect to the implementation criteria. The initial study, C1, C2 and C3 in total were just nine case studies. It is possible we overlooked important criteria simply because we were not faced with them during these cases.

Serial interpretation

A weakness with respect to the correspondence metric is the implementation: while transforming raw observational data to a correspondence metric a couple of linked interpretations are used. We tried to build in as much objectivity as possible (counting events) and document the interpretations as precisely as pragmatically possible. However, because the interpretations are linked, and therefore a bias in the beginning will work through the entire process, the procedure is perhaps still too 'interpreter sensitive'.

Stage of this research

With the outcomes of this research, we certainly do not believe to have spoken the final word with respect to the research question. So, where are we now?

During the empirical stages of this study we evaluated nine case studies: one in the initial study, one in C1, one in C2 and six in C3. Just in C3 we encountered one case in which the results came particularly close to the intentions and three in which the results came nearby the intentions. In one case study there was far too much technical failure, but when the technology was stable, the implemented curriculum came close to the intentions. In one case study the implementation was far from the intentions. In general, we could say that this very explorative study offers us the possibility to produce an intervention that will be stable, that is, of which the implementation in general will come close to the intention, when our recommendations, as formulated in the success criteria, are met. We would like to see a study that expands the prototype, for instance by merging our materials about the normal distribution (from the initial study) with the materials about descriptive statistics (from C1-C3), and combines this longer intervention with teacher training according to our recommendations. This teacher training should result in better feedback skills, both with respect to ASS as well to DL activities. This could diminish the feeling of insecurity that was reported by the teachers who participated in this research. A study designed like this would fine tune the intervention and the conditions that are to be met before the intervention can succeed.

The influence of ICT on procedural skills in mathematics

Although we consider ASS to be as important as DL, we focused in this study on the development of DL, because we think that DL has not received the attention it deserves. Nevertheless, we have a problem, which we will illustrate with an example. As we reported in section 7.5 we adapted a student exercise about the calculation of the standard deviation. In fact, we tailored it more to the possibilities of the GC. Before this adaptation we offered an authentic data set in a spreadsheet. The authenticity in this case among others meant 387 entries, making the use of ICT almost indispensable. In the old exercise text we explained what the standard deviation was, and which steps have to be set in order to calculate this measure of spread (see figure 8.1). In the student manual we described how the spreadsheet on the GC worked.

Op de volgende pagina 2.2 zie je de lengte van 387 12–jarige Nederlandse jongens in september 2000.
Bereken in cel B1: het gemiddelde van die lengtes.
Bereken in kolom C: $lengtes\ van\ jongens\ van\ 12\,jaar - gemiddelde$
Bereken in kolom D: $(lengtes\ van\ jongens\ van\ 12\,jaar - gemiddelde)^2$
Bereken in cel E1: de som van alle $(lengtes\ van\ jongens\ van\ 12\,jaar - gemiddelde)^2$ (=D1+D2+…+D386+D387)
Bereken in cel F1: $\frac{de\ som\ van\ alle\ (lengtes\ van\ jongens\ van\ 12\,jaar - gemiddelde)^2}{387}$
Bereken in cel G1: $\sqrt{\frac{de\ som\ van\ alle\ (lengtes\ van\ jongens\ van\ 12\,jaar - gemiddelde)^2}{387}}$
De uitkomst in cel G1 noem je de standaardafwijking (ook wel: standaarddeviatie) van de lengtes van 12–jarige jongens.
Hoe groot is die standaardafwijking? Rond je antwoord af op twee decimalen.

Figure 8.1 The procedural steps to be set in order to calculate the standard deviation

In our observations during C3 cases S1 and S2a we noticed some students struggling with this procedure. During S2a the teacher needed about 17 minutes of feedback just on this exercise, after which we still doubted whether the procedure was clear to most of the students. After this lesson we discussed it with the teacher. He confirmed our suspicion: this was too difficult for this target group. Therefore, we adapted the exercise, using the pre-installed procedure for calculating the SD. Ultimately, we thus black boxed the procedure. A purist would perhaps notice that using a spreadsheet was the first step on this way to black boxing. As mentioned, research points out that ASS is needed in order to develop DL (Rittle-Johnson, Siegler, & Alibali, 2001). The actual performance of the concrete steps stimulates the development of students' number sense and the intuition for mathematical procedures. And by what is it replaced?

Sparrow, Liu and Wegner (2011) studied a similar phenomenon. They compared the working of the memory when it comes to remembering plain facts (declarative knowledge) of persons who used the World Wide Web as an information source and of subjects who first consulted their own internal memory. They report a shift in the memory of those using the web from the facts themselves towards the place to find them, and called this the 'Google' effect on memory. One could ask the same for users of global positioning systems (GPS): What is the influence on their 'card and compass skills'? And are more abstract skills, like spatial awareness, influenced? A possible influence like that could be called the 'TomTom effect'. Both the Google and the TomTom effect involve mainly declarative knowledge which is in mathematics not considered as the most crucial type. In essence, we are questioning how in mathematics, procedural skills and conceptual skills are interrelated. In recent research there seems to be a revaluation of procedural knowledge as a source for conceptual knowledge (Baroody, Feil, & Johnson, 2007; Star, 2005). To which level can the one be substituted by the other? There seem to be individual differences among students (Gilmore & Papadatou-Pastou, 2009). Is there a

dependency on the learning goals? This was not part of our study, but it dawned on us as a very important and not completely solved issue for mathematics education (Peled & Zaslavsky, 2008).

Intermediary effects

When using a student questionnaire and doing student interviews it became clear that it was very hard for the students to concentrate on their perception of just the network possibilities. This is logical, of course, because the students first received a completely new GC to work with. Although we developed a manual that was to guide them from the usual GC to this new handheld device, and although we focused on a restricted part of the functionality and the new device was more user friendly than the one they were used to, it still was a completely new device with complete new possibilities. Some of the students reported that it took them a couple of lessons to understand their device. We cannot exclude the influence of this on the way the students completed the questionnaire and on their opinions reported during the interviews. A subsequent study must, if possible, try to avoid this effect.

Reliability and validity

Which measures did we take to meet these criteria?

With respect to *internal validity,* indicated as *credibility* by Guba (1981), we constantly monitored the validity of the prototype of the intervention by discussing issues with colleagues. Before conducting C1 and C3 we explicitly asked an external colleague to review the prototype with respect to the research question and the learning goals. We discussed their comments and adjusted the prototype with respect to the findings of this review process. Participating teachers, on the other hand, provided feedback on the prototype based on their prior and actual experiences. While working out their ideas we improved the actual practicality of the intervention. We concluded the teacher interview session with a group interview, looking back collectively at the experiences, in order to recognise those of each other and to validate them. We used data triangulation during three empirical stages of prototype piloting in order to optimise the chances that the intervention corresponded with the original problem. All these measures seem to have contributed to an internally valid study.

With respect to *external validity*, designated as *transferability* by Guba (1981), in qualitative research like this study, this is mostly a concern of the generalisability of the research findings. We have secured this by precisely describing the dependency of our conclusions with respect to the research context (Barab, et al., 2008). Attributes of teachers, students, disciplinary content, and technology were taken into account when interpreting our data. These can be reused when extrapolating our findings to other educational contexts (Barab & Kirshner, 2001). Further, by conducting this research in 'everyday classroom practice' we made its results more transferable to other research settings.
We claim, for example, that the main findings of this study are applicable to domains of learning other than just mathematics. In none of the used procedures is there an intrinsic domain-bounded step. Of course, the distinction between data literacy and algorithmic statistical skills is a typical statistical phenomenon. However, we used this distinction as an instance of the more abstract distinction between conceptual and procedural/declarative knowledge. This abstract distinction could be used in order to develop a comparable feedback matrix for another learning domain. The restriction to be made here is that the distinction of conceptual versus declarative-procedural has to be

meaningful in this learning domain. We cannot oversee other learning domains, but those rather similar to mathematics, for instance physics, should be able to apply the same methodology in order to get comparable results. At the same time as our initial study, we supervised a couple of trials of this way of working in English, economics and geography classes by teacher students. Their reactions were positive and in line with our observation that teacher feedback was to be a very promising characteristic when working this way.

What have we done in order to ensure *internal reliability*, interpreted as *dependability* by Guba (1981)? In our view, internal reliability in qualitative research is mainly a matter of consistency. We discussed the coding scheme, of which we used a more primitive version during C1 and C2, until we were satisfied with it. We used it to code parts of protocols of earlier recorded lessons simultaneously and independently until we agreed on the interpretation of the scheme. However, coding terabytes of video data is arduous and ensuring consistency was not always easy. For example, when starting coding we had to go back a few times to the start in order to make the coding consistent with our latest insights. We noticed that after a couple of revisions, we lost the sense of urgency to adapt the coding system as it converged to a usable system for coding the events we saw happen on the observation videos. After all of the lessons were coded, we developed a metric to measure the correspondence between HTT and the realised curriculum. It was especially hard to map different types of feedback sessions on the 7 point scale of the standard feedback session code. After this, we used the correspondence metric to convert the coded implemented feedback to a number between 0 and 7. During these processes, we did our best, but in the trajectory from classroom discourse to a correspondence number, there are multiple interpretations to be made. In order to make these interpretations more objective we based them as much as possible on counting procedures of well described teacher or student actions. Expressing a correspondence score like this into a number with two decimal places is actually hazardous. In a strictly methodological sense, one may even not calculate a mean with data on an ordinal scale (Stevens, 1951), although the counting part of the correspondence metric is of course at a ratio scale. As long as we interpret the outcomes of the calculations (mean and standard deviation) with caution, we consider the outcomes of these calculations to be meaningful. Afterwards, one could perhaps be happy that the differences in correspondence scores were as large as they were. This made the interpretation and the combining with the results of other research instruments easier. During the interviews we discussed standard items. During the subsequent discussion we played the role of devil's advocate, trying to tackle student or teacher reasoning that could be supportive for a positive evaluation of the intervention. We always stressed at the beginning of each interview and at the start of each questionnaire that participants could only support the investigation by being completely honest. In this way, we tried to reduce *response bias* (Paulhus, 1991).

External reliability, expressed as *confirmability* by Guba (1981), expresses the concern that it should be possible for the reported research to be reconstructed ('confirmed') by other researchers. We ensured this by being very descriptive on process and procedures as well as on products. An important contribution to a "*thick description*" (Ryle, 1971) is the fragments of classroom discourse from all of the case studies during C3, selected by underpinned criteria. With these fragments, it is possible to see 'the depth' of the intervention. The representation by correspondence scores gives an overview of the intervention. Both sides contribute to a complete image of what has happened. A detailed and explicit hypothetical teaching trajectory (HTT) should give other researchers the tools needed to replicate or extend this study. In addition we sometimes used an illustrative style of writing, explicitly trying not to lose preciseness, in order to appeal as much as

possible to the imagination of other researchers so they could 'virtually replicate' (Smaling, 1990) the presented study.

8.4 Recommendations

8.4.1 Practice of mathematics education

Professional development; organisation at macro level

During the empirical stages of the initial study, C1, C2 and C3 we have worked very intensively with teachers. Besides the macro preparation and training, we also provided micro support. Every evening before each lesson we called the teacher and discussed the preparation of the next lesson with respect to the HTT. We discussed each lesson more or less directly after conduction. Working this way, we were able to see the extent to which this project was hard work for the teachers, who, of course, also had their other professional obligations. However, irrespective of the hard work, they enjoyed participating in the project. One of the teachers mentioned: “You have to do things like this in order to continuously develop your teaching.” We would very much like to see mathematics education organised in such a way that experiences comparable to those the teachers had during this project would come within reach of every teacher. In short, teachers should be facilitated to become key participants in research projects. This requires an effort from school administrations and probably even from national policy makers.

Professional development; organisation at the micro level

We see mathematics as the science of patterns and structures. This study taught us, among other issues, that structuring mathematics education along a chosen principle (in our study: feedback in order to evoke a meaningful classroom discourse) is very important too: “It's the structure, stupid” (van der Hilst, 2010). We encourage teachers and other practitioners to really rethink the design of the mathematics education they are responsible for. Designing means almost inevitable learning and understanding (Wiggins & McTyghe, 2005). Backwards designing a concrete curriculum from the learning goal (in this study: development of data literacy) along the structuring factor to teaching activities could be a cleansing experience.

8.4.2 Design of classroom networks

Better feedback function on the handheld side

On the handheld side of the classroom network immediate feedback is generated on the work of the student. Unfortunately, this feedback is still very rudimentary:

- knowledge of *results*: just informs the students whether they provided the right answers or not;
- knowledge of *correct response*: just informs the students about what the correct response to the assignment was.

It is remarkable that with the classroom network we used on multiple choice answers there was no specific differentiated feedback with respect to the specific answers. It kept 'Right/wrong' feedback and the possibility to ask the correct answer. Students and teachers were actually somewhat disappointed about this. They had apparently expected more sophisticated feedback.

A big step further in the same direction is the integration of some kind of intelligent tutoring system (ITS; see section 2.2.1) (Sarrafzadeh, et al., 2008; Sleeman & Brown, 1982) on handhelds. Through the World Wide Web, intelligent feedback on students doing mathematics is already deliverable to systems running a web browser. This development has not made the step to the graphing calculator devices yet, but with the emergence of very portable tablet computers like the iPad, the Samsung Galaxy pad or the BlackBerry Playbook, all connected to the web using WiFi or a 3G connection, it will be just a matter of time (and some internet interactivity supplying middleware like Java, Flash or HTML5) that ITS will be within reach for devices like this. Machines like the TI Nspire, which we used during C3, are not yet connected to the web, but through the cradle they can utilise WiFi connections. If the WiFi access point in the classroom can be connected with the internet, a browser and interactivity support are the only things needed in order to update the GC with the possibilities of online ITS. This would enhance the richness of the immediate feedback enormously.

The support of feedback the way this study has investigated could possibly be realised by other ICT classroom settings, such as with interactive whiteboards combined with voting systems (Higgins, Beauchamp, & Miller, 2007). Future research could investigate this transferability.

Tracking function of feedback

With the current state of development of classroom networks it is not possible to track the interaction a student has with the mathematical content (exercises) on his GC. This is a loss of a valuable source of information with which the teacher could further improve the feedback in the classroom discourse. To generate on the handheld side some kind of dynamic representation, for instance a screen movie that shows all student activity, combined with software on the network side that is able to condense this data flow into an understandable representation for the teacher, would be an innovative step facilitating further fine tuning of the teacher feedback.

8.4.3 Further research

Continuation of this study

We recommend an expansion of this intervention study using the materials and research design we developed. This would preferably be for a longer period, for instance three chapters, if possible in schools where the Nspire is already the standard GC. With an expansion like that the 'start bias' that comes with working in a new learning environment could be reduced to a minimum.

From a research perspective, we especially hope for a further outlining of the coding scheme and the correspondence metric, which we see as the main methodological contributions of this study.

Conducting this research among substantially more than six teachers, for instance between ten and twenty, should be enough to test the ‘success conditions’ we formulated and to collect data for the fine tuning of the prototype. When the intervention is adapted according to the findings of a study like this, it would then be appropriate to conduct a quantitative effect study. In the training of the participating teachers, the videos collected and analysed in this study could be very useful, as we can conclude from the high appreciation of the videos shared with the teachers in our study using YouTube. Specific behaviour and critical events in the classroom discourse can be highlighted by using our materials.

It would be very interesting after the first intervention to repeat it three years later, meanwhile letting the teachers optimise their teaching according to the principles we formulated. This should show significantly better results (Adey, 2006).

Collaborative learning

A substantial number of the interviewed students mention that they experienced better learning through the input of their peers, as made accessible through the classroom network. Although we were aware that peer feedback is nowadays considered as a valuable learning process (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010), we did not design our prototype on this specific possibility. The fact that it is nevertheless mentioned spontaneously by quite some students indicates that in future design research studies this deserves structural attention. The same counts for peer assessment (Vickerman, 2009), a specification of peer feedback.

ICT mediated relationship between ASS and DL

As we stated in section 8.3.4 in the subsection 'The influence of ICT on procedural skills in mathematics' our study did not problematise the influence of a change in procedural skills (ASS in this study) on conceptual skills (DL in this study) as is influenced by the use of ICT. This can be seen in the framework of what in recent years has been called '21st century skills' (Silva, 2009; Trilling & Fadel, 2009). For mathematics education, traditionally quite strongly involved with the possibilities of ICT, this is a major domain of research. Specifically in our view, research is needed to investigate how the mediation of ICT can contribute to the mutual reinforcement of DL and ASS.

Terminology and frequently used acronyms

Acronym	Stands for	Description
ADDIE	Analysis, Design, Development, Implementation, Evaluation	Well known phasing of scientific engineering-like activities.
ASS	Algorithmic statistical skills	Procedural knowledge needed in order to solve statistical problems.
CA	ClassAnalysis	Tool of the CN for evaluation and representation of the students' work.
CAF	ClassAnalysis feedback	Teacher feedback based on information generated by ClassAnalysis
CC	Custom choices	An exercise type that lets students choose between an adjustable number of alternative answers.
CH	Character	A parameter referring to the character of e specific feedback session.
CN	Classroom network	The configuration of hardware and software in order to establish a wireless connection between students' handhelds and the teacher computer.
DL	Data literacy	Reasoning and sense making with and about data.
EDR	Educational design research	A series of approaches, with the intent of producing new theories, artefacts, and practices that account for and potentially impact learning and teaching in naturalistic settings (Barab & Squire, 2004).
FITB	Fill in the blank	An exercise type in which the students have to fill in a numerical answer.
GC	Graphing calculator	A handheld calculator capable of plotting graphs, solving simultaneous equations, and performing numerous other tasks with variables.
GSS	Get students' screens	Tool of the classroom network used to get the displays of the students graphing calculators.
HLT	Hypothetical Learning Trajectory	A coherent set of the goal for the students' learning, the mathematical tasks that will be used to promote student learning, and hypotheses about the process of the students' learning.
HTT	Hypothetical Teaching Trajectory	A coherent set of the goal for the students' learning, the mathematical tasks that will be used to promote student learning, and hypotheses about the process of teaching.
ICT	Information and communication technology	In this study we consider ICT as a tool in order to facilitate the process of teaching and learning.
ITS	Intelligent tutoring system	A computer system that utilises interaction with the user for the representation of its output.
IQR	Interquartile range	A measure of statistical dispersion (spread), being equal to the difference between the upper and lower quartiles. $IQR = Q_3 - Q_1$
IWB	Interactive whiteboard	A large interactive display that connects to a computer and projector and projects the computer's desktop onto the board's surface where users control the computer using a pen, finger, stylus, or other device.
JAC	Just answer checking	A marginal use of the classroom network in order to see if a clear majority of the students did not perceive difficulties with a specific exercise.
LP	Live presenter	Tool of the classroom network used to project the display of the graphing calculator of a specific contributor.
LPF	Live presenter feedback	Teacher feedback that utilises the tool Live presenter to monitor student progress (usually with respect to some specific algorithmic skill).
MC	Multiple choice	An exercise type that is to be completed by making a choice among a set of given alternative answers.
NCN	No classroom network used	The teacher did not use the classroom network for a

Acronym	Stands for	Description
		specific element of the implemented feedback session.
ND	No data	The research instruments did not generate any valid data.
NF	No feedback	Classroom situation in which the teacher did not provide feedback.
OR	Open response	An exercise type that calls for a solution in the respondent's own words.
RAM	Random access memory	Type of internal computer memory of which each physical place has the same access time.
RME	Realistic mathematics education	RME “takes students’ initial understanding as a starting point, provides them with problem situations which they can imagine, scaffolds the learning process via models, and evokes reflection by offering the students opportunities to share their experiences". (van den Heuvel-Panhuizen, 2005)
QP	QuickPoll	A classroom network tool to pose an improvised question to the students.
SD	Standard deviation	The sample standard deviation is the square root of the sample variance of a set of *n* values: $s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \bar{x} = \sum_{i=1}^{n} x_i$
SE	Statistics education	The practice of the teaching and learning of statistics supported by the scholarly discipline that researches this practice.
SII	Students interaction input	Indicates the number of reacting students, the number of student reactions, and the character of the reaction (focussed on data literacy or algorithmic statistical skills).
SLG	Statistical learning goal	The learning goal a specific statistics content unit has.
SV	SmartView	A tool that emulates a graphing calculator on a computer.
SVF	SmartView feedback	Teacher feedback based on a demonstration with SmartView
TC	Teacher conclusion	The conclusion with which the teacher ends a feedback session.
TF	True - false	A question type that is to be answered by making a choice among two alternative answers: 'True' and 'False'.
TI	Texas Instruments	Supplier of hard- and software used in this study.

References

Abrahamson, A. L. (1999). Teaching with classroom communication system - What it involves and why it works. Retrieved December 29, 2011, from http://www.bedu.com/Publications/PueblaFinal2.html

Adey, P. (2006). A model for the professional development of teachers of thinking. *Thinking Skills and Creativity, 1*(1), 49-56. doi: 10.1016/j.tsc.2005.07.002

Ainley, J., Nardi, E., & Pratt, D. (1998). *Graphing as a computer mediated tool.* Paper presented at the 22nd Annual Conference of the International Group for the Psychology of Mathematics, Stellenbosch, South Africa.

Alessi, S. M., & Trollip, S. R. (2001). *Multimedia for learning: methods and development*. Boston, MA: Allyn and Bacon.

Alexander, R. (2008). *Towards dialogic thinking: rethinking classroom talk* (4 ed.). York: Dialogos.

Alexander, S., Sarrafzadeh, A., Masoodian, M., Jones, S., & Rogers, B. (2004). Interfaces that adapt like humans. *Computer Human Interaction* (Vol. 3101/2004, pp. 641-645). Berlin / Heidelberg: Springer

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13*(4), 467-505.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*(2), 167-207.

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology, 62*, 148-156.

Anohina, A. (2007). Advances in intelligent tutoring systems: problem-solving modes and model of hints. *International Journal of Computers, Communications & Control, 2*(1), 48 -55.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as individual resource - Personal strategies of creating information. *Organizational Behavior and Human Performance, 32*(3), 370-398.

Atkins, D.E., Bennett, J., Brown, J.S, Chopra, A., Dede, C., Fishman, B., Gomez, L., Honey, M.Kafai, Y., Luftglass, M., Pea, R., Pellegrino, J. W., Rose, D., Thille, C., & Williams, B. (2010). *Transforming American education; learning powered by technology*. Alexandria, VA: U.S. Department of Education, Office of Educational Technology.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*(2), 111-127.

Bakker, A. (2004). *Design research in statistics education: on symbolizing and computer tools*. Utrecht: CD Bètapress.

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

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*(3), 333-358.

Bangert-Drowns, R. L., Kulik, C. L., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.

Barab, S. A., Baek, E.-O., Schatz, S., Scheckler, R., & Moore, J. (2008). Illuminating the braids of change in a web-supported community; A design experiment by another name. In A. E. Kelly, R. A. Lesh & J. Y. Baek (Eds.), *Handbook of design research methods in education* (pp. 320-352). New York, NY: Routledge.

Barab, S. A., & Kirshner, D. (2001). Rethinking methodology in the learning sciences. *Journal of the Learning Sciences, 10*(1&2), 5-15.

Barab, S. A., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*(1), 1-14.

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education, 38*(2), 115-131.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking & Learning, 2*(1), 127-155.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning and thinking: goals, definitions and challenges *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-16). Berlin: Springer.

Berch, D. B., & Mazzocco, M. M. M. (Eds.). (2007). *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities*. Baltimore, MD: Paul H. Brookes Publishing Co.

Berger, M. (1998). Graphic calculators: an interpretative framework. *For the Learning of Mathematics, 18*(2), 13-20.

Berkvens, J. B. Y. (2009). *Developing effective professional learning in Cambodia*. Enschede: University of Twente.

Biehler, R. (1991). Computers in probability education. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: probability in education* (pp. 169-211). Amsterdam: Kluwer Academic.

Biehler, R. (1997). Software for learning and for doing statistics. *International Statistical Review / Revue Internationale de Statistique, 65*(2), 167-189.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Maidenhead, Berkshire: Open University Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practices, 5*(1), 7-74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

Bliwise, N. G. (2005). Web-based tutorials for teaching introductory statistics. *Journal of Educational Computing Research, 33*(3), 309.

Bloch, J. (2002). Student/teacher interaction via email: the social context of Internet discourse. *Journal of Second Language Writing, 11*(2), 117-134.

Block, J. H. (1972). Student learning and the setting of mastery performance standards. *Educational Horizons, 50(4)*, 183-191.

Block, J. H., & Burns, R. B. (1976). Mastery learning. *Review of Research in Education, 4*, 3-49.

Bloom, B. S. (1968). Learning for mastery. *UCLA Evaluation Comment, 1*(2), 1-12.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (Eds.). (1971). *Handbook on the Formative and Summative Evaluation of Student Learning* New York, NY: McGraw-Hill.

Bogdan, R. C., & Biklen, S. K. (1992). *Qualitative research for education: An introduction to theory and methods.* (2nd ed.). Boston, MA: Allyn and Bacon, Inc.

Bork, A. (1980). Preparing student-computer dialogs: Advice to teachers. In R. Taylor (Ed.), *The computer in the school: tutor, tool, and tutee* (pp. 15-52). New York, NY: Teachers College Press.

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical reasoning. *The American Statistician, 50*(1), 69-78.

Bronfenbrenner, U. (1976). The experimental ecology of education. *Educational Researcher, 5*(9), 5-15.

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

Bunderson, V. (1981). Courseware. In H. F. O'Neil (Ed.), *Computer-assisted instruction: A state of the art assessment*. New York, NY: Academic Press.

Bunt, L. N. H. (1956). *Statistiek voor het voorbereidend hoger en middelbaar onderwijs*. Groningen: Wolters.

Burns, H. L., & Capps, C. G. (1988). Foundations of Intelligent Tutoring Systems. In J. J. R. Martha C. Polson, Elliot Soloway (Ed.), *Foundations of intelligent tutoring systems: An introduction* (pp. 1-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Burrill, G. (1997). Graphing calculators and their potential for teaching and learning statistics. In J. B. Garfield & G. Burill (Eds.), *Research on the role of technology in teaching and learning statistics*. Voorburg: International Statistical Institute.

Burrill, G., Allison, J., Breaux, G., Kastberg, S., Leatham, K., & Sanchez, W. (2002). *Handheld graphing technology in secondary mathematics*. Dallas, TX: Texas Instruments.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273-281.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational research, 65*(3), 245-281.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474-482.

Chance, B. L., & Rossman, A. J. (2005). *Investigating statistical concepts, applications, and methods*. Pacific Grove, CA: Duxbury Press.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293 - 332.

Clariana, R. B. (1999). *Differential memory effects for immediate and delayed feedback: A delta rule explanation of feedback timing effects*. Paper presented at the Association of Educational Communications and Technology annual convention, Houston, TX.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531-554.

Cobb, G. W. (1991). Teaching statistics: More data, less lecturing. *Amstat News, 182*, 1-4.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801-823.

Cobb, P., & Bauersfeld, H. (1995). *Emergence of mathematical meaning: Instruction in classroom cultures*. Hillsdale, NJ: Erlbaum.

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

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), The Challenge of Developing Statistical Literacy, Reasoning and Thinking (pp. 375-395). Dordrecht: Kluwer Academic Publishers.

Cognition and Technology Group at Vanderbilt (1997). *The Jasper project: Lessons in curriculum, instruction, assessment, and professional development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. R. B. (2007). *Research methods in education* (6th ed.). London and New York, NY: Routledge.

Cohen, V. B. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. *Educational Technology, 25*(1), 33-37.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 13*(1), 15-42.

Collis, B., de Boer, W., & Slotman, K. (2001). Feedback for web-based assignments. *Journal of Computer Assisted Learning, 17*(3), 306-313.

Corbett, A. T., & Anderson, J. R. (2001). *Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Seattle, WA, United States.

Davis, S. (2003). Observations in classrooms using a network of handheld devices. *Journal of Computer Assisted Learning, 19*(3), 298-307.

Davis, W. D., Carson, C. M., Ammeter, A. P., & Treadway, D. C. (2005). The interactive effects of goal orientation and feedback specificity on task performance. *Human Performance, 18*(4), 409-426.

De Corte, E., Verschaffel, L., & Eynde, P. O. (2000). Self-regulation: A characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 687-726). San Diego, CA: Academic Press.

Decoo, W. (1994). In defence of drill and practice in CALL: A reevaluation of fundamental strategies. *Computers & Education, 23*(1-2), 151-158.

Dede, C. (2000). Emerging influences of information technology on school curriculum. *Journal of Curriculum Studies, 32*(2), 281-303.

Dede, C. (2008). Theoretical perspectives influencing the use of information technology in primary and secondary education. In J. Voogt & G. Knezek (Eds.), *International handbook of information technology in education* (pp. 43-62). New York, NY: Springer.

Denzin, N. K. (2006). *Sociological methods: A sourcebook* (5th ed.). New Jersey, NJ: Transaction Publishers.

Denzin, N. K., & Lincoln, Y. S. (1994). Entering the field of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 1-17). Thousand Oaks, CA: Sage.

Dewey, J. (1929). *The quest for certainty: A study of the relation of knowledge and action*. New York, NY: Minton, Balch.

Doerr, H. M., & Zangor, R. (2000). Creating meaning for and with the graphing calculator. *Educational Studies in Mathematics, 41*(2), 143-163.

Doorman, L. M. (2005). *Modelling motion: From trace graphs to instantaneous change*. Utrecht: CD-ß Press / Freudenthal Institute.

Doyle, W., & Ponder, G. (1977). The practical ethic and teacher decision-making. *Interchange, 8*(3), 1-12.

Drijvers, P. (2003). *Learning algebra in a computer algebra environment*. Utrecht: CD-ß Press / Freudenthal Institute.

Drijvers, P., & Doorman, M. (1996). The graphics calculator in mathematics education. *Journal of Mathematical Behavior, 15*(4), 425.

Drijvers, P., Doorman, M., Boon, P., Reed, H., & Gravemeijer, K. (2010). The teacher and the tool: Instrumental orchestrations in the technology-rich mathematics classroom. *Educational Studies in Mathematics, 75*(2), 213-234. doi: 10.1007/s10649-010-9254-5

Dufresne, R. J., Gerace, W. J., Leonard, W. J., Mestre, J. P., & Wenk, L. (1996). Classtalk: A classroom communication system for active learning. *Journal of Computing in Higher Education, 7*, 3-47.

Dunleavy, M., Dextert, S., & Heinecke, W. F. (2007). What added value does a 1 : 1 student to laptop ratio bring to technology-supported teaching and learning? *Journal of Computer Assisted Learning, 23*(5), 440-452.

Elliott, J. (1991). *Action research for educational change*. Milton Keynes: Open University Press.

Even, R. (1993). Subject-matter knowledge and pedagogical content knowledge: Prospective secondary teachers and the function concept. *Journal for Research in Mathematics Education, 24*(2), 94-116.

Fies, C., & Marshall, J. (2006). Classroom response systems: A review of the literature. *Journal of Science Education and Technology, 15*(1), 101-109.

Foley, M. (2002). Instant messaging reference in an academic library: A case study. *College & Research Libraries, 63*(1), 36-45.

Freudenthal, H. (1971). Geometry between the devil and the deep sea. *Educational Studies in Mathematics, 3*(3/4), 413-435.

Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht: Reidel Publishing Company.

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

Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht: Reidel Publishing Company.

Freudenthal, H. (1991). *Revisiting mathematics education. China Lectures*. Dordrecht: Kluwer Academic Publishers.

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

Gal, I., & Garfield, J. B. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-13). Amsterdam: IOS.

Garfield, J. B. (1995). How students learn statistics. *International Statistical Review, 63*(1), 25-34.

Garfield, J. B., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44-63.

Garfield, J. B., & Gal, I. (2007). Assessment and statistics education: Current challenges and directions. *International Statistical Review, 67*(1), 1-12.

Geary, D. C. (2010). Missouri longitudinal study of mathematical development and disability. *British Journal of Educational Psychology Monograph Series II, Number 7 - Understanding number development and difficulties, 1*(1), 31-49.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304-315.
Gilmore, C. K., & Papadatou-Pastou, M. (2009). Patterns of individual differences in conceptual understanding and arithmetical skill: A meta-analysis. *Mathematical Thinking and Learning, 11*(1-2), 25-40. doi: 10.1080/10986060802583923
Ginsburg, H. P. (1997). Mathematics learning disabilities: A view from developmental psychology. *Journal of Learning Disabilities, 30*, 20-33.
Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Hawthorne, NY: Aldine de Gruyter.
Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology, 89*(5), 809-821.
Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology, 89*(2), 248-262.
Grant, L. K., & Courtoreille, M. (2007). Comparison of fixed-item and response-sensitive versions of an online tutorial. *Psychological Record, 57*(2).
Gravemeijer, K. P. E. (1994). *Developing realistic mathematics education*. Utrecht: CD-ß Press / Freudenthal Institute.
Gravemeijer, K. P. E. (1998). Developmental research as a research method. In J. Kilpatrick & A. Sierpinska (Eds.), *Mathematics education as a research domain: a search for identity (An ICMI study)* (Vol. 2, pp. 277-295). Dordrecht: Kluwer Academic Publishers.
Gravemeijer, K. P. E., & Cobb, P. (2006). Design research from a learning design perspective. In J. J. H. v. d. Akker, K. P. E. Gravemeijer, S. McKenney & N. Nieveen (Eds.), *Educational Design Research* (pp. 17-51). Abingdon: Routledge.
Guba, E. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational technology research and development, 29*(2), 75-91. doi: 10.1007/bf02766777
Gustafson, K. L., & Branch, R. M. (2002). *Survey of instructional development models* (4th ed.). Syracuse, NY: ERIC Clearinghouse on Information and Technology.
Hakkarainen, K. A. I., Muukonen, H., Lipponen, L., Ilomaki, L., Rahikainen, M., & Lehtinen, E. (2001). Teachers' information and communication technology (ICT) skills and practices of using ICT. *Journal of Technology and Teacher Education, 9*(2), 181-197.
Hannafin, M., Philips, T., Rieber, T., & Garhart, C. (1987). The effects of orienting activities and cognitive processing time on factual and inferential learning. *Educational Communications and Technology Journal, 35*(2), 75-84.
Harackiewicz, J. M., Manderlink, G., & Sansone, C. (1984). Rewarding pinball wizardry: Effects of evaluation and cue value on intrinsic interest. *Journal of Personality and Social Psychology, 47*(2), 287-300.
Hartman, H. J. (2002). Scaffolding and cooperative learning *Human learning and instruction* (pp. 23-69). New York, NY: City College, University of New York.
Hativa, N. (1988). Computer-based drill and practice in arithmetic: widening the gap between high- and low-achieving students. *American Educational Research Journal, 25*(3), 366-397.
Hattie, J. A. (1999). Influences on student learning Retrieved December 22nd, 2011, from http://www.education.auckland.ac.nz/uoa/fms/default/education/staff/Prof.%20John%20Hattie/Documents/Presentations/influences/Influences_on_student_learning.pdf

Hattie, J. A. (2009). *Visible learning: A synthesis of meta-analyses in education*. London: Routledge.

Hattie, J. A., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hegedus, S. J., & Kaput, J. (2001). *New activity structures exploiting wirelessly connected graphing calculators.* Paper presented at the 23rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Columbus, OH.

Hegedus, S. J., & Kaput, J. (2002). *Exploring the phenomenon of classroom connectivity.* Paper presented at the 26th Meeting for the International Conference for the Psychology of Mathematics Education, Atlanta, GA, USA.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*(1), 33-46.

Hemelrijk, J. (1968). *Report on the desirability en possibility for the introduction of statistics in secondary mathematics education: Committee modernization mathematics curriculum* [Dutch: Rapport over de wenselijkheid en mogelijkheid van het invoeren van statistiek in het onderwijs voor M.A.V.O., H.A.V.O. en V.W.O.].

Higgins, S., Beauchamp, G., & Miller, D. (2007). Reviewing the literature on interactive whiteboards. *Learning, Media and Technology, 32*(3), 213-225.

Hoefakker, R. (2002). Like real life; Reaction on the paper 'Authentic learning with ICT' [Dutch: Levensecht; reactie op artikel 'Authentiek leren met ICT']. *Onderwijsinnovatie*(4), 121-123.

Hollingsworth, S. (1997). *International action research: A casebook for educational reform*. London: Falmer Press.

Hong, E., & Karstensson, L. (2002). Antecedents of state test anxiety. *Contemporary Educational Psychology, 27*(2), 348-367.

Hughes, J. (2005). The role of teacher knowledge and learning experiences in forming technology-integrated pedagogy. *Journal of Technology and Teacher Education, 13*(2), 277-302.

Jones, G. A., Langrall, C. W., & Mooney, E. S. (2007). Research in probability: Responding to classroom realities. In F. K. Lester (Ed.), *The second handbook of research on mathematics* (pp. 909-956). Reston, VA: National Council of Teachers of Mathematics (NCTM).

Karabenick, S. A., & Knapp, J. R. (1988). Effects of computer privacy on help seeking. *Journal of Applied Social Psychology, 18*(6), 461- 472.

Kelchtermans, G. (1993). Teachers and their career story: A biographical perspective on professional development. In C. Day, J. Calderhead & P. Denicolo (Eds.), *Research on teacher thinking: Understanding professional development* (pp. 198-220). London: The Falmer Press.

Kelly, A. E. (2003). Theme issue: The role of design in educational research. *Educational Researcher, 32*(1), 3.

Kelly, A. E. (2004). Design research in education: Yes, but is it methodological? *Journal of the Learning Sciences, 13*(1), 115-128.

Kelly, A. E. (2007). When is design research appropriate. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 73-88). Enschede: SLO.

Kelly, A. E., Lesh, R. A., & Baek, J. Y. (Eds.). (2008). *Handbook of design research methods in education; Innovations in science, technology, engineering, and mathematics learning and teaching*. New York, NY: Routledge.

Klecker, B. M. (2007). The impact of formative feedback on student learning in an online classroom. *Journal of Instructional Psychology, 34*(3), 161-165.
Kleijne, W. (2008). The introduction of statistics and probability theory in secondary education. *Euclides, 83*(4), 135-138.
Kluger, A. N., & Adler, S. (1993). Person- versus computer-mediated feedback. *Computers in Human Behavior, 9*(1), 1-16.
Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.
Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67-72.
Knoblauch, C. H., & Brannon, L. (1981). Teacher commentary on student writing: The state of the art. *Freshman English News, 10*(2), 1-4.
Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14*(4), 511-550.
Koedinger, K. R., & Anderson, J. R. (1993). *Effective use of intelligent software in high school math classrooms.* Paper presented at the Proceedings of the World Conference on Artificial Intelligence in Education, Charlottesville, VA.
Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30-43.
Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education, 9*(1).
Kort, B., Reilly, R., & Picard, R. W. (2001). *An affective model of interplay between emotions and learning: Reengineering educational pedagogy - building a learning companion.* Paper presented at the IEEE International Conference on Advanced Learning Technologies: Issues, Achievements and Challenges, Madison, WI.
Korthagen, F. A. J., & Kessels, J. P. A. M. (1999). Linking theory and practice: changing the pedagogy of teacher education. *Educational Researcher, 28*(4), 4-17. doi: 10.3102/0013189x028004004
Kromhout, O. M. (1972). Effect of computer tutorial review lessons on exam performance in introductory college physics. *Tech Memo* (Vol. 64). Tallahassee, FL: Florida State University Computer-Assisted Instruction Center.
Küchemann, D. E. (1981). Algebra. In K. M. Hart (Ed.), *Children's understanding of mathematics: 11-16* (pp. 102-119). London: John Murray.
Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211-232.
Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology, 10*(3), 285-291.
Kulik, C. L., & Kulik, J. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*, 75-94.
Kulik, J. A., & Kulik, C. L. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79-97.
Lesh, R., Amit, M., Schorr, R. Y., Gal, I., & Garfield, J. (1997). Using "real-life" problems to prompt students to construct conceptual models for statistical reasoning. *The Assessment Challenge in Statistics Education* (pp. 65-84). Amsterdam: IOS Press.
Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Liu, N.-F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279-290.

Loether, H. J., & McTavish, D. G. (1988). *Descriptive and inferential statistics*. Boston, MA: Allyn and Bacon Division of Simon and Schuster.

Lovett, M. C. (2001). A collaborative convergence on studying reasoning processes: a case study in statistics. In S. Carver & D. Klahr (Eds.), *Cognition and Instruction: Twenty-Five Years of Progress* (pp. 347-384). Mahwah, NJ: Erlbaum.

Lynch, C. F., Ashley, K., Aleven, V., & Pinkwart, N. (2006). *Defining "ill-defined" domains: A literature survey.* Paper presented at the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems, Jhongli, Taiwan.

Mao, X., & Li, Z. (2010). Agent based affective tutoring systems: A pilot study. *Computers & Education, 55*(1), 202-208.

Martin, G. W., Carter, J., Forster, S., Howe , R., Kader, G., Kepner, H., et al. (2009). *Focus in high school mathematics: reasoning and sense making*. Reston, VA: NCTM.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: research-based strategies for increasing student achievement.* . Alexandria, VA: Association for Supervision and Curriculum Development.

Mayer, R. E. (2009). *Multimedia Learning* (2nd ed.). Cambridge, MA: Cambridge University Press.

Mazur, E. (2009). Farewell, lecture? *Science, 323*(5910), 50-51. doi: 10.1126/science.1168927

McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine, Part I. *Communications of the ACM, 3*, 184-195.

McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanisms in number processing and calculation: Evidence from dyscalculia. *Brain and Cognition, 4*, 171-196.

McColskey, W., & Leary, M. R. (1985). Differential effects of norm-referenced and self-referenced feedback on performance expectancies, attribution, and motivation. *Contemporary Educational Psychology, 10*, 275-284.

McIntosh, M. E. (1997). Formative assessment in mathematics. *The Clearing House, 71*(2), 92-96.

McIntyre, D., & Brown, S. (1979). Science teachers' implementation of two intended innovations. *Scottish Educational Review, 11*(1), 42-57.

Merriam, S. B. (1988). *Case study research in education*. San Francisco, CA: Jossey Bass.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: a review of the literature. *Journal of Statistics Education, 10*(1), http://www.amstat.org/publications/jse/v10n1/mills.html

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: a framework for teacher knowledge. *Teachers College Record, 108*(6), 1017-1054.

Molenda, M. (2003). In search of the elusive ADDIE model. *Performance Improvement, 42*(5), 34-36.

Molenda, M., Pershing, J. A., & Reigeluth, C. M. (1996). Designing instructional systems. In R. L. Craig (Ed.), *The ASTD training and development handbook* (4th ed., pp. 266-293). New York, NY: McGraw-Hill.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: new approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Moore, D. S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review, 65*(2), 123-137.
Moore, M. G. (1989). Three types of interaction. *American Journal of Distance Education, 3*(2), 1-6.
Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*, 99-113.
Morris, E. J., Joiner, R., & Scanlon, E. (2002). The contribution of computer-based activities to understanding statistics. *Journal of Computer Assisted Learning, 18*(2), 116-126.
Mory, E. H. (2004). Feedback research revisited. In D. Jonassen & P. Harris (Eds.), *Handbook of research on educational communications and technology* (2nd ed., pp. 745-783). Mahwah, NJ: Lawrence Erlbaum.
Mory, E. H., & Jonassen, D. H. (1996). Feedback Research. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology*. London: Prentice Hall International.
Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33-52.
Nagel, S. M., & Grant, L. K. (Producer). (2007, December 21 2011). Introductory biological psychology tutorials. Retrieved from http://psych.athabascau.ca/html/Psych289/Biotutorials/index.shtml?sso=true
Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated Learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199.
Nieveen, N. (2009). Formative evaluation in educational design research. In T. Plomp & N. Nieveen (Eds.), *An introduction to educational design research* (pp. 89-101). Enschede: SLO.
Nieveen, N., McKenney, S., & Van den Akker, J. J. H. (2006). Educational design research; the value of variety. In J. J. H. Van den Akker, K. P. E. Gravemeijer, S. McKenney & N. Nieveen (Eds.), *Educational design research* (pp. 151-158). Oxford: Routledge.
Onderwijsinspectie. (2001). *The second stage one stage further*. [Dutch: De Tweede Fase een fase verder]. Utrecht: Onderwijsinspectie.
Onderwijsinspectie (2011). *The disappearance of Dutch mathematics teachers through the rise of self-supported learning*. [Dutch: Het verdwijnen van wiskundeleraren door de opkomst van zelfstandig leren] Utrecht: Onderwijsinspectie.
Park, O. (1987). Conventional CBI versus intelligent CAI: Suggestions for the development of future systems. *Educational Technology, 27*(5), 15-21.
Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 3-8.
Passey, D., Rogers, C., Machell, J., McHugh, G., & Allaway, D. (2004). *The motivational effect of ICT on pupils*. Lancaster: Department of Educational Research Lancaster University.
Paulhus, D. (1991). Measurement and control of response bias. In P. R. S. J. Robinson, and L. S. Wrightsman (Ed.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17-59). New York, NY: Academic Press.
Paulos, J. A. (1988). *Innumeracy, mathematical illiteracy and its consequences*. London: Penguin Books.

Peled, I., & Zaslavsky, O. (2008). Beyond local conceptual connections: Meta-knowledge about procedures. *For the Learning of Mathematics, 28*(3), 28-35.

Penuel, W. R., Boscardin, C. K., Masyn, K., & Crawford, V. M. (2007). Teaching with student response systems in elementary and secondary education settings: A survey study. *Educational Technology Research and Development, 55*(4), 315-336.

Pereira-Mendoza, L., & Swift, J. (1981). Why teach statistics and probability? In A. Schulte & J. Smart (Eds.), *Teaching statistics and probability. 1981 Yearbook of the National Council of Teachers of Mathematics* (pp. 1-7). Reston, VA: NCTM.

Phillips, T., Hannafin, M., & Tripp, S. (1988). The effects of practice and orienting activities on learning from interactive video. *Educational Communication and Technology Journal, 36*, 93-102.

Phye, G. D., & Sanders, C. E. (1994). Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology, 19*(3), 286-301.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Plato (2001). Five dialogues (G. M. A. Grube, Trans.). In J. M. Cooper (Ed.), (2nd ed.). Indianapolis, IN: Hackett Publishing Company Inc.

Pol, H. J., Harskamp, E. G., & Suhre, C. J. M. (2005). Solving physics problems with the help of computer-assisted instruction. *International Journal of Science Education, 27*(4), 451 - 469.

Pol, H. J., Harskamp, E. G., & Suhre, C. J. M. (2008). The effect of the timing of instructional support in a computer-supported problem-solving program for students in secondary physics education. *Computers in Human Behavior, 24*(3), 1156-1178.

Polanyi, M. (1966). *The tacit dimension*. New York, NY: Anchorday Books.

Pressey, S. L. (1926). A simple apparatus which gives tests and scores and teaches. *School and Society, 23*, 373-376.

Pressey, S. L. (1950). Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. *The Journal of Psychology, 29*, 417 - 447.

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology, 20*(4), 444-450.

Quesada, A. R., & Maxwell, M. E. (1994). The effects of using graphing calculators to enhance college students' performance in precalculus. *Educational Studies in Mathematics, 27*(2), 205-215.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13.

Rapp, B., & Goldrick, M. (2004). Feedback by any other name is still interactivity: a reply to Roelofs (2004). *Psychological Review, 111*(2), 573-578.

Reeves, T. C. (2006). Design research from a technology perspective. In J. J. H. van den Akker, K. P. E. Gravemeijer, S. McKenney & N. Nieveen (Eds.), *Educational Design Research* (pp. 52-67). Abingdon: Routledge.

Richey, R., & Nelson, W. (1996). Developmental research. In D. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology* (pp. 1213-1245). London: McMillan.

Rieber, L. (1996). Animation as feedback in a computer-based simulation: Representation matters. *Educational technology research and development, 44*(1), 5-22.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155-169.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346-362.

Roback, P. J. (2003). Teaching an advanced methods course to a mixed audience. *Journal of Statistics Education, 11*(2).

Roelofs, E. C., & Houtveen, A. A. M. (1999). Pedagogy of authentic learning in lower secondary education. [Dutch: Didactiek van authentiek leren in de basisvorming]. *Pedagogische Studiën 76*(4), 237 - 257.

Romberg, T. A., Carl, I. M., Hirsch, C. R., Crosswhite, J. F., Lappan, G., Dossey, J. A., et al. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.

Roschelle, J. (2003). Unlocking the learning value of wireless mobile devices. *Journal of Computer Assisted Learning, 19*(3), 260-272.

Roschelle, J., & Pea, R. (2002). A walk on the WILD side: How wireless handhelds may change computer-supported collaborative learning. *International Journal of Cognition and Technology, 1*(1), 145-168.

Roschelle, J., Tatar, D., Vahey, P., Kaput, J., & Hegedus, S. J. (2003). *Five key considerations for networking in a handheld-based mathematics classroom.* Paper presented at the Joint Meeting of PME and PMENA, Honolulu, Hawai.

Rosnick, P., & Clement, J. (1980). Learning without understanding: the effect of tutorial strategies on algebra misconceptions. *Journal of Mathematical Behaviour, 3*(1), 3-27.

Rowe, M. B. (1974). Wait time and rewards as instructional variables, their influence on language, logic and fate control. *Journal of Research in Science Teaching, 11*(2), 81-94.

Rubin, A. V., Rosebery, A. S., & Bruce, B. (1988). *ELASTIC and reasoning under uncertainty. Research report no. 6851.* Boston, MA: BBN Systems and Technologies Corporation.

Rumsey, D. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3).

Ryle, G. (1971). University Lectures *Collected Papers* (Vol. 2, pp. 480-496). London: Hutchinson.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77 - 84.

Salisbury, D. F. (1990). Cognitive psychology and its implications for drill and practice for computers. *Journal of Computer Based Instruction, 17*(1), 23-30.

Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., & Bigdeli, A. (2008). "How do you know that I don't understand?" A look at the future of intelligent tutoring systems. *Computers in Human Behavior, 24*(4), 1342-1363.

Schield, M. (2002). *Three kinds of statistical literacy: what should we teach?* Paper presented at the ICOTS-6: The Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

Schofield, J. W., Evans-Rhodes, D., & Huber, B. R. (1990). Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students. *Social Science Computer Review, 8*(1), 24-41.

Schroth, M. L. (1992). The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary Educational Psychology, 17*(1), 78-82.

Scott, P. H., Mortimer, E. F., & Aguiar, O. G. (2006). The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Science Education, 90*(4), 605-631.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1-36.

Shaughnessy, J. M. (2010). Statistics for all - the flip side of quantitative reasoning. *President's Corner* Retrieved 2011-12-29, from http://www.nctm.org/about/content.aspx?id=26327

Shaughnessy, J. M., Garfield, J. B., & Greer, B. (1996). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick & C. Laborde (Eds.), *International handbook of mathematics education* (Vol. 1, pp. 205-237). Dordrecht: Kluwer Academic Publishers.

Shute, V. J. (2008). The focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Silva, E. (2009). Measuring skills for 21st-century learning. *Phi Delta Kappan, 90*(9), 630-634.

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

Simon, M. A. (2009). Amidst multiple theories. *Journal for Research in Mathematics Education, 40*(5), 477-490.

Simpson, V., & Oliver, M. (2007). Electronic voting systems for lectures then and now: A comparison of research and practice. *Australasian Journal of Educational Technology, 23*(2), 187-208.

Sleeman, D., & Brown, J. (1982). *Intelligent tutoring systems*. London: Academic Press.

Smaling, A. (1990). Some aspects of qualitative research and the clinical interview (Dutch: Enige aspecten van kwalitatief onderzoek en het klinisch interview). *Tijdschrift voor nascholing en onderzoek van het reken-wiskundeonderwijs, 8*(3), 4-10.

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education, 6*(3).

Snell, L. (1999). *Using Chance media to promote statistical literacy.* Paper presented at the Joint Statistical Meetings, Dallas, TX.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776-778.

Staatsblad. (1997). *Law of 2 July 1997 amending the profiles in secondary education*. [Dutch: Wet van 2 juli 1997 tot wijziging van de profielen voortgezet onderwijs]. The Hague: SDU.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education, 36*(5), 404-411.

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

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York, NY: Wiley.

Streibel, M. J. (1986). A critical analysis of the use of computers in education. *Educational technology research and development, 34*(3), 137-161.

Stroup, W. M., Ares, N. M., & Hurford, A. C. (2005). A dialectic analysis of generativity: Issues of network-supported design in mathematics and science. *Mathematical Thinking & Learning, 7*(3), 181-206.
Sweeney, J., O'Donoghue, T., & Whitehead, C. (2004). Traditional face-to-face and web-based tutorials: A study of university students' perspectives on the roles of tutorial participants. *Teaching in Higher Education, 9*(3), 311-323.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science: A Multidisciplinary Journal, 12*(2), 257-285.
Sykes, E. (2007). Developmental process model for the Java intelligent tutoring system. *Journal of Interactive Learning Research, 18*(3), 399-410.
Tolboom, J. L. J. (2005). Wireless network in the mathematics classroom [Dutch: Draadloos netwerk in de wiskunde klas]. *Euclides, 81*(3), 108-112.
Trees, A. R., & Jackson, M. H. (2007). The learning environment in clicker classrooms: student processes of learning and involvement in large university-level courses using student response systems. *Learning, Media and Technology, 32*(1), 21-40.
Treffers, A. (1987). *Three dimensions: a model of goal and theory description in mathematics instruction - The Wiskobas project*. Dordrecht: Reidel Publishing Company.
Treffers, A. (1993). Wiskobas and Freudenthal realistic mathematics education. *Educational Studies in Mathematics, 25*(1/2), 89-108.
Trilling, B., & Fadel, C. (2009). *21st century skills: learning for life in our times*. San Fransisco, CA: Jossey-Bass.
Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.
van den Akker, J. J. H. (1988). *Design and implementation of science education* [Dutch: Ontwerp en implementatie van natuuronderwijs]. Amsterdam / Lisse: Swets & Zeitlinger.
van den Akker, J. J. H. (1999). Principles and methods of development research. In J. J. H. van den Akker, R. M. Branch, K. L. Gustafson, N. Nieveen & T. Plomp (Eds.), *Design approaches and tools in education and training* (pp. 1-14). Dordrecht: Kluwer Academic Publishers.
van den Akker, J. J. H. (2003). Curriculum perspectives; An introduction. In J. J. H. van den Akker, W. Kuiper & U. Hameyer (Eds.), *Curriculum landscapes and trends* (pp. 1-11). Dordrecht: Kluwer Academic Publisher.
van den Akker, J. J. H., Gravemeijer, K. P. E., McKenney, S., & Nieveen, N. (Eds.). (2006). *Educational design research*. Oxford: Routledge.
van den Akker, J. J. H., & Kuiper, W. (2008). Research on models for instructional design. In J. M. Spector, M. D. M.D. Merrill, J. v. Merriënboer & M. P. Driscoll (Eds.), *Handbook of research on educational communication and technology* (Third ed., pp. 739-748). New York, NY: Lawrence Erlbaum Associates.
van den Heuvel-Panhuizen, M. (2005). Can scientific research answer the 'what' question of mathematics education? *Cambridge Journal of Education, 35*(1), 35-53.
van der Hilst, B. (2010). *Organising learning* [Dutch: Het leren organiseren]. Amsterdam: Center for professional development for teachers [Dutch: Centrum voor nascholing).
van Streun, A. (2001). *Promoting thinking* [Dutch: Het denken bevorderen]. Groningen: University of Groningen.
Vanderlinde, R., & van Braak, J. (2010). The gap between educational research and practice: views of teachers, school leaders, intermediaries and researchers. *British Educational Research Journal, 36*(2), 299 - 316.

Verschaffel, L., & de Corte, E. (1996). Number and arithmatic. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 99-137). Dordrecht: Kluwer Academic Publishers.

Vickerman, P. (2009). Student perspectives on formative peer assessment: an attempt to deepen learning? *Assessment & Evaluation in Higher Education, 34*(2), 221-230.

Vinsonhaler, J., & Bass, R. (1972). A summary of ten major studies on CAI drill practice. *Educational Technology, 12*(7), 29-32.

Voogt, J. (2003). Consequences of ICT for aims, contents, processes and environments of learning. In J. J. H. van den Akker, W. Kuiper & U. Hameyer (Eds.), *Curriculum landscapes and trends* (pp. 217-236). Dordrecht: Kluwer Academic Publishers.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT press.

Vygotsky, L. S. (1978). *Mind and society: the development of higher mental processes*. Cambridge, MA: Harvard University Press.

Wagemann, E. (1935). *Narrenspiegel der Statistik*. Hamburg: Hanseatische Verlagsanstalt.

Waimon, M. D. (1962). Feedback in classrooms - a study of corrective teacher responses. *Journal of Experimental Education, 30*(4), 355-359.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research & Development, 53*(4), 5-23.

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation, 1*(3), 247-275.

Watson, J. M., Gal, I., & Garfield, J. B. (1997). Assessing statistical thinking using the media. In J. B. Garfield & I. Gal (Eds.), *The Assessment Challenge in Statistics Education*. Amsterdam: IOS Press and International Statistical Institute.

Weissglass, J., & Cummings, D. (1991). Dynamic visual experiments with random phenomena. In W. Zimmermann & S. Cunningham (Eds.), *Visualization in teaching and learning mathematics* (pp. 215-223). Washington, DC: Mathematical Association of America Committee on computers in mathematics education.

Wiberg, M. (2009). Teaching statistics in integration with psychology. *Journal of Statistics Education, 17*(1).

Wiggins, G., & McTyghe, J. (2005). *Understanding by design* (2 ed.). Alexandria, VA: ASCD.

Wijers, M., Jonker, V., & Kemme, S. (2004). Authentic contexts in mathematics textbooks in vocational education [Dutch: Authentieke contexten in wiskundemethoden in het vmbo]. *Tijdschrift voor Didactiek der β-wetenschappen, 21*(1), 1-19.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

Wiliam, D. (2007). Keeping learning on track: classroom assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053-1098). Greenwich, CT: Information Age Publishing.

Williams, S. E. (1997). *Teachers' written comments and students' responses: A socially constructed interaction*. Paper presented at the Annual meeting of the Conference on College Composition and Communication, Phoenix, AZ.

Witterholt, M., Van Streun, A., Goedhart, M. J., & Beijaard, D. (2007). Statistical investigation in grade 9; Acquiring skills for statistical investigations [Dutch: Statistisch onderzoek door 3 havo; het leren van statistische

onderzoeksvaardigheden]. *Tijdschrift voor Didactiek der β-wetenschappen, 24*(1-2), 31-58.
Wong, W. L., Shen, C., Nocera, L., Carriazo, E., Tang, F., Bugga, S., et al. (2007). *Serious video game effectiveness.* Paper presented at the International conference on advances in computer entertainment technology, Salzburg, Austria.
Woods, P. (1993). Critical events in education. *British Journal of Sociology of Education, 14*(4), 355-371.
Yarnall, L., Shechtman, N., & Penuel, W. (2006). Using handheld computers to support improved classroom assessment in science: Results from a field trial. *Journal of Science Education and Technology, 15*(2), 142-158.
Yin, R. K. (2003). *Case study research. Design and methods* (4th ed.). Thousand Oaks, CA: Sage.

Summary

One of the persistent problems in mathematics education is the perception of mathematics teachers that they lack sufficient time to teach their students mathematics good enough. The basic idea behind this study is to improve the quality of the lesson time in mathematics education. A vast body of research states that one of the most effective ways to improve this lesson time is to focus on feedback. Computer devices connected in a network have strong feedback possibilities. During an initial study, this is confirmed in a setting with graphing calculators that were connected in a classroom network. Therefore, in this study we tried to answer the main research question "*What are the potentials of a classroom network in supporting teachers in providing feedback in statistics education?*"

In order to answer this question we conducted a literature review on feedback and on statistics education in order to define our position with respect to these topics.

In this study, we considered 'feedback' to be the teacher's, content's and peers' comments on students' work in order to improve the students' learning, while the students' work and behaviour are a source of feedback for the teacher in order to modify instruction.

We chose to divide the statistical learning activities in statistics education into:

1. reasoning and sense making (Martin, et al., 2009) with and about data, (Cobb's (1991) 'data and concepts'), to be called *data literacy* (DL);
2. *algorithmic statistical skills* (ASS), (Cobb's (1991) recipes).

Data literacy belongs to conceptual knowledge which is, in our opinion, too often ignored in statistics education. Without denying the vital importance of procedural skills (in this study denoted by algorithmic statistical skills), in this study we tried to foster DL. We restricted ourselves to the sub-domain of *descriptive* statistics. We have no reason to believe that feedback in descriptive statistics is more important than in other sub-domains of statistics education. The main reason is that descriptive statistics is usually the introduction to statistics education, making it interesting because it is possible to give a good impression of statistics. The use of descriptive statistics is that widespread that it is hard to open up a newspaper without being exposed to it. Apart from the fact that descriptive statistics is used in a wide variety of sciences, it is important to have knowledge of its concepts and methods for those who want to be a member of today's information-based society (Gal & Garfield, 1997). But utility for and future study of secondary school students are not the only reasons. Pereira-Mendoza and Swift (1981) added aesthetics to the curricular goals of statistics, which we adopt as a third reason for choosing it as a learning domain to conduct our research. In our view, the beauty should be sought, for the target group in this study (grade 10 senior secondary education), primarily in the possibility to analyse real life phenomena, caught in authentic contexts (Wijers, Jonker, & Kemme, 2004), with the use of mathematical techniques.

The approach that seemed best suitable to generate data analysis of which could answer our main research question is educational design research (EDR). A process of analysis, design, development, implementation and evaluation (ADDIE) of a prototype was iteratively deployed.

We specified the main research question into four subquestions:

1. Is technological support by means of the classroom network adequate for the intended feedback in the lessons? (*Conditional* question)

2. Is it possible for a mathematics teacher to implement the prototype in accordance with the intentions? (*Existence* question)
3. Is the feedback support of the classroom network equal for algorithmic statistical skills (ASS) and data literacy (DL)? (*Didactical* question)
4. Which teacher characteristics promoted/hindered the implementation of the CN as intended? (*Identificatio*n question)

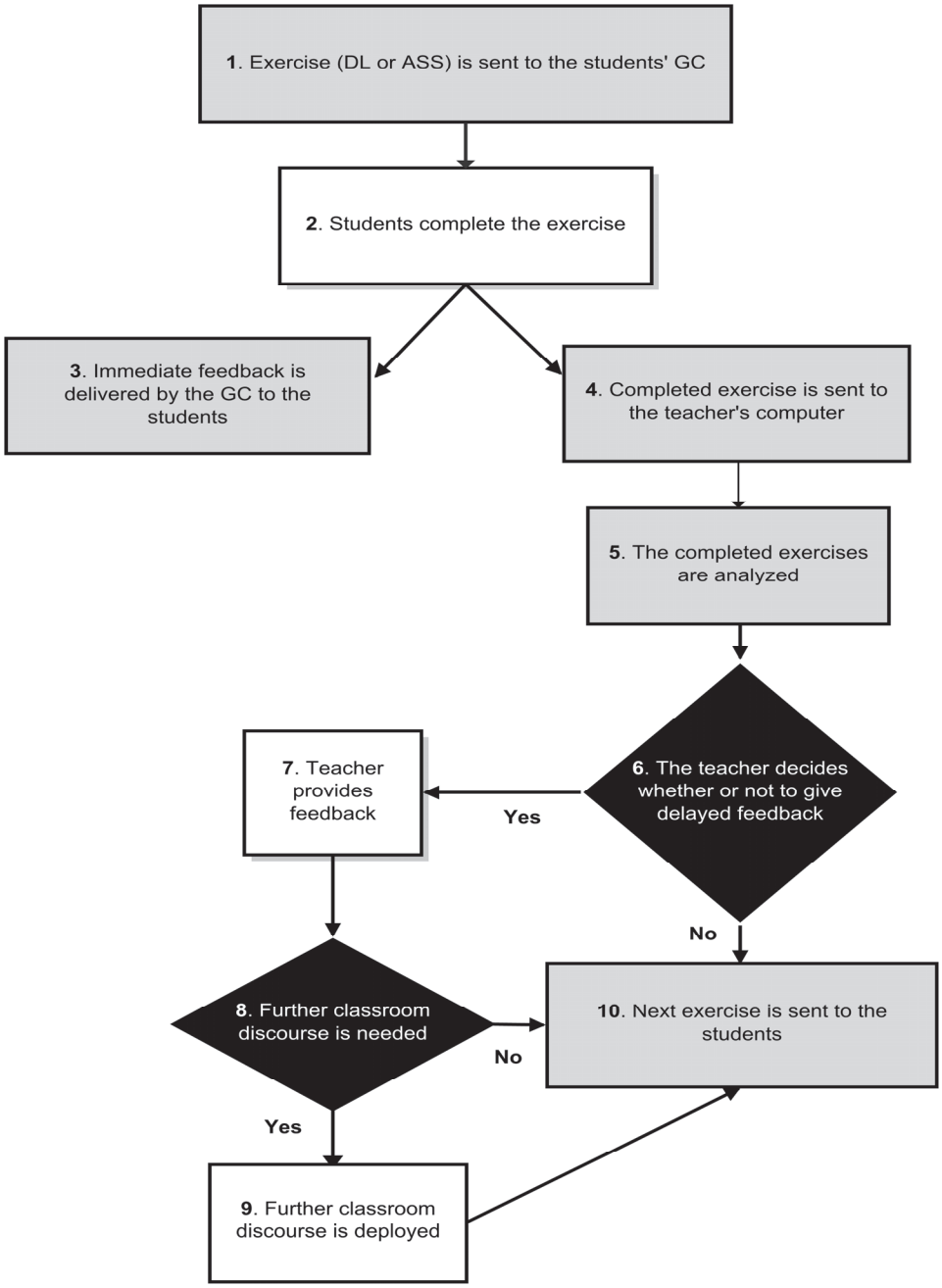
The development of the prototype was guided by design principles with respect to feedback and statistics education (with ICT interwoven). The most important of these principles were framed into the *feedback matrix for statistics education* (see below), that categorises statistical exercises with respect to the nature of the intended learning goal (data literacy or algorithmic statistical skills) and the timing of the feedback (immediate or delayed).

Timing of feedback / Learning goal	Immediate	Delayed
Data literacy	Type I	Type II
Algorithmic statistical skills	Type III	Type IV

Feedback matrix for statistics education

In order to make students as active as possible in this study we used the *questionising approach*: try to represent as many of the learning and teaching activities as possible as questions to the students. Students' responses are a feedback possibility for the teacher and they are themselves feedback for the teacher, because they inform the teacher about how well students are performing.

We used exercise types like ‘multiple choice’ and ‘fill in the blank’ (asking numerical input from the student) in order to facilitate type I and type III feedback. Exercise type ‘open response’(asking an essay input form the student) was more suitable for type II and type IV feedback. Roughly, immediate feedback is delivered on the students' graphing calculators, and delayed feedback is given by the teacher during the next class. A scheme for the process of feedback is depicted below.



Scheme of the feedback process

During three empirical stages (C1, C2, C3) the prototype was implemented, evaluated and adapted. We developed hypothetical teaching trajectories (HTT) in order to make our intentions with respect to the teacher behaviour explicit. All the lessons were observed, videotaped and scored. We developed a coding system to compress the –lengthy- intentions and a correspondence metric with which we expressed the correspondence between the realised intervention (as observed) and the intended intervention (as described in the HTT). Each stage is also evaluated using a questionnaire for all participating students, interviews with participating students (three from each group) and teacher interviews. After C3 a group interview with all participating teachers was organised in order to tune experiences.

After C3, during which the third prototype was implemented in six case studies, this eventually yielded in a correspondence scheme as presented below. The correspondence

is expressed in a score varying from 0 (no correspondence whatsoever) to 7 (perfect correspondence) with respect to the feedback sessions as realised during the case studies. This correspondence is relative to the HTT.

Case	Mean	SD	Mean_DL	Mean_ASS	M_DL-M_ASS	%-Missing
S1	5.14	1.70	4.89	5.60	-0.71	68.09
S2A	4.40	2.29	3.66	5.71	-2.05	14.89
S3	5.38	1.85	5.34	5.46	-0.12	0.00
S4A	3.60	2.13	3.65	3.44	0.21	31.91
S4B	3.89	2.09	4.11	3.33	0.78	34.04
S2B	2.04	1.70	1.95	2.25	-0.30	59.57

Correspondence scheme for the six case studies during C3

We see a considerable range with respect to the mean correspondence as well as to the percentage missing. In order to understand this more or less quantitative overview in a qualitative way, we analysed the classroom discourse also on a more microscopic level. For each case study we presented some examples. A major criticism on classical reports of educational design research studies concerns the exemplary character of the results. We think we met this criticism by the presentation of the correspondence scores for a great deal. But besides this, for C3, we explicitly described the way we sampled these examples from the abundance of classroom discourse we had collected. First of all, the selected example had to be *substantial*: that is, it had to contain one or more events that are interesting from the perspective of our research question. This criterion for sampling makes it likely that the correspondence score of the selected events is higher than the mean correspondence score of the case study. Further, we selected examples concerning mainly DL. As mentioned, we consider ASS to be very important but the instruction of DL is perhaps even more tedious. Improving that is a main concern of this study. Then, the selected example had to be 'somehow representative' for the case study as a whole. That is, the classroom discourse had to contain elements that were more or less typical for the specific case study. Hence, we decided to present two examples of classroom discourse. In order to make a comparison of the case studies more valid, we tried to find an exercise on DL that gave rise to substantial classroom discourse in all of the six case studies. Due to the considerable problems in the first and the last case studies, this was not realisable. The classroom discourse that sparked from the feedback on exercise 8.8 came, over six case studies, most close to this criterion. With this exercise, asking the students to draw a conclusion about the computer behaviour of boys and girls, we have a point of comparison for the first five case studies, which were the five best out of six with respect to the mean correspondence between the HTT and the implemented curriculum.

Reasoning this way, we came to answer the four research questions as follows:

1. Is technological support by means of the classroom network adequate for the intended feedback in the lessons? (*Conditional* question)

   Because of technological instability it was not when we started this study. But after C3, where in case S3 it had been possible to conduct every intended

feedback session in the classroom setting, we concluded that technological support of the classroom network was adequate.

2. Is it possible for a mathematics teacher to implement the prototype in accordance with the intentions? (*Existence* question)

   We consider case S3 to be a convincing implementation of the intended intervention. This had a high mean correspondence score, both with respect to ASS and DL. Students and the teacher were equally enthusiastic about the improvement of feedback. All of the feedback sessions were carried out as intended, demonstrating that the technology served the intervention very well. Besides a convincing case study being a *proof of existence* for the goals of the intervention it is remarkable that in every single case study there were feedback sessions with a convincing correspondence score. This means that every teacher, under the right circumstances, had been able to conduct a feedback session as planned. We consider these as '*micro proofs of existence*': the teacher succeeded in conducting at least one feedback session sufficiently according to the intentions while the students were convinced of the feedback potential.

3. Is the feedback support of the classroom network equal for algorithmic statistical skills (ASS) and data literacy (DL)? (*Didactical* question)

   The support for ASS proved to be better, but, with a highly specified HTT, we managed to support the teachers in giving feedback on DL in a satisfactory way. The slightly better support for ASS is shown by the fact that the mean difference in correspondence score between DL and ASS was 0.37 (in the advantage of ASS). We consider this gap, with respect to a variable on a scale from 0 to 7, to be quite small.

   The 'built in' support of the CN for developing students' DL has to be completed by specific teaching methods and by more directing teacher preparation.

4. Which teacher characteristics promoted/hindered the implementation of the CN as intended? (*Identification* question)

   We concluded in chapter 7 that the data as collected during C3 proved that improving feedback in statistics education by the use of a classroom network was possible. The little difference in corresponding scores between S4a and S4b (with the same teacher for different groups) suggests that correspondence is more teacher dependent than group dependent. With respect to the teacher, this brings up an interesting question related to our main research question: what are the strong teacher influences that cause this variation in correspondence score?

   We concluded that there are at least *four conditions* that have to be met before a teacher, trained and supported as we did during C3, in a learning environment that is technically stable, can fully utilise the feedback potential of the classroom network in statistics education.

   First of all, there should be a relationship between teacher and group that is based on *sufficient mutual trust*. If this trust is lacking, all education is to fail, however well-resourced the learning environment potential may be. Good education is an intimate process. Feedback and classroom discussion are perhaps the most vulnerable parts of it. Mutual trust is indispensable for making these succeed. The teacher is the one who

Secondly, the teacher has to have deep *conversational skills*, including the attitude (or is it even 'personality'?) to apply them as productively as possible in the classroom discourse. This means that she or he has to be a 'conductor' (Drijvers, Doorman, Boon, Reed, & Gravemeijer, 2010) of the classroom discourse, which in this context should be interpreted as 'the spider in the web of the educational process'. A sufficient level of *functional extraversion* is needed in order to be able fulfill this. In general the teacher's repertoire on formative assessment (Black, et al., 2003; Black & Wiliam, 2009) and dialogic teaching (Alexander, 2008) has to be at a sufficient level.

Thirdly, besides these conversational skills, the teacher should have competence in *quickly interpreting students' answers* as he has a greater number of these to handle than without the use of a classroom network. He should be able to make 'statistical sense' of much more student input than before, for example by making a fast and rough 'feedback scheme' based on students' answers. This capacity in interpretation of students' input requires sufficient *subject matter (mathematical) knowledge and pedagogical content knowledge (PCK)*.

Fourthly, the teacher should have *skills with respect to ICT*. Using technology, both on the handheld side as well as on the network side, should only result in a low cognitive load so that the teacher is able to concentrate on giving feedback and directing the classroom discourse towards meaningful interaction with respect to statistics. It takes a sustainable effort to maintain these skills in order to be able to smoothly switch to new tools or to new versions of familiar tools. PCK is nowadays supplemented by technological pedagogical content knowledge (TPCK) (Koehler & Mishra, 2009).

What is to be recommended based on the results of this study?

With respect to the **practice of mathematics education**:

Professional development; organisation at macro level

One of the participating teachers mentioned during this research project: "You have to do things like this in order to continuously develop your teaching." We would very much like to see mathematics education organised in such a way that experiences comparable to those the teachers had during this project would come within reach of every teacher. In short, teachers should be facilitated to become key participants in research projects. This requires an effort from school administrations and probably even from national policy makers.

Professional development; organisation at the micro level

We see mathematics as the science of patterns and structures. This study taught us, among other issues, that structuring mathematics education along a chosen principle (in our study: feedback in order to evoke a meaningful classroom discourse) is very important too: "It's the structure, stupid" (van der Hilst, 2010). We encourage teachers and other practitioners to really rethink the design of the mathematics education they are responsible for.

With respect to the **design of classroom networks:**

Better feedback function on the handheld side

On the handheld side of the classroom network immediate feedback is generated on the work of the student. This feedback just concerns knowledge of *results* and

knowledge of *correct response*. It is remarkable that with the classroom network we used on multiple choice answers there was no specific differentiated feedback with respect to the specific answers. Students and teachers had expected more sophisticated feedback. A big step further in the same direction should be the integration of some kind of intelligent tutoring system (ITS) (Sarrafzadeh, et al., 2008; Sleeman & Brown, 1982) on handhelds.

Tracking function of feedback

With the current state of development of classroom networks it is not possible to track the interaction a student has with the mathematical content (exercises) on his GC. This is a loss of a valuable source of information with which the teacher could further improve the feedback in the classroom discourse. To generate on the handheld side some kind of dynamic representation, for instance a screen movie that shows all student activity, combined with software on the network side that is able to condense this data flow into an understandable representation for the teacher, would be an innovative step facilitating further fine tuning of the teacher feedback.

We recommend with respect to **further research**

Continuation of this study

We recommend an expansion of this intervention study using the materials and research design we developed. This would preferably be for a longer period, for instance three chapters, if possible in schools where the Nspire is already the standard GC. With an expansion like that the 'start bias' that comes with working in a new learning environment could be reduced to a minimum.

Conducting this research among substantially more than six teachers, for instance between ten and twenty, should be enough to test the 'success conditions' we formulated. It would be very interesting after the first intervention to repeat it three years later, meanwhile letting the teachers optimise their teaching according to the principles we formulated. This should show significantly better results (Adey, 2006) and should facilitate a further outlining of the coding scheme and the correspondence metric, which we see as the main methodological contributions of this study.

Collaborative learning

A substantial number of the interviewed students mentioned that they experienced better learning through the input of their peers, as made accessible through the classroom network. Although we were aware that peer feedback is nowadays considered as a valuable learning process (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010), we did not design our prototype on this specific possibility. The fact that it is nevertheless mentioned spontaneously by quite some students indicates that in future design research studies this deserves structural attention. The same counts for peer assessment (Vickerman, 2009), a specification of peer feedback.

ICT mediated relationship between ASS and DL

This study did not problematise the influence of a change in procedural skills (ASS in this study) on conceptual skills (DL in this study) as is influenced by the use of ICT. This can be seen in the framework of what in recent years has been called '21st century skills' (Silva, 2009; Trilling & Fadel, 2009). For mathematics education,

traditionally quite strongly involved with the possibilities of ICT, this is a major domain of research. Specifically in our view, research is needed to investigate how the mediation of ICT can contribute to the mutual reinforcement of conceptual skills and procedural skills.

Samenvatting in het Nederlands

Een van de hardnekkige problemen in het wiskundeonderwijs is de perceptie van de wiskundeleraren dat zij onvoldoende lestijd hebben om hun leerlingen goed wiskunde te leren. De onderzoekshypothese van deze studie is dat de lestijd in het wiskundeonderwijs doeltreffender en doelmatiger kan worden door middel van een betere feedback. Uit veel onderzoek blijkt (Hattie, 2009) dat een focus op feedback een van de meest effectieve manieren is om de leeropbrengsten van lestijd te verbeteren. Computerapparaten verbonden in een netwerk bieden een goede mogelijkheid om feedback te geven. Tijdens een initiële studie (Tolboom, 2005) werd dit bevestigd in een omgeving met grafische rekenmachines die zijn verbonden in een classroom network. Daarom luidt de hoofdonderzoeksvraag van deze studie: "Wat zijn de mogelijkheden van een classroom network in de ondersteuning van docenten bij het geven van feedback in statistiekonderwijs?"

Om deze vraag te beantwoorden hebben we eerst een literatuuronderzoek rondom feedback, statistiekonderwijs en informatie- en communicatietechnologie (ICT) uitgevoerd om onze positie te bepalen.

In deze studie verstaan wij onder 'feedback' enerzijds 'het commentaar van de leraar, de grafische rekenmachine (GR) en de medeleerlingen op het werk van leerlingen, om dat werk te verbeteren', terwijl anderzijds het werk van leerlingen en hun gedrag feedback voor de docent zijn om eventueel de instructie op aan te passen.

We kozen ervoor om de statistische leeractiviteiten te verdelen in:

1. *Redeneren over en betekenis geven aan* (Martin, et al., 2009) *data*, (Cobb's (1991) 'data en concepten'), dat we in deze studie *data geletterdheid* (DG) noemen;
2. *Algoritmische statistische vaardigheden* (ASV), (Cobb's (1991) recepten).

Data geletterdheid (DG) behoort tot de conceptuele kennis die, naar onze mening, te weinig aandacht krijgt in het statistiekonderwijs. Zonder het vitale belang te ontkennen van procedurele vaardigheden (in dit onderzoek aangegeven met algoritmische statistische vaardigheden), hebben we in deze studie geprobeerd om DG van leerlingen te bevorderen. We hebben ons daarbij beperkt tot het subdomein van de beschrijvende statistiek. Er is geen reden om aan te nemen dat feedback in de beschrijvende statistiek belangrijker is dan in andere deelgebieden van de statistiek. De belangrijkste reden om te kiezen voor de beschrijvende statistiek is het feit dat statistiekonderwijs vaak begint met beschrijvende statistiek. Dat maakt het een didactisch interessant onderdeel omdat het daarmee mogelijk is om een goede indruk te geven van statistiek als wetenschap. Afgezien van het feit dat beschrijvende statistiek in een groot aantal wetenschappen wordt gebruikt, is het belangrijk om kennis te hebben van de hierin gebruikte begrippen en methoden voor diegenen die moeten functioneren in de huidige informatiesamenleving (Gal & Garfield, 1997). Maar behalve dat statistiek nuttig is voor vervolgstudie en als voorbereiding op de maatschappij vinden wij met Pereira-Mendoza en Swift (1981) dat esthetiek een belangrijk leerdoel is van statistiekonderwijs. Naar onze mening kan de schoonheid, voor de doelgroep van dit onderzoek (havo 4 wiskunde A), vooral worden gevonden in de mogelijkheid om verschijnselen uit het echte leven, verwoord in authentieke contexten (Wijers, Jonker, en Kemme, 2004), te analyseren met het gebruik van wiskundige technieken.

De onderzoeksbenadering die het meest geschikt leek om onze hoofdonderzoeksvraag te beantwoorden was die van onderwijsontwikkelingsonderzoek. Een iteratief proces van analyse, ontwerp, ontwikkeling, implementatie en evaluatie van een prototype van een onderwijsinterventie werd gepland.

We hebben de belangrijkste onderzoeksvraag in vier deelvragen gespecificeerd:

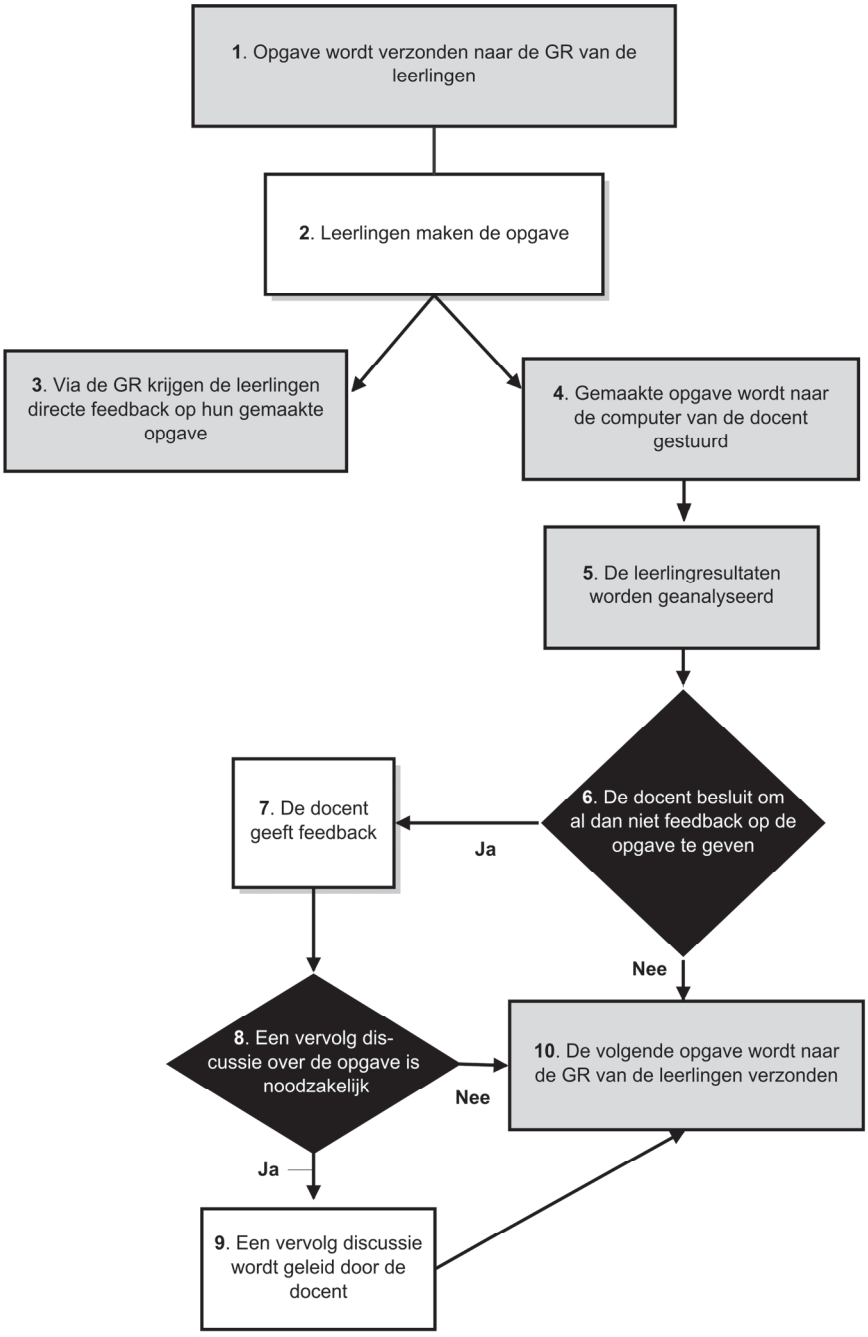
1. Is technologische ondersteuning door middel van het classroom network geschikt voor de beoogde feedback in de lessen? (*Conditie* vraag)
2. Is het mogelijk voor een wiskundedocent om het prototype van de interventie te implementeren in overeenstemming met de bedoelingen? (*Existentie* vraag)
3. Is de feedback ondersteuning van het classroom network even geschikt voor algoritmische statistische vaardigheden (ASV) als voor data geletterdheid (DG)? (*Didactische* vraag)
4. Welke docentkenmerken bevorderen / belemmeren de uitvoering van de interventie zoals bedoeld? (*Identificatie* vraag)

De ontwikkeling van het prototype van de interventie was gebaseerd op ontwerpprincipes met betrekking tot feedback en statistiekonderwijs (met ICT daar doorheen verweven). Deze ontwerpprincipes zijn voor een deel gebaseerd op aanbevelingen uit de literatuur en voor een deel op eigen keuzes. De belangrijkste van deze principes werden ingekaderd in de feedbackmatrix voor de statistiek onderwijs (zie hieronder). Hierin worden statistische opgaven gecategoriseerd met betrekking tot het beoogde leerdoel (DG of ASV) en de timing van de feedback (onmiddellijk of uitgesteld).

Timing van feedback / Leerdoel	Onmiddellijk	Uitgesteld
Data geletterdheid	Type I	Type II
Algorithmische statistische vaardigheden	Type III	Type IV

Feedbackmatrix voor statistiekonderwijs

We gebruikten opgavetypen als ‘meerkeuze’ en ‘korte invoer’ (vragend om een numeriek antwoord van de leerling) om type I en type III feedback te ondersteunen. Het opgavetype ‘open antwoord’ (vragend om een zelf geformuleerd antwoord van de leerling) was geschikter voor het ondersteunen van type II en type IV feedback. Onmiddellijke feedback werd geleverd door de grafische rekenmachines van de leerlingen en vertraagde feedback werd gegeven door de docent tijdens de erop volgende les. Een schematisering van het feedbackproces is hieronder afgebeeld.



Schema van het feedbackproces

Om de leerlingen zo actief mogelijk te laten deelnemen aan de lessen gebruiken we de *doorvraag aanpak*: probeer zo veel mogelijk van de leer-en onderwijsactiviteiten in vraagvorm te gieten. Leerlingreacties zijn een mogelijkheid voor de docent om feedback op te geven en ze zijn zelf feedback voor de docent, omdat zij hem of haar informatie verschaffen over de kennis van de leerlingen.

Gedurende drie empirische fasen (C1, C2, C3) werden de drie prototypes van de interventie uitgevoerd, geëvalueerd en aangepast. We ontwikkelden hypothetische doceer trajecten (HDT) om onze intenties met betrekking tot het gewenste docentgedrag expliciet te maken. Alle lessen werden geobserveerd, op video opgenomen en geanalyseerd. We ontwikkelden een coderingssysteem om het HDT en de gerealiseerde feedback mee te comprimeren en een metriek om de correspondentie tussen de beoogde interventie (zoals

beschreven in het HDT) en de gerealiseerde interventie (zoals waargenomen) uit te drukken. Elke empirische fase werd ook geëvalueerd met behulp van een vragenlijst voor alle deelnemende studenten, interviews met deelnemende studenten (drie uit elke groep) en interviews met de docenten. Na C3 werd een groepsinterview met alle deelnemende docenten georganiseerd om een kader voor de opgedane ervaringen te geven.

Fase C3, waarin het derde prototype van de interventie is uitgevoerd gedurende zes casestudies, kan worden samengevat in een correspondentieschema als hieronder weergegeven. De kolommen vertegenwoordigen respectievelijk de naam van de case studie, de gemiddelde correspondentiescore, de standaardafwijking van de correspondentiescore, de gemiddelde correspondentie met betrekking tot DG, de gemiddelde correspondentie met betrekking tot ASV, het verschil tussen de laatste twee en het percentage ontbrekende feedback sessies. De correspondentie van de gerealiseerde feedbacksessies wordt uitgedrukt in een score variërend van 0 (geen enkele correspondentie) tot 7 (perfecte correspondentie). Deze correspondentie geldt ten opzichte van het HDT.

Case	Gemiddelde	SD	GEM_DG	GEM_ASV	GEM_DG-GEM_ASV	%-Ontbrekend
S1	5.14	1.70	4.89	5.60	-0.71	68.09
S2A	4.40	2.29	3.66	5.71	-2.05	14.89
S3	5.38	1.85	5.34	5.46	-0.12	0.00
S4A	3.60	2.13	3.65	3.44	0.21	31.91
S4B	3.89	2.09	4.11	3.33	0.78	34.04
S2B	2.04	1.70	1.95	2.25	-0.30	59.57

Correspondentieschema van de casestudies in C3

We zien een aanzienlijke variatie van de gemiddelde correspondentie en van het percentage ontbrekende feedbacksessies tussen de verschillende case studies. Om dit min of meer kwantitatieve overzicht op een kwalitatieve manier te duiden, analyseerden we de klassikale discussie ook op een meer microscopisch niveau. Uit elke casus presenteerden we een aantal voorbeelden van gerealiseerde feedbacksessies. Een veelgehoorde kritiek op de klassieke rapportages van onderwijsontwikkelingsonderzoek betreft de exemplarische manier waarop de resultaten worden gepresenteerd. Wij denken dat we deze kritiek voor een groot deel tegemoet komen door de rapportage van alle correspondentiescores. Daarnaast hebben we voor C3 expliciet de wijze van steekproef trekken beschreven. In de eerste plaats moest het geselecteerde voorbeeld substantieel zijn. Dat wil zeggen, de voorbeelden moesten één of meer gebeurtenissen bevatten die belangrijk zijn vanuit het perspectief van onze hoofdonderzoeksvraag. Door dit criterium te hanteren voor de wijze van steekproef trekken is het waarschijnlijk dat de correspondentiescore van de geselecteerde gebeurtenissen hoger is dan de gemiddelde correspondentie score van de case studie. Verder hebben we hoofdzakelijk voorbeelden gekozen die betrekking hadden op DG. Wij zijn van mening dat ASV zeer belangrijk is, maar dat de instructie van DG problematischer is. Het verbeteren van die instructie is een van de belangrijkste doelen van deze studie. Vervolgens moest het geselecteerde

voorbeeld ‘op één of andere manier representatief' voor de casus als geheel zijn. Dat wil zeggen dat de klassikale discussie elementen moest bevatten die min of meer typerend zijn voor de specifieke casus. Daarom hebben we besloten om per casus twee voorbeelden van klassikale discussie presenteren. Om een meer valide vergelijking te maken tussen de case studies, hebben we geprobeerd om een opgave te vinden met DG als statistisch leerdoel, die bovendien aanleiding gaf tot serieuze klassikale discussie in alle van de zes case studies. Door de grote problemen bij de eerste en de laatste case studies was dit niet realiseerbaar. De klassikale discussie voortvloeiend uit de feedback op opgave 8.8 kwam het meeste in de buurt van dit criterium. Met deze opgave, met daarin de vraag om een conclusie te trekken over het verschil in computergedrag van jongens en meisjes, hebben we een punt voor de vergelijking van de feedback tijdens de eerste vijf case studies. Deze vijf waren de beste van de zes case studies gelet op de gemiddelde correspondentie tussen het HDT en het gerealiseerde curriculum.

Na het synthetiseren van alle resultaten (vragenlijsten, interviews, lesobservaties) kwamen we tot de volgende beantwoording van de vier onderzoeksvragen:

1. Is technologische ondersteuning door middel van het classroom network geschikt voor de beoogde feedback in de lessen? (*Conditie* vraag)

   Door technologische instabiliteit was dit niet het geval toen we begonnen aan deze studie. Maar na C3, waarin het tijdens case studie S3 mogelijk was om alle beoogde feedback sessies uit te voeren, hebben we geconcludeerd dat technologische ondersteuning van het classroom network adequaat was.

2. Is het mogelijk voor een wiskundedocent om het prototype van de interventie te implementeren in overeenstemming met de bedoelingen? (*Existentie* vraag)

   We beschouwen case studie S3 als een overtuigende uitvoering van de voorgenomen interventie. Deze had een hoge gemiddelde correspondentiescore (5,38 op een schaal van 0 tot 7), zowel met betrekking tot ASV als tot DG. Leerlingen en leraar waren even enthousiast over de verbetering van de feedback. Alle feedback sessies werden uitgevoerd, waaruit blijkt dat de technologie de ingreep uitstekend ondersteunde. Naast de in zijn geheel overtuigende case studie S3 was het opmerkelijk dat in elke andere case studie er feedbacksessies waren met een overtuigende correspondentie score. Dit betekent dat elke docent, onder de juiste omstandigheden, in staat is geweest om een feedbacksessie uit te voeren zoals gepland. We beschouwen die feedbacksessies als ‘micro existentie bewijzen’: de docenten zijn er in geslaagd ten minste één feedbacksessie uit te voeren volgens de intenties, terwijl de leerlingen overtuigd waren van de kwaliteit van de gerealiseerde feedback.

3. Is de feedback ondersteuning van het classroom network even geschikt voor algoritmische statistische vaardigheden (ASV) als voor data geletterdheid (DG)? (*Didactische* vraag)

   Dat de ondersteuning voor ASV beter is dan die voor DG wordt aangetoond door het feit dat het gemiddelde verschil in de correspondentiescore tussen ASV en DG 0,37 was (in het voordeel van ASV). We beschouwen dit verschil in gemiddelde, van een variabele met een bereik van 0 tot 7, als klein.

De ‘ingebouwde’ ondersteuning van het classroom network voor de ontwikkeling van DG bij de leerlingen moet worden aangevuld met een specifieke didactische voorbereiding van de docent.

Met een zeer specifiek HDT zijn we er op een bevredigende manier in geslaagd om de docenten te ondersteunen bij het geven van feedback op DG.

4. Welke docentkenmerken bevorderen / belemmeren de uitvoering van de interventie als bedoeld? (*Identificatie* vraag)

   De resultaten van C3 maken aannemelijk dat het verbeteren van feedback in statistiekonderwijs door het gebruik van een classroom network mogelijk is. Het kleine verschil in scores tussen de case studies S4A en S4b (met dezelfde leraar voor verschillende groepen) suggereert dat de correspondentie meer docentafhankelijk is dan groepsafhankelijk. Dit brengt ten aanzien van de docent een interessante vraag te berde: wat zijn de docentkenmerken die de variatie in de correspondentiescore veroorzaken?

   We concludeerden dat er minstens *vier voorwaarden* zijn waaraan moet worden voldaan voordat een docent, ondersteund zoals tijdens C3, in een leeromgeving die is technisch stabiel is, ten volle de feedback mogelijkheden van het classroom network in statistiekonderwijs kan gebruiken.

   Allereerst moet er tussen docent en klas een relatie bestaan die berust op voldoende wederzijds vertrouwen. Als dit vertrouwen ontbreekt, is alle onderwijs gedoemd te mislukken, hoe goed de mogelijkheden van de rest van de leeromgeving ook zijn. Goed onderwijs is een intiem proces. Feedback en klassikale discussie zijn misschien wel de meest kwetsbare delen van dat proces. Wederzijds vertrouwen is onontbeerlijk voor het slagen hiervan. De docent is hiervoor de meest bepalende enkelvoudige factor.

   Ten *tweede*, de leraar moet sterke communicatieve vaardigheden hebben, beginnend met een communicatieve houding (of misschien zelfs ‘persoonlijkheid’?) Met deze vaardigheden kan zij of hij de klassikale discussie zo productief mogelijk maken. Dit betekent dat hij of zij een ‘dirigent’ (Drijvers, Doorman, Boon, Reed, & Gravemeijer, 2010) moet zijn van de klassikale discussie, die in deze context moet worden geïnterpreteerd als ‘de spin in het web van het onderwijsproces’. Een voldoende *functionele extraversie* is nodig om deze communicativiteit ook onder druk te kunnen etaleren. In het algemeen geldt dat het repertoire van de docent wat betreft formatieve toetsing (Zwart, et al., 2003;. Black & Wiliam, 2009) en dialogisch onderwijs (Alexander, 2008) op een voldoende niveau moet zijn.

   Ten *derde* moet de docent goed zijn in het snel interpreteren van de leerlingantwoorden omdat hij daarvan een groter aantal te verwerken krijgt dan zonder het gebruik van een classroom network. Hij of zij moet in staat zijn om ‘statistische zin’ te maken van veel meer leerlinginbreng dan voorheen. Hij of zij moet snel een ruw ‘feedback schema’ kunnen maken waarvoor hij of zij die leerlingantwoorden selecteert die samen de volledige antwoordenruimte opspannen en daarmee een goede start kunnen zijn voor een brede klassikale discussie. Dit vereist een grote mate van kennis van het onderwerp en van wiskundedidactische kennis (pedagogical content knowledge, PCK).

Ten *vierde* moet de docent over ICT vaardigheid beschikken. Het gebruik van technologie, zowel aan de GR kant als aan de netwerkkant, mag niet resulteren in een hoge cognitieve belasting, zodat de docent in staat is om zich te concentreren op het geven van feedback en het leiden van de klassikale discussie naar zinvolle interactie over statistiek. Het vergt een duurzame inspanning om deze vaardigheden te verwerven en te behouden om vlot te kunnen overschakelen naar nieuwe hulpmiddelen of nieuwe versies van vertrouwde tools. De vereiste PCK wordt tegenwoordig aangevuld met technologische vakdidactische kennis tot TPCK (Koehler & Mishra, 2009).

Welke aanbevelingen volgen uit de conclusies van deze studie?

Met betrekking tot de **praktijk van wiskundeonderwijs**:

Professionele ontwikkeling; organisatie op macro niveau

Een van de deelnemende leerkrachten merkte tijdens dit onderzoek op: "Je moet dit soort dingen doen om je eigen onderwijs continu te ontwikkelen." Wij bevelen aan dat wiskundeonderwijs op een zodanige manier georganiseerd wordt dat ervaringen vergelijkbaar met die van de docenten tijdens dit project binnen handbereik komen van elke docent. Docenten moeten, kortom, worden gestimuleerd om belangrijke deelnemers te worden aan onderzoeksprojecten. Dit vereist een inspanning van schooldirecties en waarschijnlijk zelfs van nationale beleidsmakers.

Professionele ontwikkeling; organisatie op micro niveau

We zien wiskunde als de wetenschap van patronen en structuren. Deze studie heeft ons geleerd, onder andere, dat het structureren van wiskundeonderwijs volgens een gekozen principe (in ons onderzoek: feedback met het doel een zinvolle klassikale discussie op te roepen) erg belangrijk is: "Het is de structuur, sufferd!" (Van der Hilst, 2010). We moedigen docenten aan om het ontwerp van het wiskundeonderwijs waar ze voor verantwoordelijk zijn van tijd tot tijd serieus tegen het licht te houden.

Met betrekking tot het **ontwerp van classroom networks:**

Betere feedback mogelijkheden op de handheld

Aan de handheldkant van de classroom network wordt onmiddellijke feedback gegeven op het werk van de leerling. Deze feedback betreft alleen maar kennis van de resultaten en kennis van het juiste antwoord. Het is opmerkelijk dat het systeem dat we gebruikten in geval van meerkeuzeantwoorden geen specifieke feedback mogelijk maakte op de specifieke antwoorden. Leerlingen en docenten hadden meer geavanceerde feedback verwacht. Een grote stap in deze richting zou de integratie van een intelligent tutoring systeem (ITS) (Sleeman & Brown, 1982 Sarrafzadeh, et al., 2008) op de handhelds zijn.

Leerlingvolgfunctie

Met de huidige generatie classroom networks is het niet mogelijk om de interactie die een leerling heeft met de wiskundige inhoud (opgaven) op zijn GR bij te houden. Dat betekent een verlies van een waardevolle bron van informatie waarmee de docent de feedback en de klassikale discussie verder kan verbeteren. Een volgende stap voor de ondersteuning van feedback door het classroom network zou kunnen zijn om aan de handheldkant een dynamische representatie van die interactie op te slaan, bijvoorbeeld een schermfilm die alle leerlingactiviteit laat zien, in combinatie met software aan de netwerkkant die in staat is om de gegevensstroom die deze dynamische interactie oplevert begrijpelijk weer te geven voor de leraar,.

Met betrekking tot **verder onderzoek**

Vervolg op deze studie

Wij raden aan deze interventie studie uit te breiden, daarbij gebruik makend van de materialen en onderzoeksopzet zoals ontwikkeld tijdens deze studie. Deze uitbreiding kan een langere interventie behelzen, bijvoorbeeld gedurende drie hoofdstukken, indien mogelijk in scholen waar de TI Nspire al de standaard GR is. Met een uitbreiding als deze kan de 'start ruis', die hoort bij het werken in een nieuwe leeromgeving, worden beperkt.

Het uitvoeren van dit onderzoek onder aanzienlijk meer dan zes docenten, bijvoorbeeld tussen de tien en twintig, moet voldoende zijn om de succes voorwaarden die wij hebben geformuleerd verder te toetsen. Het zou zeer interessant zijn om die interventie na drie jaar te herhalen, terwijl de docenten ondertussen hun onderwijs optimaliseren op basis van de principes die wij hebben geformuleerd of de eventueel bijgestelde versies daarvan. Dit moet significant betere feedbacksessies opleveren (Adey, 2006) en dit moet leiden tot een verdere verfijning en stabilisering van het codeerschema en de correspondentiemetriek, die wij zien als de belangrijkste bijdragen van deze studie aan de methodologie van onderwijsontwikkelingsonderzoek.

Samenwerkend leren

Een aanzienlijk aantal van de ondervraagde leerlingen heeft tijdens interviews gezegd dat ze beter leren door middel van de inbreng van hun medeleerlingen, toegankelijk gemaakt door het classroom network. Hoewel we wisten dat peer feedback tegenwoordig wordt beschouwd als een waardevol leerproces (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010), hebben we het ontwerp van ons prototype niet op deze specifieke mogelijkheid afgestemd. Het feit dat het toch door veel leerlingen spontaan wordt genoemd geeft aan dat het aspect van samenwerkend leren in de toekomst tijdens onderwijsontwikkelingsstudies als deze structurele aandacht verdient. Hetzelfde geldt voor peer assessment (Vickerman, 2009), een specificatie van peer feedback.

ICT ondersteunde relatie tussen ASV en DG

Tijdens deze studie werd de verandering van procedurele vaardigheden (ASV in deze studie) als gevolg van een verandering in conceptuele vaardigheden (DG in deze studie) onder invloed van ICT gebruik niet geproblematiseerd. Dit kan worden gezien in het licht van wat in de afgelopen jaren is genoemd '21ste eeuwse vaardigheden' (Silva, 2009; Trilling & Fadel, 2009). Voor wiskundeonderwijs, traditioneel sterk geïnteresseerd in de mogelijkheden van ICT, zou dit een belangrijk domein van onderzoek moeten zijn. In onze ogen is met name onderzoek nodig om te onderzoeken hoe het gebruik van ICT kan bijdragen aan de wederzijdse versterking van conceptuele vaardigheden en procedurele vaardigheden.

Curriculum vitae

Jos Tolboom werd geboren op 3 september 1966 in Emmen. In 1984 ging hij aan de Rijksuniversiteit Groningen natuurkunde studeren. De propedeuse maakte hem duidelijk dat in hem geen experimenteel natuurwetenschapper school. Daarom begon hij in 1985 aan de studie econometrie. Hij werd gegrepen door het wiskundige en statistische deel van die opleiding. Langzaam kwam hij er achter dat het econometrische gebruik van de wiskunde hem weliswaar technisch uitdaagde, maar dat hij de filosofische insteek op de wiskunde miste. Dat was één van de redenen waarom hij na zijn doctoraal in 1991 begon aan de postdoctorale eerstegraads lerarenopleiding wiskunde. Hij combineerde dat met een aanstelling als wiskundeleraar aan het Rölingcollege in Groningen en haalde in 1992 zijn bevoegdheid. Vanwege zijn fascinatie voor de mogelijkheden van computers en met name hun educatieve toepassingen begon hij in 1998 aan de eerstegraadsopleiding tot docent informatica bij het Consortium Omscholing Docent Informatica (CODI, een samenwerking van de UTwente, Open Universiteit en de Rijksuniversiteit Groningen met Hogeschool Windesheim en Fontys Hogeschool). Deze opleiding rondde hij in 2000 af. Van 1994 tot 2001 was hij leider van de sectie wiskunde en informatica aan het Rölingcollege. In de periode 1997 tot 2001 combineerde hij zijn leraarschap met een aanstelling aan de Rijksuniversiteit Groningen als lerarenopleider wiskunde. Van 1999 tot 2001 was hij bestuurslid van de Vereniging *i*&i, de vereniging voor ICT in het onderwijs en de vakvereniging voor docenten informatica. In 2001 verliet hij het voortgezet onderwijs en werd hij docent wiskunde in de bacheloropleiding van de Faculteit der Wiskunde en Natuurwetenschappen aan de RUG en docent educatie en communicatie in de masteropleiding van diezelfde faculteit. Van 2001 tot 2006 was hij redacteur informatie- en communicatietechnologie van Euclides, het tijdschrift van de Nederlandse Vereniging van Wiskundeleraren. In de periode 2006-2009 was hij als vaksectievoorzitter van het College voor Examens (CvE) verantwoordelijk voor de Nederlandse centrale examens wiskunde A en C voor havo en vwo. In september 2010 verliet hij de RUG en trad hij in dienst bij SLO, het nationaal expertise centrum leerplanontwikkeling, als leerplanontwikkelaar wiskunde bij de afdeling tweede fase havo-vwo.

Dankwoord

Na gedane zaken is het goed danken. Terugblikkend kan ik zeggen dat veel mensen een rol hebben gespeeld bij de totstandkoming van dit proefschrift.

In mei 1978 fietste ik met mijn moeder naar het Gemeentelijk Lyceum in Emmen. Ik had de Cito eindtoets basisonderwijs gemist door een blindedarmontsteking. Het strenge lyceum hechtte veel waarde aan het advies van Bé Wieringa, hoofdonderwijs aan de Titus Brandsmaschool, om mij "een school voor gymnasiaal onderwijs te laten volgen". Maar ik moest natuurlijk nog wel een toelatingsexamen maken. Op de gevel van de school hing een ornament met daarin een passer. Het onderschrift luidde 'Navigare necesse est' (*Varen* is noodzakelijk). In het Nederlands krijgt dit ook wel de betekenis '*Navigeren* (of zelfs: sturen?) is noodzakelijk'. Van de docente klassiek talen Monique van der Hoeven leerde ik in 1983 dat Gnaeus Pompeius rond 50 voor Christus deze woorden sprak en toen de matrozen op het schip met voedsel van Tunesië naar Rome een vliegende storm in stuurde met de toevoeging '*Vivere* non est necesse' (Leven is niet noodzakelijk).

Eerst de navigatie. In september 1991 startte ik met de lerarenopleiding wiskunde. Ik had twee begeleiders, die samen mijn visie op wiskundeonderwijs grotendeels hebben gevormd. Op de universiteit leerde Sieb Kemme mij didactisch te denken vanuit de doelen van wiskunde en daarbij iedere les, ik herhaal: iedere les, grondig voor te bereiden. Op school leerde Menno van Steenis mij dat een goede wiskundeleraar elke dag leert van zijn leerlingen om zo zijn onderwijs te verbeteren. Sieb en Menno staan daarmee aan de basis van deze studie.

In september 2001 startte Anne van Streun een onderzoekslijn bètadidactiek aan de faculteit der wiskunde en natuurwetenschappen van de RUG. Ik mocht daarin een dag per week aan de slag met een project dat ik noemde 'Feedback in digital learning environments'. Dat project zou kunnen uitmonden in een promotie. Anne gaf mij alle vrijheid om dit onderzoek vorm te geven. Van hem leerde ik de basis van het ontwerpen van wiskundeonderwijs. Mijn vrijheid bleef toen Anne terug trad en Jan van Maanen in 2003 de leiding van de bètadidactiek overnam. Jan en ik verworven een NWO-subsidie voor een ondersteunend onderzoek 'Inhoud voor WWWiskunde', dat werd uitgevoerd door Léon Tolboom (ja, wel familie). Tijdens dat onderzoek, gericht op de interactieve mogelijkheden van het world wide web in wiskundeonderwijs, besloot ik de koers van mijn eigen onderzoek te verleggen naar het bestuderen van de mogelijkheden van mobiele apparatuur in wiskundeonderwijs. Van Jan leerde ik preciezer te zijn. Van Léon leerde ik dat alle geleerde redeneringen niet alleen aan een normale wiskundedocent uit te leggen moeten zijn, maar hem of haar zelfs moeten kunnen stimuleren in het nadenken over wiskundeonderwijs. In 2004 trad Martin Goedhart aan als hoogleraar bètadidactiek aan de RUG. Martin heeft een duidelijke focus op onderzoek en een eigenzinnige visie daarop. Samen met Joke Voogt (UTwente) probeerde hij mij op een pad te krijgen waarvan steeds duidelijker werd dat het niet het mijne was. In november 2009 eindigde onze samenwerking. Van Martin en Joke had ik inmiddels al wel geleerd nog kritischer naar mijn eigen teksten te kijken.

Na het afscheid van Martin en Joke nam ik contact op met Jan van den Akker. Hij zag het belang van mijn onderzoek. Van hem heb geleerd hoe belangrijk het is dat een onderzoeksvraag aantoonbaar relevant is. Jan koppelde me aan Wilmad Kuiper die vanaf begin 2010 de zwaarste stem heeft gehad in de begeleiding van deze studie. Wilmad leerde mij hoe je consistentie aanbrengt in een studie.

Ondertussen gingen er stemmen op die vonden dat dit proefschrift toch in Groningen moest worden afgerond. Henk Broer was bereid als eerste promotor op te treden. Wilmad ging daarmee zonder enige aarzeling akkoord, terwijl hij de promotie net zo goed naar zijn eigen club aan de Universiteit Utrecht had kunnen halen. Dat gebrek aan profileringsdrang nam mij nog sterker voor hem in. Henk Broer kende ik al sinds mijn tijd als student aan de RUG. Er is niemand ter wereld die zoveel van wiskunde weet en met een bulderende lach daarbij zo hard op een tafel kan slaan. Dat hij als promotor van dit proefschrift wil optreden is voor mij een grote eer en, zo mogelijk, een nog groter genoegen.

Intussen was mij geleidelijk aan duidelijk geworden dat ik aan de RUG niet het werk kon doen dat ik graag zou willen. In september 2010 trad ik daarom in dienst bij SLO, waar ik onder andere mijn oud-docente klassieke talen Monique van der Hoeven als collega kreeg. Jan van den Akker, algemeen directeur, Hetty Mulder, manager afdeling tweede fase havo-vwo, en Wilmad Kuiper, manager afdeling onderzoek en advies, gaven mij bij SLO direct het gevoel dat ik van essentiële waarde voor de organisatie kon zijn. Ik mocht een dag per week besteden aan mijn promotieonderzoek. Ik ben hen en de organisatie daarvoor erkentelijk.

Texas Instruments heeft mij alle mogelijkheden geboden om de hardware en software die in deze studie onder de loep liggen te gebruiken en mij in de implementatie daarvan ondersteund. In het bijzonder Mark de Hiep, Pieter Schadron, Epi van Winsen en Gert Treurniet hebben zich zeer ingespannen. Het welslagen van deze studie hing voor een belangrijk deel af van de laatste empirische ronde van het onderzoek. Maar op de allereerste school van deze ronde kregen we het netwerk niet aan de praat in het lokaal waar de lessen waren gepland. Gert bleef drie nachten bij mij op zolder logeren, het hele huis was bezaaid met grafische rekenmachines, computers, hubs, routers en oplaadapparatuur. We kregen alles overal aan de praat. Behalve in het bewuste lokaal. 's Avonds om half elf voor de eerste les heb ik de roostermaker gebeld en om een ander lokaal gevraagd. Daar werkte het wel. Nooit leek een overwinning meer op een nederlaag voor de techneuten Gert en Jos.

Deze studie was onmogelijk geweest zonder de enthousiaste inzet van wiskundedocenten die iets zagen in deze manier van lesgeven. Ik dank Léon Tolboom, Martin Traas, Bert Wikkerink, Alfred van den Berg, Hugo Bronkhorst, Tom Roedema en Gea Passies voor de vele uren die zij in dit project hebben gestoken.

Tot slot ben ik professionele dank verschuldigd aan Jan van den Akker, Jan van Maanen en Ernst Wit voor het beoordelen van dit proefschrift en de verbeteringen waarin dit heeft geresulteerd.

Dan het leven. Weg met Gnaeus Pompeius! (Zegt Jantine.) Mijn moeder Lia en mijn vader Frank hebben mij nooit overdreven gestimuleerd om te leren. Dat was ook niet nodig. Zij hebben mij laten zien wat liefde is. Voor elkaar, voor Léon en mij, en voor mensen in het algemeen. Dat is belangrijkste wat ik heb meegekregen.

In de periode dat ik werkte aan deze studie werd het lief-en-leed-potje niet gespaard. Lief waren natuurlijk de geboortes van Krijn (2006) en Huib (2008). Leed kwam vooral door de ziektes van mijn schoonvader Wim en mijn vader Frank. Wim overleed in 2009. Frank heeft vele medici verbaasd leeft gelukkig vrolijk verder. Heel hard loopt hij niet meer, maar dat geldt eigenlijk ook voor mijzelf. Dat ik nog kan hardlopen met andere oud-GVAV-ers zoals Arjen Sein, Joke Spikman, Joost Wessels en Okko Jan Bosker stemt mij echter dankbaar. Dat Joost en Okko Jan mij op de dag van de promotie mij als paranimfen

terzijde wilden staan was de best mogelijke verzekering voor een goede afloop. Samen hadden we immers wel zwaardere inspanningen volbracht.

Jantine, ondertussen, zat niet stil tijdens mijn promotietraject. Naast haar eigen loopbaan bij het openbaar ministerie en recentelijk de zittende magistratuur had zij nog de puf mij er keer op keer op te wijzen dat voor het afronden van een promotietraject focus nodig is, en soms inschikkelijkheid. Bovendien heeft zij de Nederlandse teksten in dit proefschrift geredigeerd. Streng doch rechtvaardig, zoals het een rechter betaamt. Het belangrijkste van alles was uiteraard dat ze mij voor leuke dingen in weekeindes en vakanties bij het onderzoek weg plukte. Ik hoop en verwacht dat Krijn en Huib later niet zullen zeggen "We hebben pa pas leren kennen toen we zijn proefschrift lazen."

Men zegt wel eens dat praktische ondersteuning in tijden van stress het meest waardevol is. Daarom wil ik mijn schoonmoeder Mineke Nolta bedanken. Het praktisch en handig meedenken met de problematiek van het tweebanengezin en daar dan hulp op maat bij bieden beheerst zij als de beste.

Dankbaar ben ik dus voor iedereen die heeft geholpen bij het voltooien van wat mij voor ogen stond: uitzoeken hoe draagbare technologie wiskundedocenten kan helpen bij het verbeteren van feedback in hun lessen.

Ik heb geleerd, ik heb geleefd. Ik heb gezegd.